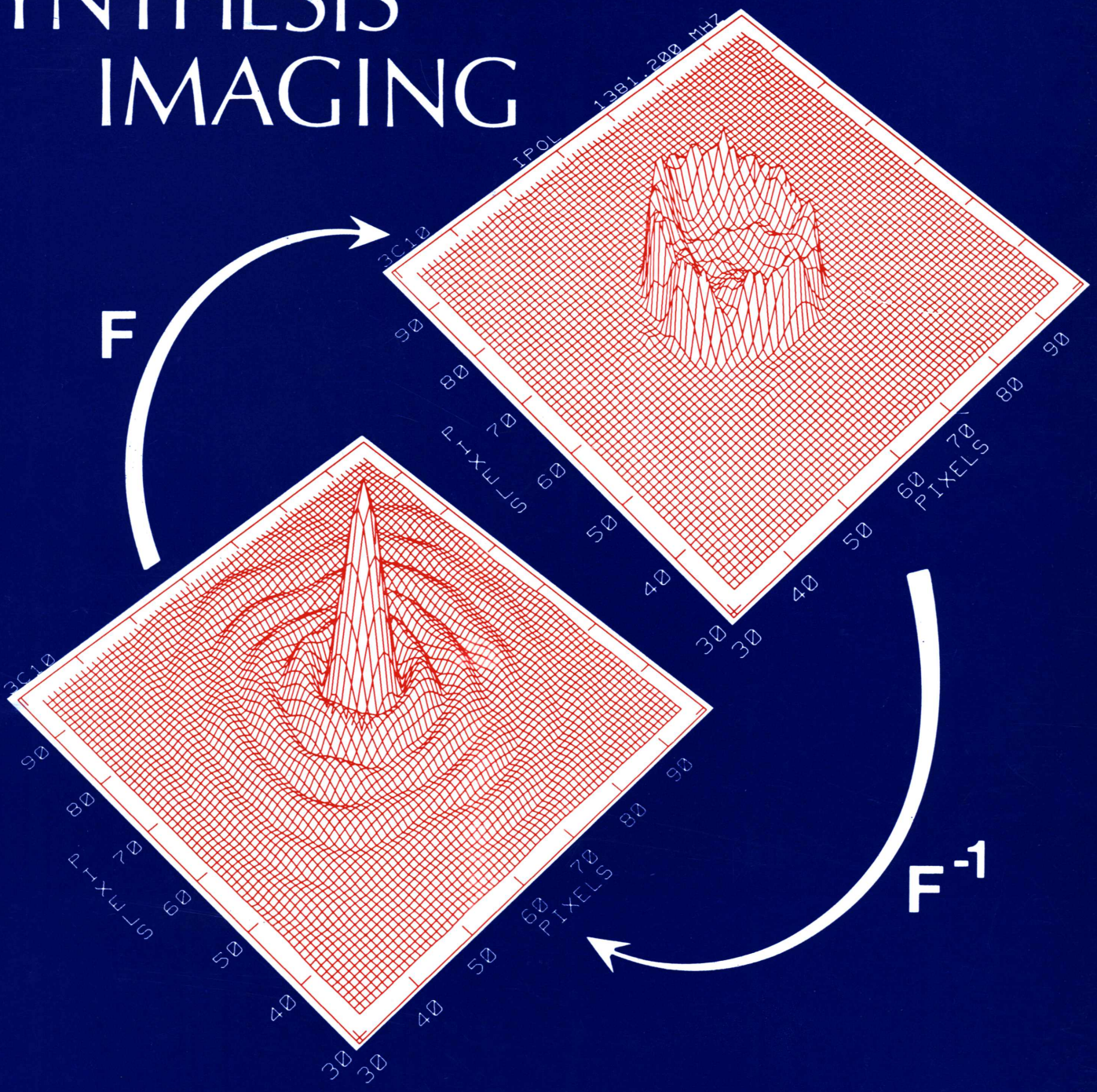
SYNTHESIS IMAGING



Course Notes from an NRAO Summer School
Held in Socorro, New Mexico
August 5–9, 1985

Edited by Richard A. Perley, Frederic R. Schwab, and Alan H. Bridle

SYNTHESIS IMAGING

Course Notes from an NRAO Summer School held in Socorro, New Mexico August 5-9, 1985

Edited by
Richard A. Perley, Frederic R. Schwab and Alan H. Bridle

Workshop No. 13

Distributed by:
National Radio Astronomy Observatory
P.O. Box 2
Green Bank, WV 24944-0002 USA

The National Radio Astronomy Observatory is operated by Associated Universities, Inc., under contract with the National Science Foundation.

Cover illustration.

(Top) A computer-generated perspective drawing of a synthesis image of the radio source 3C 10, a galactic supernova remnant. This image was obtained from VLA observations at 1381 MHz.
(Bottom) A perspective drawing of the visibility data from which this image of 3C 10 was derived. The visibility data are samples of the inverse Fourier transform of the distribution on the sky of the radio emission from this object. (Only the visibility amplitude is represented in this drawing.)

(These data were provided by Tim Cornwell.)

ACKNOWLEDGMENTS

Organizing a lecture series involving nearly 200 attendees was a hectic, interesting, and rewarding exercise. Probably the most important lesson learned from this has been how important it is to be able to rely on others for assistance in the planning and running of such an endeavor. Many individuals greatly assisted in the organization, and it is with great pleasure that we name their roles in this short section.

First, Ron Ekers provided constant and excellent advice about topics and logistics. His support was essential to the lecture series, and he should be considered the 'hidden' organizer.

Bob Dorr and Don Swann organized the logistics. It was they who ensured that meals and snacks appeared on time, that buses—and drivers—appeared where and when they should, and that essential materials were where they were needed.

The cheerful cooperation of the secretarial staff of the VLA was very important. They stuffed letters, collated the late lectures, duplicated materials, etc. We wish to especially thank Alison Patrick, who typed many of the first drafts.

Thanks go to the Macey Center staff, especially to Edrie Romans and Shannon Sandstrom, for their efforts in preparing the Macey Center for the meeting. Preparing the NMIMT dormitories was done by John Starckey, who did an excellent job.

Many of our summer students provided considerable needed assistance. We thank Brian Ritter, Dave Murphy, John Hibbard and Robin Price, who stapled lectures and were the guides who helped you get through the Albuquerque airport. We are particularly grateful to Rick Wietfeldt, who was of great assistance in practically every aspect of the organization of the meeting. Greg Hennessy, working alone, was responsible for 'Lecture 17'. You may thank him directly for his efforts.

Pat Palmer organized and led the hike to the top of South Baldy.

Thanks go to the lecturers, for getting their drafts in on time. Their enthusiasm for the series was essential to its success, for, without their cooperation, the lectures could not have taken place.

It is a great pleasure to thank Eva Fomalont for her outstanding assistance in organizing the lecture series. She worked very long hours, tending to all the details that we didn't, couldn't, or wouldn't notice. The organizational efforts would have been considerably more difficult without her help.

Enormous thanks are due the NRAO Graphics Department: to George Kessler and Patricia Smiley, for their manifold contributions of skillful draftmanship and art work—Pat was responsible for the cover design—and to the Green Bank staff, for the photographic services which they provided.

We wish also to acknowledge the contribution of Donald E. Knuth, inventor of the public-domain computerized typesetting program, TEX, which was used in producing this volume; and we would like to thank the American Mathematical Society for their financial support of his endeavor.

THE EDITORS

Table of Contents

PREFACE

1. The Purpose of the Course.

The NRAO Summer School on *Synthesis Imaging* held in Socorro from August 5th to 9th, 1985 was the second occasion on which NRAO staff prepared a series of lectures for serious students of synthesis imaging and image processing. NRAO operates one of the world's most powerful synthesis telescopes—the Very Large Array (VLA)—and is building another—the Very Long Baseline Array (VLBA). The main purpose of this course, like that of its predecessor in 1982, was to inform potential users of these telescopes about the principles on which synthesis instruments operate, about the subtleties of data acquisition, calibration and processing associated with them, and about techniques for getting the best results from them.

As such, it is aimed primarily at radio astronomers who are relative newcomers to the field of synthesis imaging, e.g., beginning graduate students or those whose research has hitherto not employed synthesis techniques. The subject matter is also of interest to people outside the traditional radio astronomy community—to astronomers whose expertise is primarily in observations at shorter wavelengths, to astrophysicists who wish to interpret the data from synthesis telescopes, and to researchers employing Fourier methods or deconvolution techniques in other fields of imaging, such as physics, medicine or remote sensing. We have therefore confined the detailed discussion of topics relating to NRAO's instruments to a few portions of the course, and have attempted to emphasize general principles wherever possible.

Nevertheless, the lectures reflect the close association of the lecturers with NRAO's instruments, especially with the VLA. We hope that readers will find this a generally beneficial influence, as the VLA is an environment in which many boundaries of image processing techniques in radio astronomy are being pushed back, and as many of you will be reading these notes because you intend to use the VLA for your own research. Those of you with broader interests will, we hope, find the VLA-specific sections of these lectures easy to identify and to skip over, if you wish.

The lectures do not appear here exactly as given. These written versions were prepared after the lecturers had reviewed comments from the course participants and from other NRAO staff. 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. We have also standardized notation and rearranged material where we felt that this made the course as a whole more coherent.

2. Subject matter.

The first lecture, by Barry Clark, develops the fundamental principles and equations of synthesis imaging, with particular attention to the assumptions which underlie them. The second, by Richard Thompson and the third, by Larry D'Addario, discuss the practical implementations of instruments to image the radio sky based on these fundamental principles. These lectures are written from the standpoint of the engineers who build the instruments, and are essential reading for anyone wishing to understand how the design of a synthesis array interacts with the quality and credibility of the data which are obtained from it. The fourth lecture, by Carl Bignell and Rick Perley, reviews the many instrumental, atmospheric and environmental effects that must be calibrated or compensated before the data from a synthesis array are ready to be passed to the imaging procedures.

The fifth lecture, by Dick Sramek and Fred Schwab, describes the primary computational steps involved in making an image from the data collected by a synthesis array. It forms the basis for all of the subsequent discussions.

The sixth lecture, by Pat Crane and Peter Napier, examines the factors that affect the sensitivity of synthesis images to various kinds of structure, with particular attention to the calculations relevant for the VLA.

The next five lectures examine the imperfections of the "dirty" images made by the straightforward techniques of Lecture 5, and discuss the battery of methods that radio astronomers bring to bear on diagnosing and suppressing these imperfections. The seventh lecture, by Tim Cornwell, reviews the strengths and weaknesses of deconvolution algorithms currently in use in astronomy. The eighth, by Bill Cotton, describes procedures for dealing with the problems faced when some of the simplifying assumptions made in Lecture 1 are violated, and for reducing the computing requirements of some difficult imaging cases. The ninth lecture, by Tim Cornwell, treats the powerful technique known as "self-calibration" whereby data obtained from the source itself are used to calibrate its own image. The tenth lecture, by Ron Ekers, describes some common image defects, how to recognize what causes them, and how to eliminate or reduce them. The eleventh lecture, by Rick Perley, discusses the techniques that are now available for extremely high-fidelity imaging in the presence of initially corrupted data from synthesis arrays.

The next two lectures treat important special topics in radio interferometry. The twelfth, by Jacqueline van Gorkom and Ron Ekers, discusses problems specific to spectral line synthesis; the thirteenth, by Craig Walker, describes the special features of synthesis imaging with arrays of antennas that are not physically connected—very long baseline interferometry (VLBI).

The fourteenth and fifteenth lectures, by Ed Fomalont and Arnold Rots, treat different aspects of image analysis; that is, the extraction of useful information from the final images once they have been computed.

The sixteenth and final lecture, by Alan Bridle, describes an orderly approach to using the information from the previous lectures when planning and executing observing programs at the VLA.

3. Terminology and Notation.

Some of the terminology used in these lectures departs from the established traditions of radio astronomy—in ways that we hope will reduce confusion for newcomers to this field. The process of image construction in radio astronomy has been known for decades as *mapping*, not *imaging*, as here. Generations of radio astronomers have regarded isophotal maps (contour maps) as the prime display of their data, and have adopted the term *map* because of the analogy with topographical mapping. In most other fields of research, the distribution of intensities across a field of view is described as an *image*, however, and we were persuaded to use the more common terminology throughout this course despite the radio astronomy tradition.

The distinctions between images made by radio telescopes and by telescopes operating at other wavelengths are minor compared with the impediment to understanding that comes from using different terminology in different applications. We have therefore used "image" in most places that "map" would occur normally in the radio astronomy literature, with a few exceptions, e.g., when describing contour displays, or when discussing techniques such as *fringe rate mapping*, which determine the layout of a source without imaging it.

We have also generally avoided the traditional term "aperture synthesis" for the imaging technique, as most modern synthesis telescopes have no equivalent aperture. The common term "Earth rotation synthesis" also seems unnecessarily restrictive, as many of the

principles described here can be employed without making use of Earth rotation to generate the sampling pattern. We have therefore adopted the brief term "synthesis", which may be thought of as a shorthand for "Fourier synthesis", throughout.

Finally, we found no *a priori* standard among the lecturers about the sign of the phase term in the Fourier transform integral, or about the direction termed the "forward" or the "inverse" Fourier transform. To make the course internally consistent (and so, we hope, to minimize confusion), we converted the notation and language of all lectures to the convention that was adopted initially by the majority. This defines the *forward* transform direction as that with the positive sign of the phase, which in these notes is the transform from the spatial frequency domain to the image domain. This convention is common in mathematics texts, but the reader should note that it is the inverse of the convention most commonly found in the engineering literature. The reader will therefore encounter the opposite convention in some of the referenced literature, but we hope the internal consistency of these lectures will avoid confusion.

4. Some NRAO Lore.

There are references to NRAO internal publications and to NRAO software throughout this course. This is inevitable, given that many important topics covered by the Summer School are not published in the regular literature or in textbooks. These references will also be important further reading for those of you who eventually pursue an interest in synthesis imaging to the point of making observations with the VLA or VLBA. Copies of memoranda in the various VLA technical and scientific series are available on request from Alison Patrick at NRAO, P.O. Box O, Socorro, NM 87801, USA.

The reader will also find frequent references to software in the 'AIPS' package. This is a software system for calibration, imaging and analysis of astronomical data that is tailored to the needs of synthesis imaging (though parts of it are much more general). 'AIPS' stands for *A*stronomical *I*mage *P*rocessing *S*ystem, and both the software and documentation describing it can be obtained on request from Nancy Wiener, NRAO, Edgemont Road, Charlottesville, VA 22903-2475, USA.

RICHARD A. PERLEY
FREDERIC R. SCHWAB
ALAN H. BRIDLE

OPENING REMARKS

I would like to welcome you to the Macey Center on the campus of the New Mexico Institute of Technology in Socorro for our second summer course on Radio Astronomical Imaging with Synthesis Telescopes.

Why have another course on "Synthesis Imaging"? Synthesis radio telescopes are playing an increasingly important role in radio astronomy. The VLA has been in full operation since 1981, two major new synthesis telescopes are under construction—the Australia Telescope and the VLBA—and further synthesis telescopes are being planned (e.g., QUASAT and NRAO's millimeter wavelength array). It is our intention to keep this discussion sufficiently general to apply to any of these instruments, although almost all examples will be taken from the VLA. Synthesis telescopes are unusually powerful and are very flexible, but they are so different from the conventional telescopes which form images directly in their focal planes that they may at first seem more difficult to understand. However, once the underlying principles are clear you will be able to exploit their flexibility.

The VLA is a national facility. National observatories make it possible to concentrate resources into single large instruments, but they also deprive many of you of the educational experience of building and using your own telescope. One aim of this course is to try to compensate for the imbalance between the very large number of users who have little opportunity to obtain hands-on experience and the relatively small number of our staff who often feel they get more hands-on experience than they need.

Especially in the area of image processing, there have been many new developments which have not been included in any existing textbook or course material on radio astronomy. Examples are the enhancements of images through deconvolution algorithms such as 'CLEAN' and MEM, the removal of atmospheric blurring by the application of antenna-based self-calibration, and the techniques used for the production and processing of spectral line synthesis data. These topics will be covered during this course, and the lecture notes will be made available to you.

This is also a rich field for cross-fertilization between disciplines. Many of you and many of the users of the VLA are not radio astronomers, and as soon as you can penetrate the barrier of the jargon of this subfield you will find that many of the underlying principles apply in a wide range of situations. Obvious examples include: optical interferometry, adaptive optics, and indirect imaging in medicine, radar, seismology and other fields.

The vast majority of observations made with the VLA use it as an analytic tool to observe known phenomena and to make specific measurements relating to a hypothesis about the object or class of objects under study. The information we cover in this course will provide the background needed to plan for this type of observation. We will describe techniques which will enable you to extract all the useful information in your data and to get it into a form suitable for interpretation, and we will try to indicate how to do this in a reasonable amount of time. Fundamental to an observational science is the quality of the results produced. The NRAO engineering staff will do everything they can to make the instrument work reliably, however it is the observers who are ultimately responsible for the results which they publish. A basic knowledge of the instrument is important in order to recognize erroneous results and to have confidence in the integrity of the final product.

It is well known that major discoveries, especially in radio astronomy, have been driven by instrumental developments. The VLA represents such a major instrumental development—with sensitivity, resolution, speed and sky coverage far greater than any previous radio telescope. New and exciting discoveries should be possible with such a telescope,

but in order to be able to recognize the unexpected you must understand the instrument well. The most likely reason for an unexpected result is an error of some kind. If you throw out everything unusual, assuming it is erroneous, you may miss something important. On the other hand, if you spend a great deal of time investigating every possible error you will never get much done. With an understanding of the basic principles by which these telescopes work you will be in a much better position to discriminate against errors and to recognize the unexpected. In his book *Cosmic Discovery* (Basic Books, New York, 1981), Martin Harwit argues that national facilities are unlikely to make major discoveries. Through courses such as these we hope to show that this need not be true.

RONALD D. EKERS

1. Introduction and Basic Theory

BARRY G. CLARK

1. INTRODUCTION

It is appropriate for this specialized workshop 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 an appreciable part of the workshop 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 workshop. 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 $\mathbf{r}$.

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

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

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

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

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

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

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

$$I^D_\nu = I_\nu * B \,, \tag{1–11}$$

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 7 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:

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

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).

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

The theoretical basis of interferometry has been discussed in the previous Lecture, and here we are concerned with some practical effects that modify the response. Other discussions of the interferometer response can be found in Swenson and Mathur (1968), Fomalont (1973), Fomalont and Wright (1974), Meeks (1976), and Christiansen and Högbom (1985); a detailed and extensive review is given by Thompson, Moran and Swenson (1986). Synthesis arrays, which produce images by Fourier synthesis from measurements of complex visibility, can be analyzed as ensembles of two-element interferometers. Many of the effects can therefore be understood from a discussion of the properties of a two-element instrument.

1. RESPONSE OF AN INTERFEROMETER

A simplified block diagram of 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 $\Delta\nu$. 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 Fourier components of 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 fringe pattern. We may assume that these oscillations are sufficiently slow that the fringes are not significantly attenuated by the averaging (an expression for the fringe frequency is given in Section 8). In contrast, the component at frequency 2ν generated in the multiplication is effectively filtered out. Note that the term $v_1 v_2$, which represents the fringe amplitude, is proportional to the received power.

We now express the interferometer output in terms of the radio brightness integrated over the sky. Let $I(\mathbf{s})$ represent the radio brightness in the direction of unit vector $\mathbf{s}$ at frequency ν. The brightness is also sometimes referred to as intensity and is measured in $\mathrm{W\,m^{-2}\,Hz^{-1}\,sr^{-1}}$. Note that I refers to the component of the radiation that is matched to the polarization of the antennas, which we assume are identically polarized. The way in

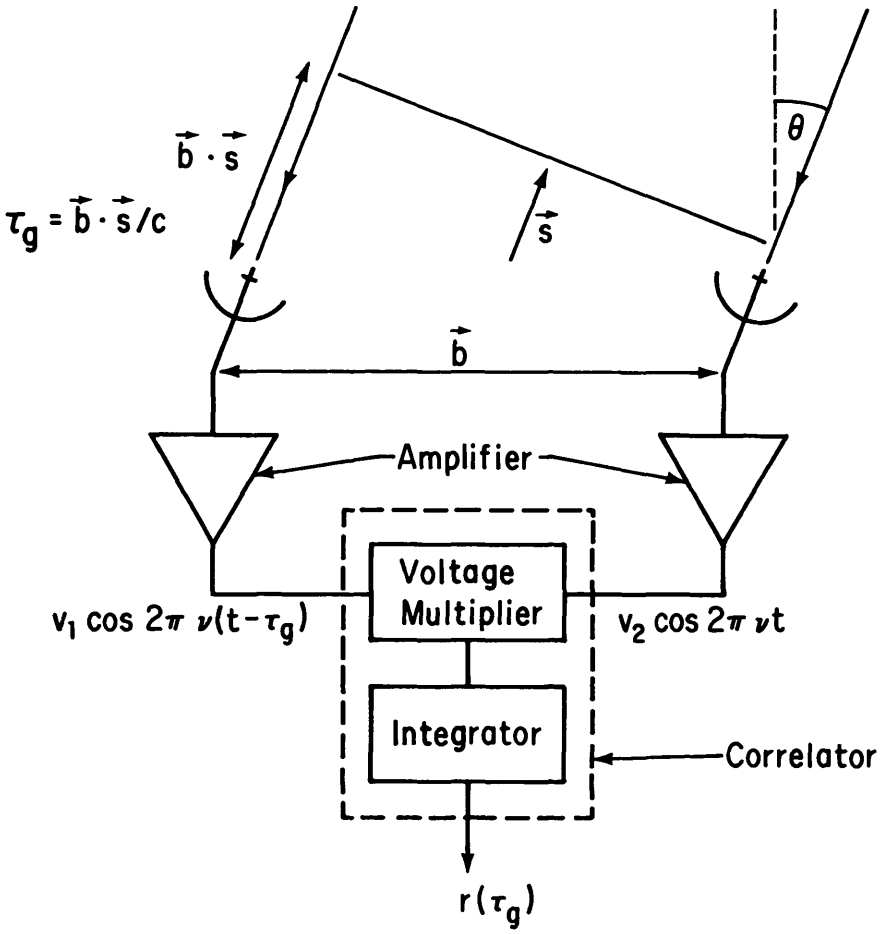


Figure 5–1. Simplified schematic diagram of a two-element interferometer.

which the antenna polarization is varied to explore the total radiation field is considered in Lecture 4. 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

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

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 8). 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

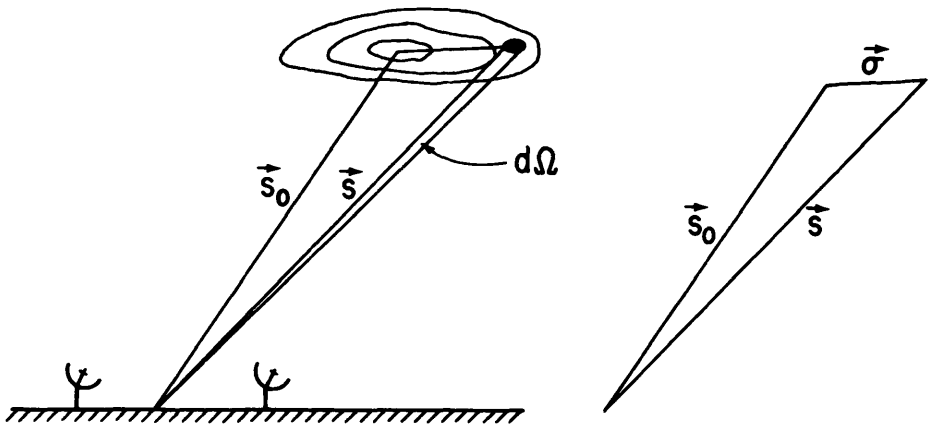


Figure 2–2. Position vectors used in deriving the interferometer response to a source. The source is represented by the contours of radio brightness $I(\mathbf{s})$ on the sky.

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 nominal source position on which the synthesized field of view is to be centered. We represent this position by the vector $\mathbf{s}_0$, as shown in Figure 2–2, and write $\mathbf{s} = \mathbf{s}_0 + \sigma$. From Equation 2–5 we then obtain

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

The complex visibility of the source is defined as

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

where $\mathcal{A}(\sigma) \equiv A(\sigma)/A_0$ 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 $\mathcal{A}(\sigma)I(\sigma)$. By separating the real and imaginary parts of V in Equation 2–7 we obtain

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

and

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

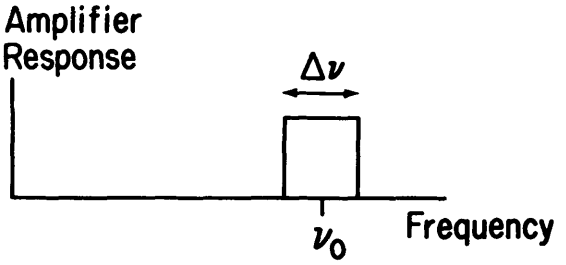


Figure 2–3. Idealised rectangular response of the receiving system.

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.

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 only observed 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

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

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.

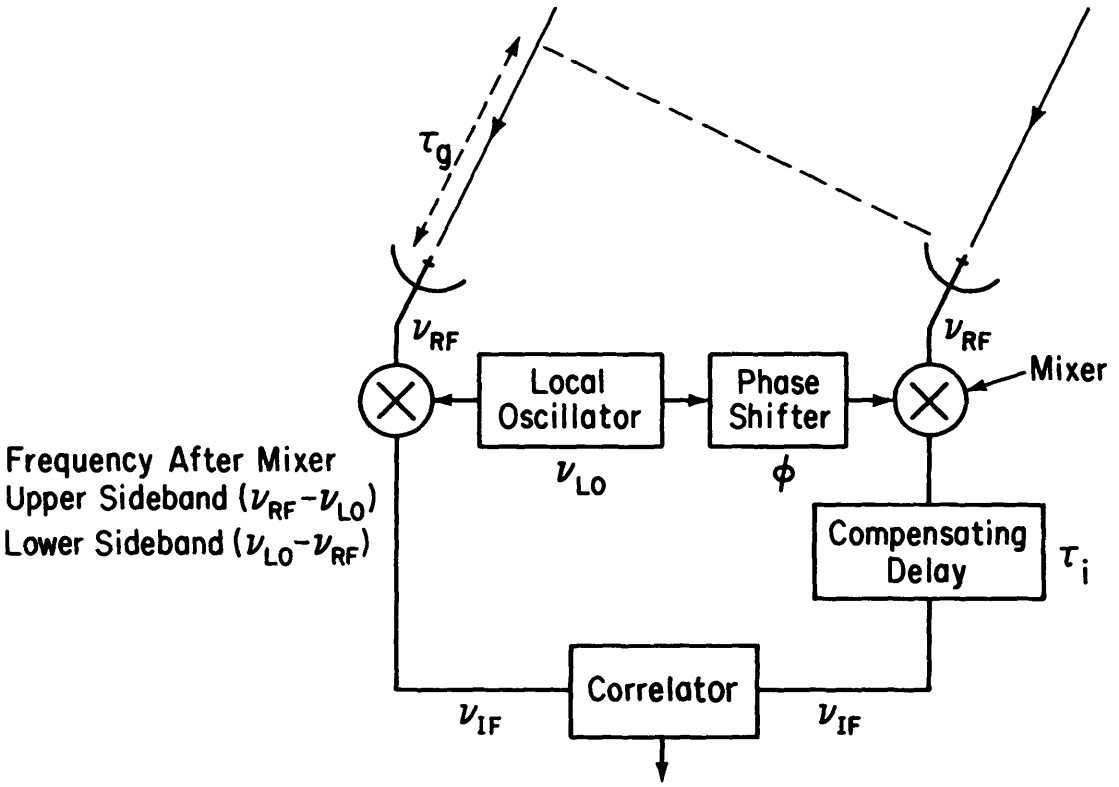


Figure 2–4. Simplified schematic diagram of an interferometer system incorporating frequency conversion and an instrumental time delay to compensate for τ_g.

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 frequency ν_{RF} with a local oscillator at frequency ν_{LO} is also included. Practical receiving systems incorporate frequency conversion because it is technically more convenient to perform such functions as amplification, filtering, delaying, and cross-correlating of the signals at an intermediate frequency that is lower than ν_{RF} and remains fixed when the observing frequency is changed. The signals at the frequencies ν_{RF} and ν_{LO} are combined in a mixer which contains a non-linear element (often a diode) in which combinations of the two frequencies are formed. The intermediate frequency ν_{IF} is related to the mixer input frequencies by

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

Thus the mixer input is 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

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

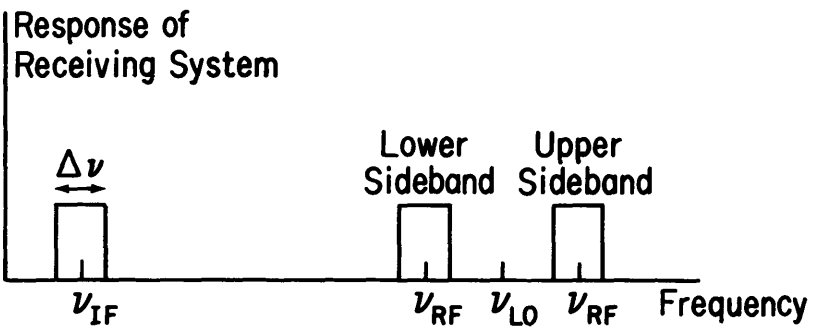


Figure 2–5. Relationship of RF, IF, and LO frequencies.

where ϕ_{LO} is the difference in the phase of the local oscillator signal at the two mixers, and τ_i is the instrumental delay that compensates for τ_g. The upper-sideband response of the interferometer is obtained by replacing the argument of the cosine function in Equation 2–10 by $\phi_1 - \phi_2 - \phi_V$:

$$r_u = A_0 \Delta\nu |V| \cos[2\pi(\nu_{LO}\tau_g + \nu_{IF}\Delta\tau) - \phi_V - \phi_{LO}] . \qquad (2\text{–}16)$$

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

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

whence

$$r_l = A_0 \Delta\nu |V| \cos[2\pi(\nu_{LO}\tau_g - \nu_{IF}\Delta\tau) - \phi_V - \phi_{LO}] . \qquad (2\text{–}18)$$

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 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 the two cases.

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:

$$r_d = r_u + r_l = 2\Delta\nu A_0 |V| \cos(2\pi\nu_{LO}\tau_g - \phi_V - \phi_{LO}) \cos(2\pi\nu_{IF}\Delta\tau) . \qquad (2\text{–}19)$$

Note that the delay-tracking error $\Delta\tau$ here appears in a separate cosine term that modulates the amplitude rather than the phase of the cosine fringe term. 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 more detailed discussion see Thompson, Moran and Swenson (1986).

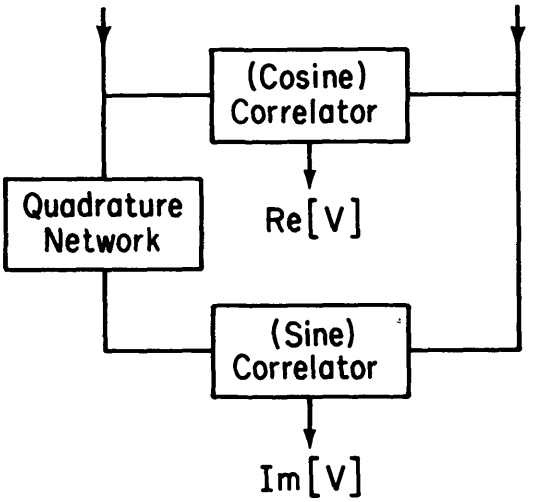


Figure 2–6. Complex correlator system. The quadrature network introduces a $\pi/2$ phase shift: a signal of the form $\cos 2\pi\nu t$ at its input becomes $\cos(2\pi\nu t - \pi/2)$ at the output.

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, 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 ϕ_{LO} so that $(2\pi\nu_{LO}\tau_g - \phi_{LO})$ 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 ϕ_{LO} is usually controlled by the same computer that regulates the delay tracking, is variously referred to as *fringe rotation* or *fringe stopping.*

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 reference position). This voltage does not provide a measure of the amplitude and phase of the fringes. 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 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 6 for an analysis of signal-to-noise ratios.

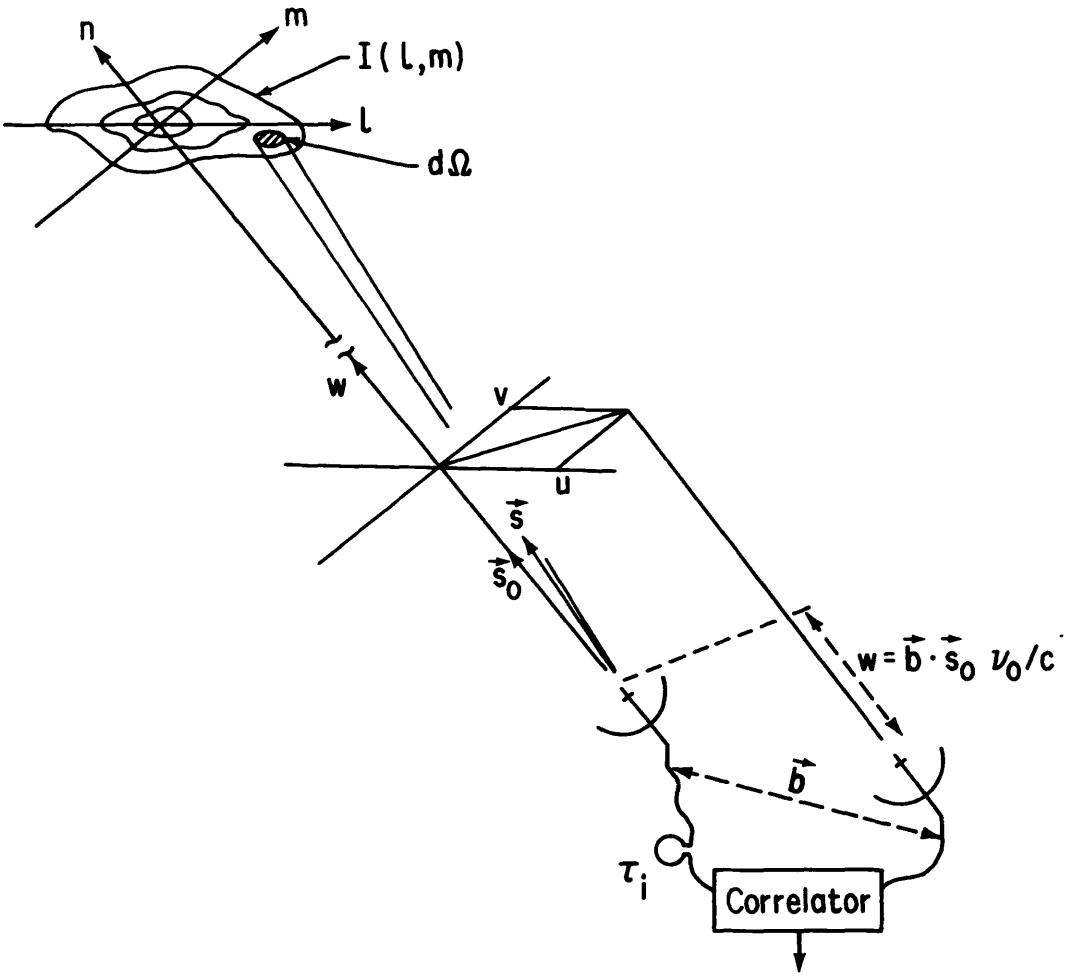


Figure 2–7. The (u, v, w) and (l, m, n) coordinate systems used to express the interferometer baselines and the source brightness distribution, respectively.

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 sign of the multiplier output in the correlator before the data are averaged. To prevent the loss of the wanted output from the radio source, the phase of the signal at one antenna of the interferometer pair is synchronously reversed by switching an extra half-wavelength of transmission line into the signal path, or, more commonly, reversing the phase of a local oscillator signal. Reversing the phase of the signal at one antenna has the effect of reversing the sign of the wanted correlator output, and this reversal cancels the reversal applied at the correlator output. 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).

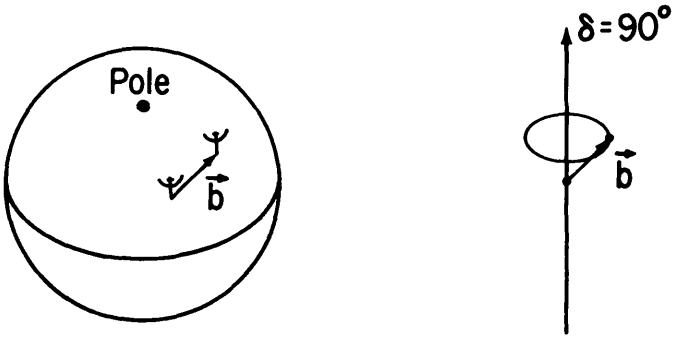


Figure 2–8. As the earth rotates, the baseline vector **b**, which represents the spacing of the two antennas, traces out a circular locus in a plane normal to the direction of declination (δ) equal to 90°. If the antennas are in an east–west line on the earth, then the vector **b** is normal to the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

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 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 map. Note that u and v are measured in wavelengths at the center frequency ν_0, and in directions towards the east and north 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 map 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

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

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

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 form occurs when $w = 0$, and the conditions required can be understood by considering the way in which the earth's rotation carries the antennas through space. It should be evident from Figure 2–8 that the rotation causes the tip of the baseline vector to trace out a circle concentric with the earth's rotation axis. The rising and setting of a point on the sky usually limit the range over which V can be measured to an arc of the circle. In general, for a two-dimensional array of antennas on the surface of the earth, the circular

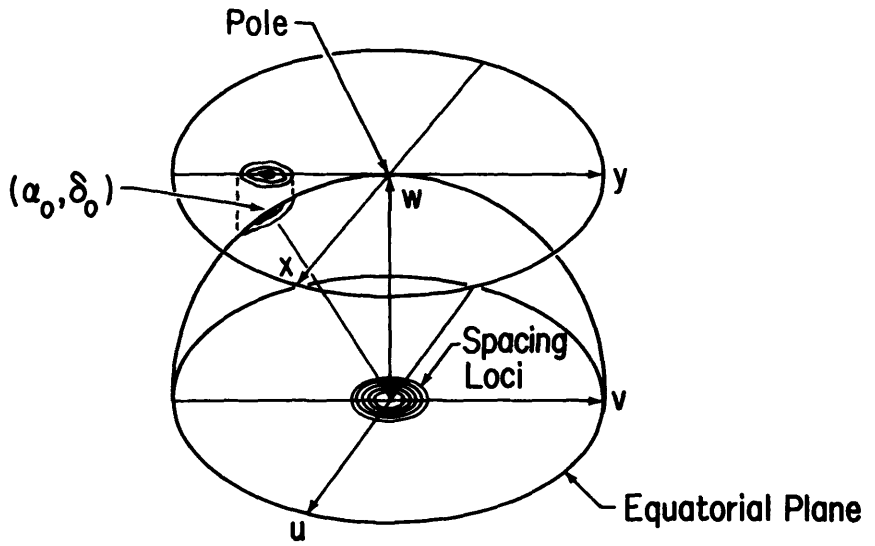


Figure 2–9. Celestial hemisphere showing the projection of a source at (α_0, δ_0) onto the tangent plane at the pole. The spacing-vector loci are for an array with east–west baselines, and lie in a plane parallel to the earth's equator. The direction of the w-axis is here chosen to be that of the pole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

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-vector loci are coplanar. Then, if we choose the w-axis to lie in the direction of the celestial pole, so that $w = 0$, Equation 2–21 becomes

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

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

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

Equation 2–23 can be applied to all parts of the hemisphere shown in Figure 2–9. Usually we want to map 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 v-axis as in Figure 2–9 so that l is small within the region of interest and is closely 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 map is compressed in the m direction by a factor $\sin\delta$. The coordinate transformation

$$\begin{aligned} l' &= l\,, \\ m' &= (m - \cos\delta_0)/\sin\delta_0\,, \end{aligned} \qquad (2\text{–}24)$$

results in a map in (l', m') in which the origin is at (α_0, δ_0), and the scale factor in the m' direction is correct at that point. However, there is still a progressive change in scale in the m' direction across a map. This can be ignored in small field maps, and in large field maps the data, which are usually computed for points at uniform increments in l and m, can be interpolated into a more desirable coordinate system (Rots 1974).

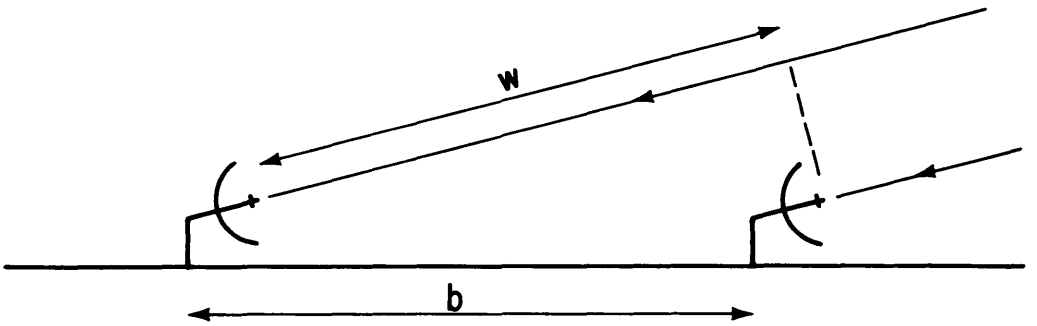


Figure 2–10. Comparison of the w-component and the antenna spacing when the direction of the source is close to that of the baseline.

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 return to Equation 2–21. The usual way in which Equation 2–21 is used for non-east–west baselines depends upon $|l|$ and $|m|$ being small enough that we can write

$$\left(\sqrt{1 - l^2 - m^2} - 1\right) w \simeq -\tfrac{1}{2}(l^2 + m^2)w \simeq 0\,. \tag{2–25}$$

Then Equation 2–21 becomes

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

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 2–26 we can write

$$\mathcal{A}(l,m) I(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(u,v) e^{2\pi i(ul+vm)}\, du\, dv\,. \tag{2–27}$$

For arrays in which the baselines do not remain coplanar as the earth rotates, the approximation in Equation 2–25 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 2–25 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 mapped 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 θ_{HPBW}, we can write

$$\frac{1}{\theta_{\mathrm{HPBW}}} \simeq \frac{b_{\max}}{\lambda} \simeq w_{\max}\,, \tag{2–28}$$

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

$$\frac{\pi\theta_{\mathrm{F}}^2}{4\theta_{\mathrm{HPBW}}}\,. \tag{2–29}$$

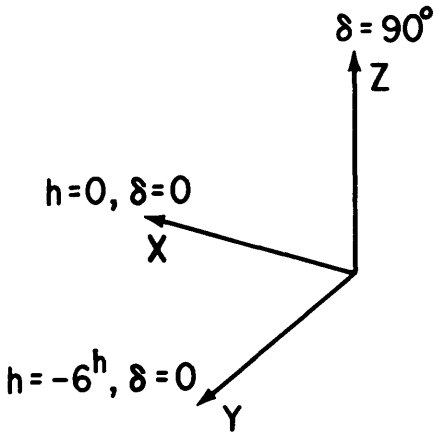


Figure 2–11. Coordinate system for specification of baseline parameters. X is the direction of the meridian at the celestial equator, Y is toward the east, and Z toward the north celestial pole.

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

$$\theta_F < \frac{1}{3}\sqrt{\theta_{HPBW}}, \tag{2–30}$$

where the two angles are measured in radians. Then, for example, if $\theta_{HPBW} = 1''$, $\theta_F < 2\overset{'}{.}5$. For fields of greater width than allowed by Equation 2–30 there are ways of avoiding or ameliorating the distortion introduced by the phase errors—see Lecture 8.

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, $(h = -6^h, \delta = 0)$ 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

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} \sin H_0 & \cos H_0 & 0 \\ -\sin\delta_0 \cos H_0 & \sin\delta_0 \sin H_0 & \cos\delta_0 \\ \cos\delta_0 \cos H_0 & -\cos\delta_0 \sin H_0 & \sin\delta_0 \end{pmatrix} \begin{pmatrix} L_X \\ L_Y \\ L_Z \end{pmatrix}, \tag{2–31}$$

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. By eliminating H_0 from the expressions for u and v we obtain the equation of an ellipse in the u-v plane:

$$u^2 + \left(\frac{v - (L_Z/\lambda)\cos\delta_0}{\sin\delta_0}\right)^2 = \frac{L_X^2 + L_Y^2}{\lambda^2}. \tag{2–32}$$

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

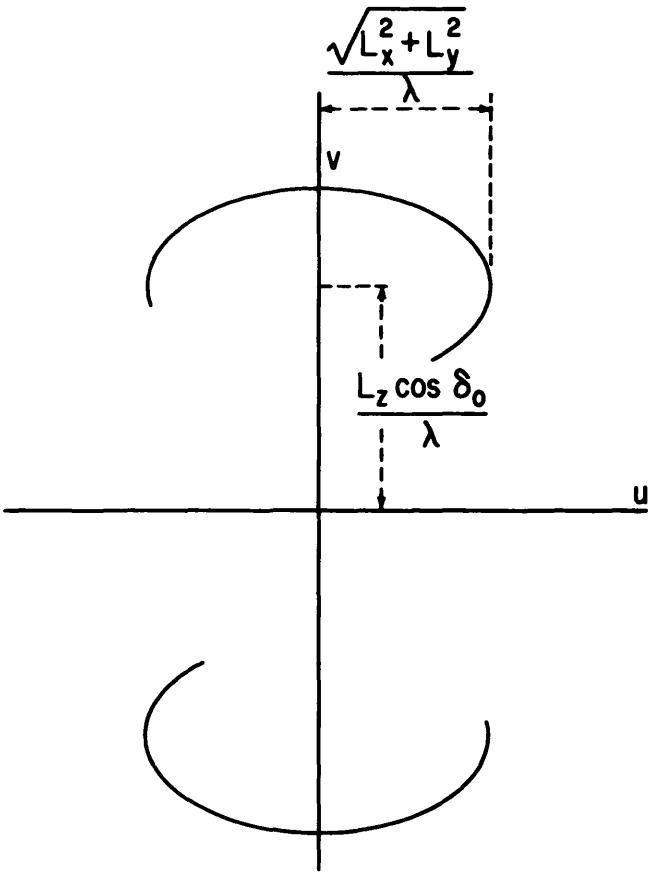


Figure 2–12. Elliptical loci representing the projection of the baseline vector onto the u-v plane as a source is tracked across the sky. The lower curve corresponds to a reversal of the direction of the baseline vector, and represents the points for which the visibility is the complex conjugate of that 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.

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*, $W(u, v)$, 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 mapping an interferometer or array is used to provide values of the complex visibility as a function of u and v, from which a brightness distribution can be derived.

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

In addition to the measurement of complex visibility, two other characteristics of the interferometer output can be used to determine astronomical data. These are principally of importance in VLBI, in which it is 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 response 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 2–31. 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 2–31 by differentiation. A useful expression for the fringe frequency ν_F is

$$\nu_F = \frac{dw}{dt} = -\omega_e u \cos\delta\,, \tag{2–33}$$

where $\omega_e = dH_0/dt$ is the angular rotation velocity of the earth. Thus 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 mapping 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

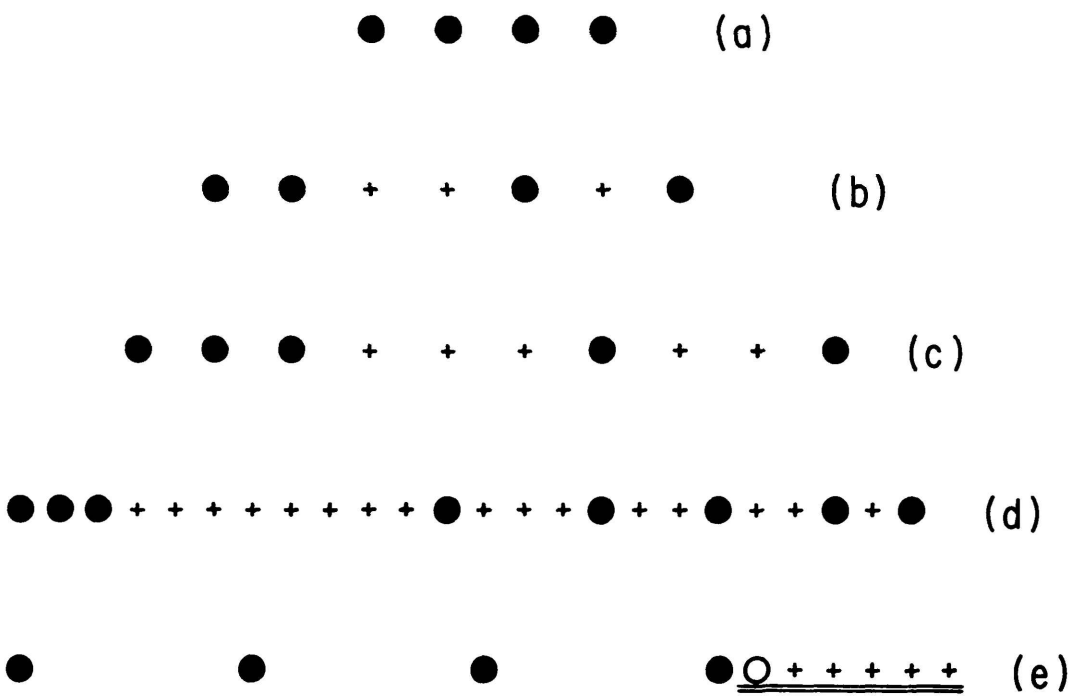


Figure 2–13. Examples of several types of linear arrays of antennas. **(a)** Uniform-spacing array, **(b)** non-redundant array (Arsac 1955), **(c)** minimum-redundancy array (Bracewell 1966), **(d)** minimum-redundancy array (Moffet 1968), and **(e)** array with movable element represented by the open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

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 make use of this principle: these include the One–Mile and Five–Kilometer arrays at Cambridge (Ryle 1962, 1972) and the Westerbork Synthesis Radio Telescope (Högbom and Brouw 1974). 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 well 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 nth antenna on each arm, counting outwards from the center, is

proportional to $n^{1.716}$. With this power-law design, no two spacings on any arm are equal. The array is rotated through 5° from the position of north–south symmetry to avoid exact east–west baselines, which would otherwise occur between antennas on the two southern arms. At declination 0° the u-v components for all east–west baselines become coincident with the u-axis. Thus the power-law spacing and the rotation are features of the VLA design that reduce redundancy in the coverage of the u-v plane.

The same considerations of uniformity of sampling in the u-v plane also apply to arrays for imaging by VLBI. The main practical difference is that since the antennas are not directly interconnected, except by telephone lines for monitor and control purposes, there is no advantage to any particular geometric pattern. Thus, after the u-v coverage, the main concern is the choice of sites for freedom from interference, low water vapor in the atmosphere, convenience for service, etc. The proposed locations for antennas in the Very Long Baseline Array (VLBA), 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).

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 point. 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.

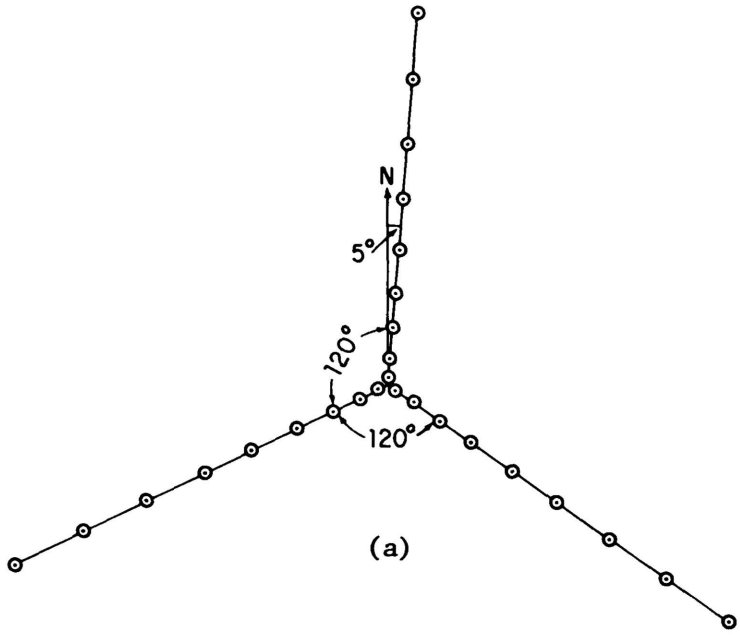
In observing continuum radiation we are interested in the mean brightness over the bandwidth $\Delta\nu$, and the visibility data are processed as though they were all observed at the center frequency ν_0 indicated in Figure 2–17a. In particular, the spatial frequency coordinates in the u-v plane are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

$$(u_0, v_0) = \left(\frac{\nu_0}{\nu}u, \frac{\nu_0}{\nu}v\right) . \qquad (2\text{–}34)$$

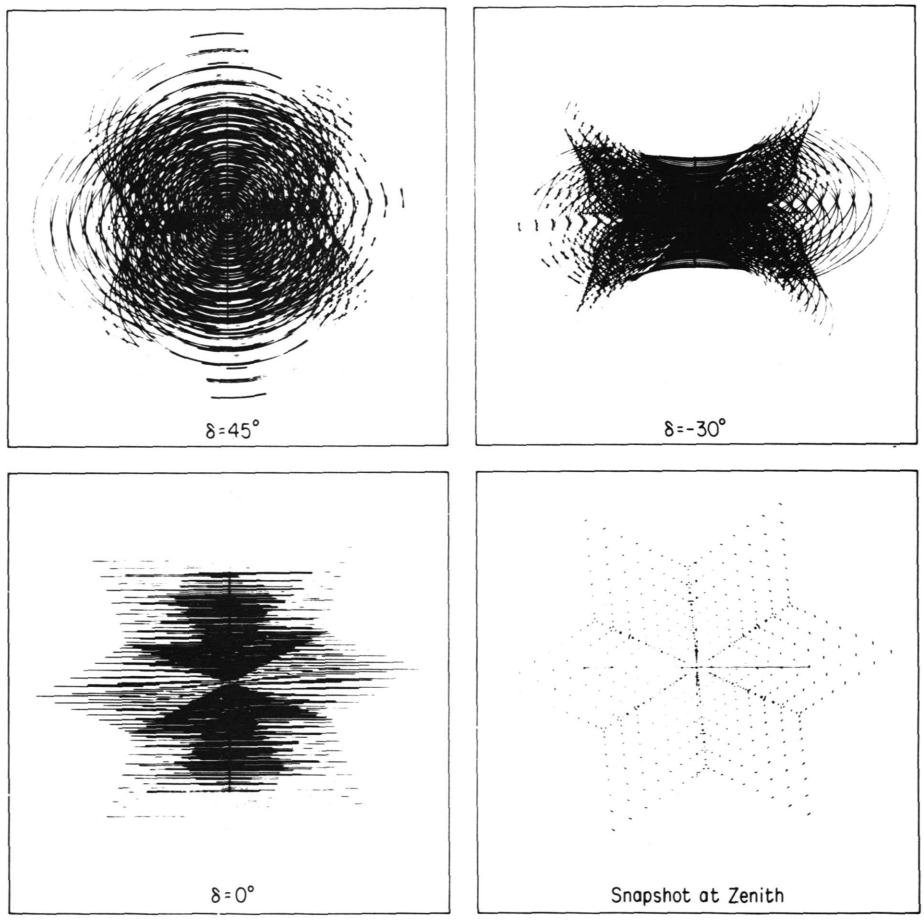
Now consider the visibility that corresponds to a small band of frequencies centered on ν as in Figure 2–17a. This band contributes a component of brightness I to the synthesized image which is related to the corresponding visibility by

$$V(u, v) \rightleftharpoons I(l, m) , \qquad (2\text{–}35)$$

where the symbol $\rightleftharpoons$ indicates that the two functions constitute a Fourier transform pair, and we have here omitted the functions $A(l, m)$ and $1/\sqrt{1 - l^2 - m^2}$ which are usually close to unity. Note that the processes of correlation and Fourier transformation are linear, and that they allow us to consider the synthesized image as the sum of a series of contributions from different parts of the frequency passband. In the derivation of the radio image we assign to V values u_0 and v_0 which are the true values multiplied by ν_0/ν (Eq. 2–34). The

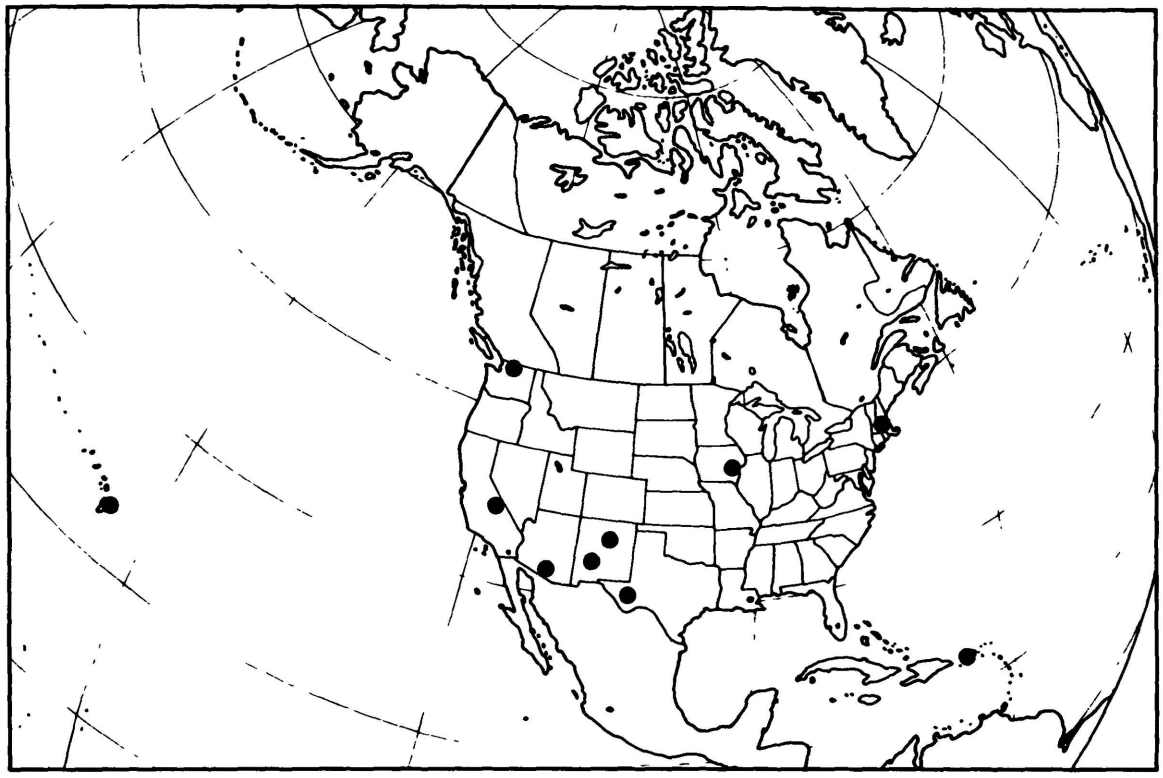


(a)

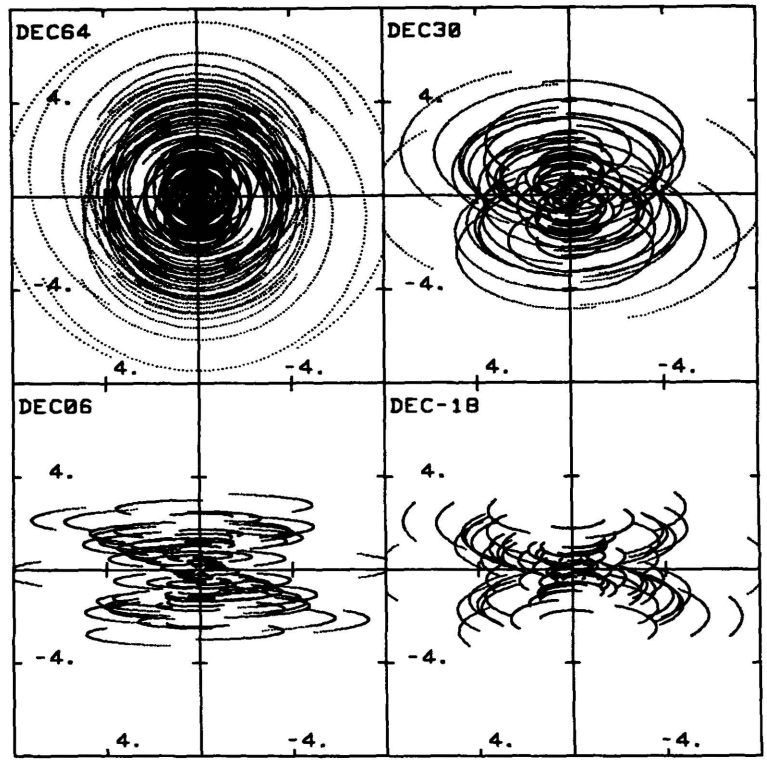


(b)

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



(a)



(b)

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).

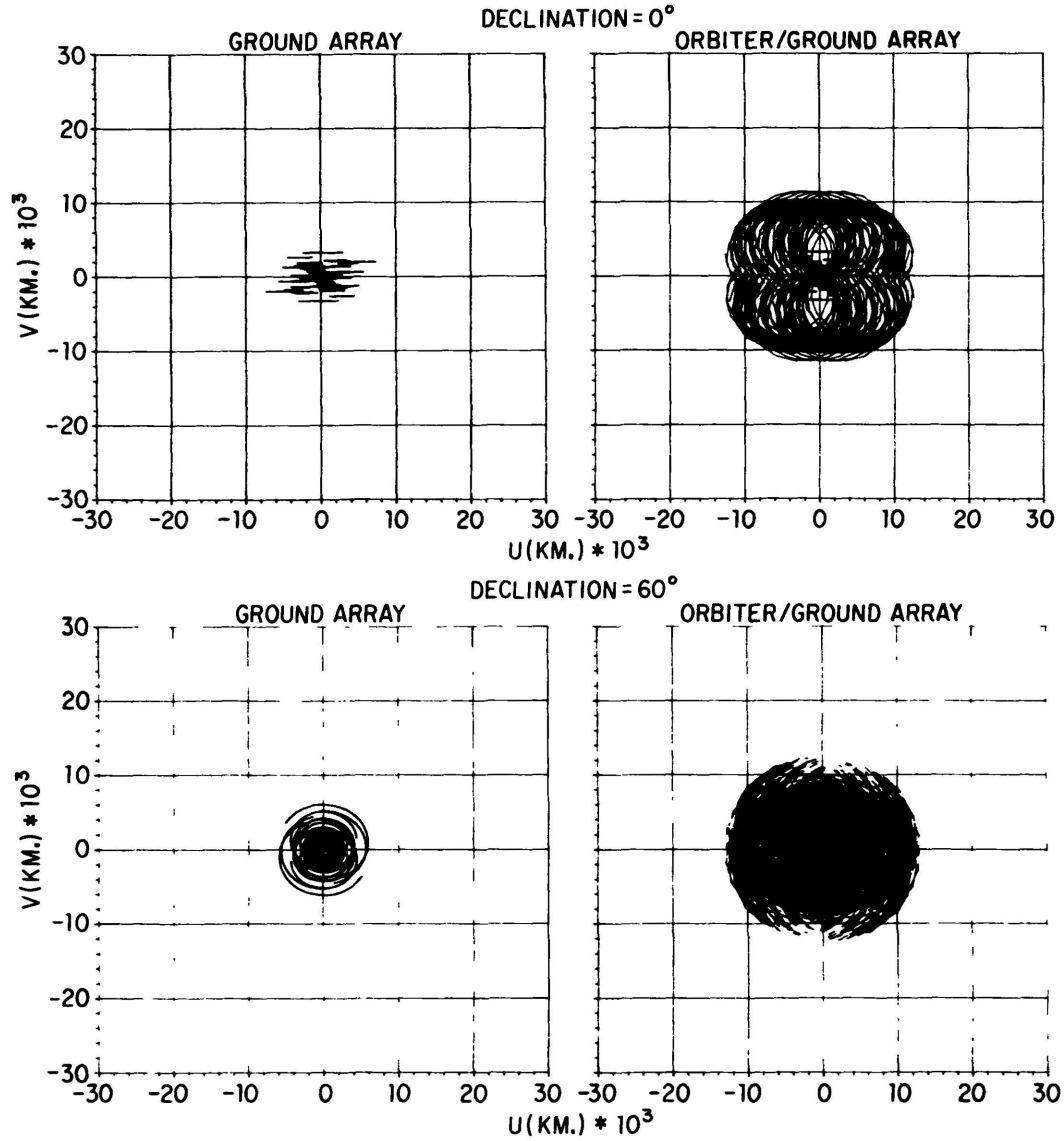


Figure 2–16. Examples of the u-v coverage obtained using a VLBI array similar to that of Figure 2–15a and one additional antenna in low earth orbit. From Preston *et al.* (1983).

effect in the image can be obtained from the similarity theorem of Fourier transforms (e.g., Bracewell 1978), using which we can write

$$V\left(\frac{\nu_0}{\nu}u, \frac{\nu_0}{\nu}v\right) \rightleftharpoons \left(\frac{\nu}{\nu_0}\right)^2 I\left(\frac{\nu}{\nu_0}l, \frac{\nu}{\nu_0}m\right) . \tag{2–36}$$

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

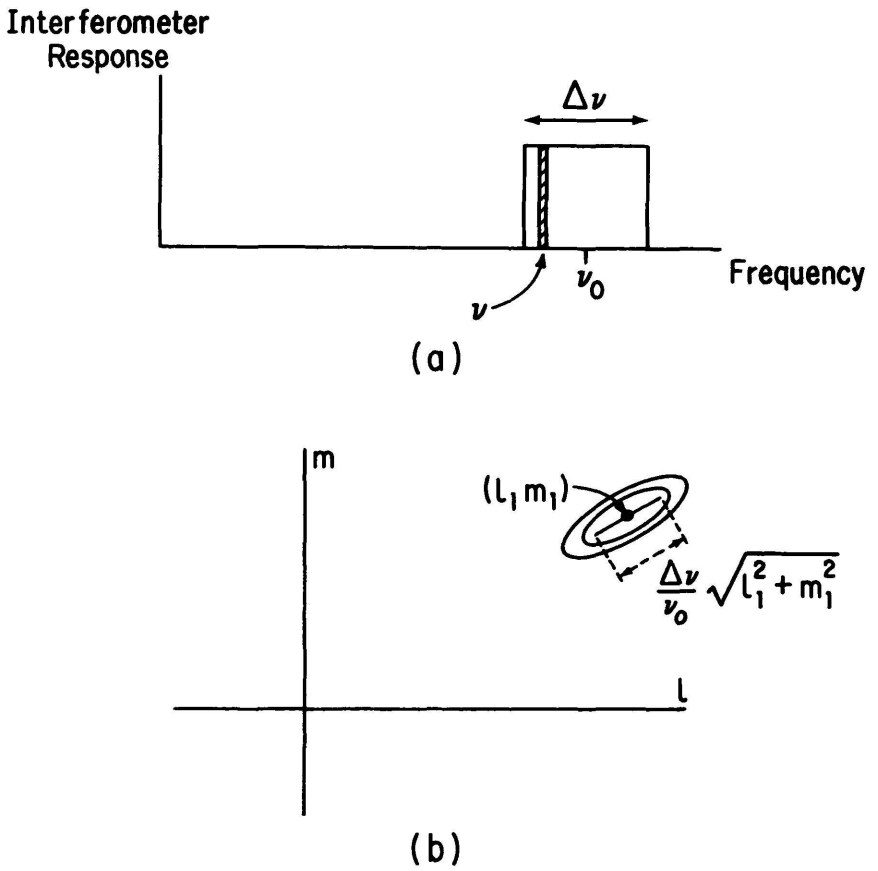


Figure 2–17. **(a)** Idealised rectangular response showing center frequency ν_0 and a narrow band at frequency ν. **(b)** The radial smearing of a point source at (l_1, m_1) in the synthesized image.

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 suppress unwanted sidelobes. Thus the use of a multi-channel system is a desirable technique in broadband image synthesis, but in practice is restricted by the increase in computing required to accommodate n times as many visibility data as in the corresponding continuum observation.

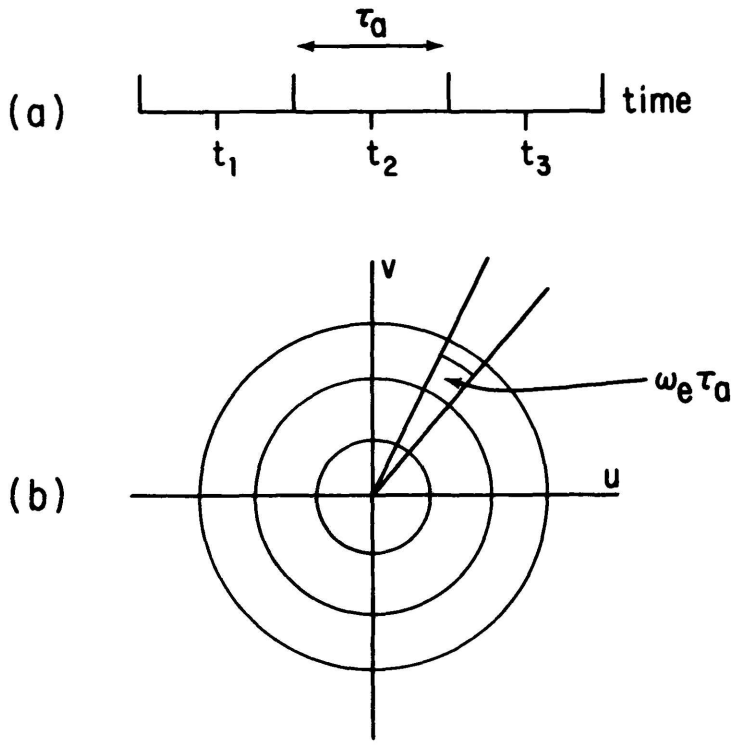


Figure 2–18. (a) Consecutive time intervals of duration τ_a over which the visibility is averaged. (b) Circular loci in the u-v plane which result from the continuous observation of a source close to the celestial pole. In a time interval τ_a, the baseline vectors which generate the loci move through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

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

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 Univ. 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*, G. L. Verschuur and K. I. Kellermann, Eds., Springer–Verlag, New York, pp. 256–290.

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.

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

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.

Rots, A. H. (1974), *Distribution and Kinematics of Neutral Hydrogen in the Spiral Galaxy M81*, Ph. D. Thesis, University of Groningen, The Netherlands.

Rowson, B. (1963), "High Resolution Observations With a Tracking Radio Interferometer", *Mon. Not. Royal 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.*, **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, 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. 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 which 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; Lecture 12 will consider these special problems in greater depth.

2. CORRELATORS IN GENERAL

The two preceding Lectures dealt mainly with the correlation of quasi-monochromatic signals. We would now like to generalize to the case of wide bandwidth signals; this leads naturally to an understanding of spectroscopic cross correlation. Sometimes one wishes to observe over a signal bandwidth that 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

$$x_{ij}(\tau) = \langle v_i(t) v_j(t+\tau) \rangle . \tag{3-1}$$

This is a real function of delay τ, and can be estimated by the simple correlator of Figure 3–1. 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. 3–2). That is, we can write

$$x_{ij}(\tau) = x_R \cos 2\pi\nu_0(\tau - \tau_0) + x_I \sin 2\pi\nu_0(\tau - \tau_0) , \tag{3-2}$$

for τ in the vicinity of reference delay τ_0. 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

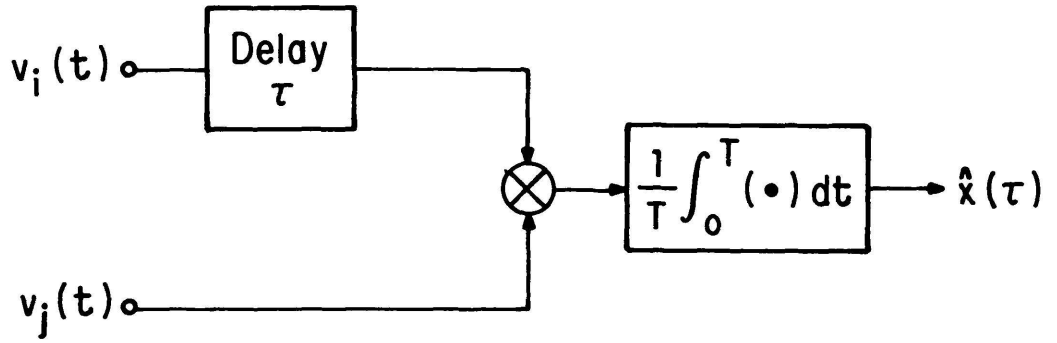


Figure 3–1. A simple (real) correlator.

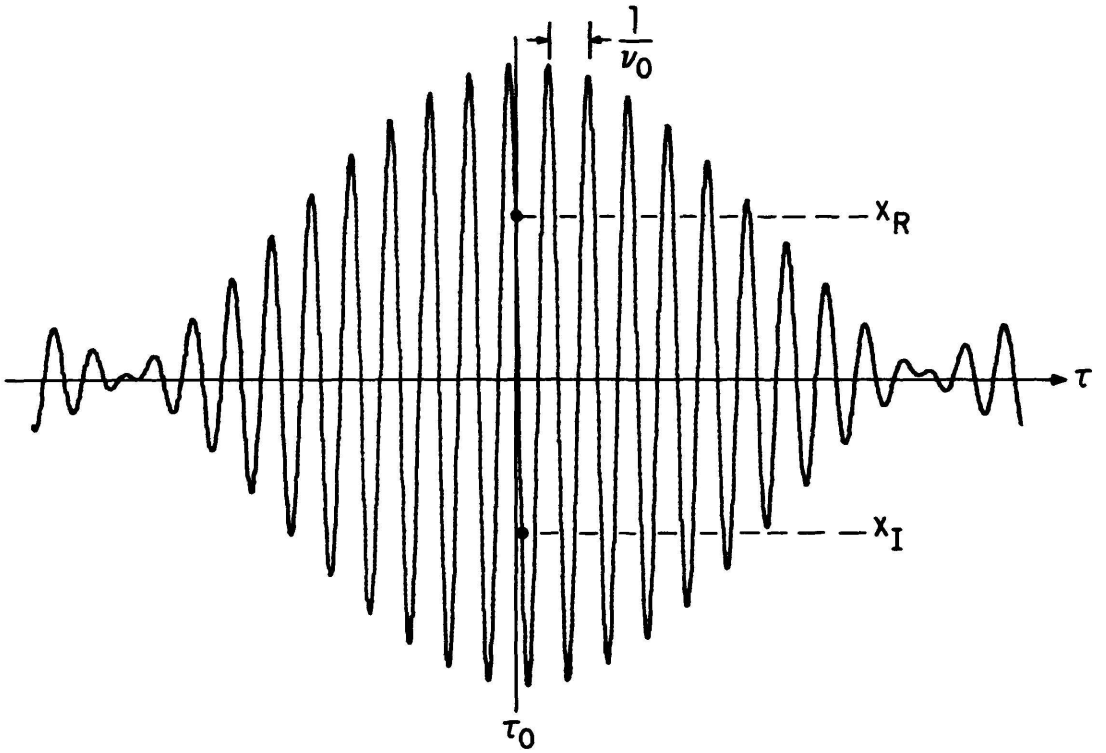


Figure 3–2. Cross correlation function of quasi-monochromatic signals with rectangular passbands centered at ν_0. In this plot, $\Delta\nu/\nu_0 = 0.2$ for clarity, but often it is much smaller.

to 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 3–3.

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 3–4. Here each box "CC" represents that part of Figure 3–3 within dashed lines, but each has a delay $\Delta\tau_k = 1/(4\nu_k)$ appropriate to its own frequency. If one is not interested in the variation of correlation with frequency, such as in the case of a continuum source, one 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

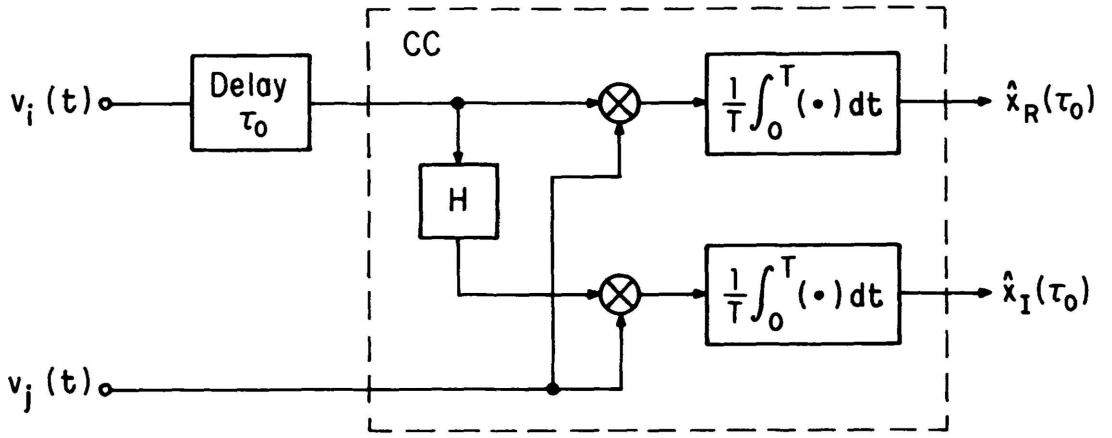


Figure 3–3. Complex correlator, for narrow bandwidth signals. "H" is a quarter-cycle delay, $\Delta\tau = 1/4\nu_0$.

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:

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

Now, it turns out that this same quantity can be measured without the elaborate filtering and multiple correlators of Figure 3–4. It is merely necessary to replace the quarter-cycle delay in Figure 3–3 (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 3–3, where "H" is a Hilbert transform filter. But the filter bank correlator of Figure 3–4 would obviously be useful for spectroscopy, where one would record the output of each complex correlator separately, rather than adding them together. However, a nearly equivalent way to obtain the spectroscopic measurements is illustrated in Figure 3–5. 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.

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 3–1. Now consider its (inverse) Fourier transform[1]

$$r_{ij}(\nu) = \int_{-\infty}^{\infty} x_{ij}(\tau)\, e^{-2\pi i\nu(\tau-\tau_0)}\, d\tau \,, \tag{3–4}$$

[1]The Fourier transform definition which is in use here is in accord with the Editors'—rather than the author's—preference. It is opposite the convention which is common in the engineering literature, particularly in much of the literature of communications engineering. — *Eds.*

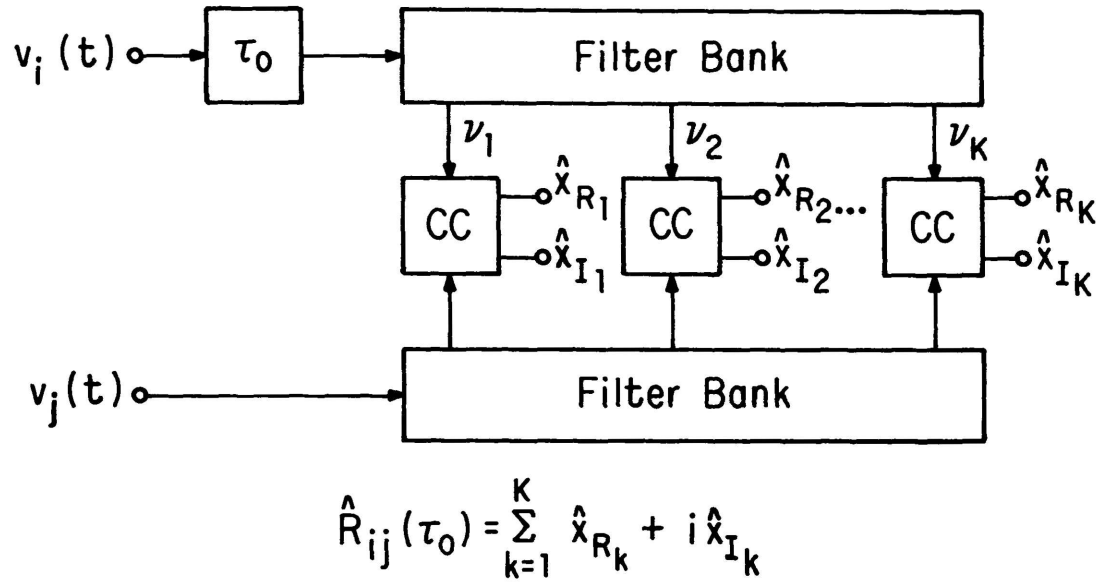


Figure 3–4. A wide-band complex correlator synthesized from narrow-band complex correlators, or a spectroscopic correlator. Each box labeled "CC" is as indicated in Figure 3–3.

which is called the cross power spectrum. (Recall that, similarly, the inverse transform of the autocorrelation 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.)[1] The complex cross correlation function is defined as

$$R_{ij}(\tau) = 2\int_0^\infty r_{ij}(\nu)e^{+2\pi i\nu(\tau-\tau_0)}\,d\nu\,; \tag{3–5}$$

i.e., it is twice the Fourier transform of Equation 3–4 with negative frequencies deleted. Notice that the correlator of Figure 3–5 approximates the right-hand side of Equation 3–4, and that adding up the outputs approximates the r.h.s. of Equation 3–5 for $\tau = \tau_0$. More precisely, the operation of the spectroscopic correlator of Figure 3–5 is described by

$$\hat{x}_k = \sum_{l=0}^{2K-1}\left[\frac{1}{T}\int_0^T v_i(t-\tau_0-l\delta\tau)v_j(t)\,dt\right]e^{-2\pi ilk/2K}\,. \tag{3–6}$$

The expression in brackets is the output of each simple correlator. Comparing Equations 3–4 and 3–6, one sees that

$$r_{ij}(k/\delta\tau) \approx \hat{x}_k\delta\tau\,. \tag{3–7}$$

It can be shown that another way to compute the continuous cross correlation function is

$$\begin{aligned} R_{ij}(\tau) &= \tfrac{1}{2}\langle[v_i(t)+i\tilde{v}_i(t)]^*\,[v_j(t+\tau)+i\tilde{v}_j(t+\tau)]\rangle \\ &= \langle v_i(t)v_j(t+\tau)\rangle + i\langle v_i(t)\tilde{v}_j(t+\tau)\rangle\,, \end{aligned} \tag{3–8}$$

where $\tilde{v}$ represents the Hilbert transform of v. Thus Equation 3–8 describes the operation of the complex correlator of Figure 3–3, except that time averages are replaced by expectations. Again, I will not give the proof here. I wish only to point out that there is a mathematical

[1] In Equation 3–4, 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.

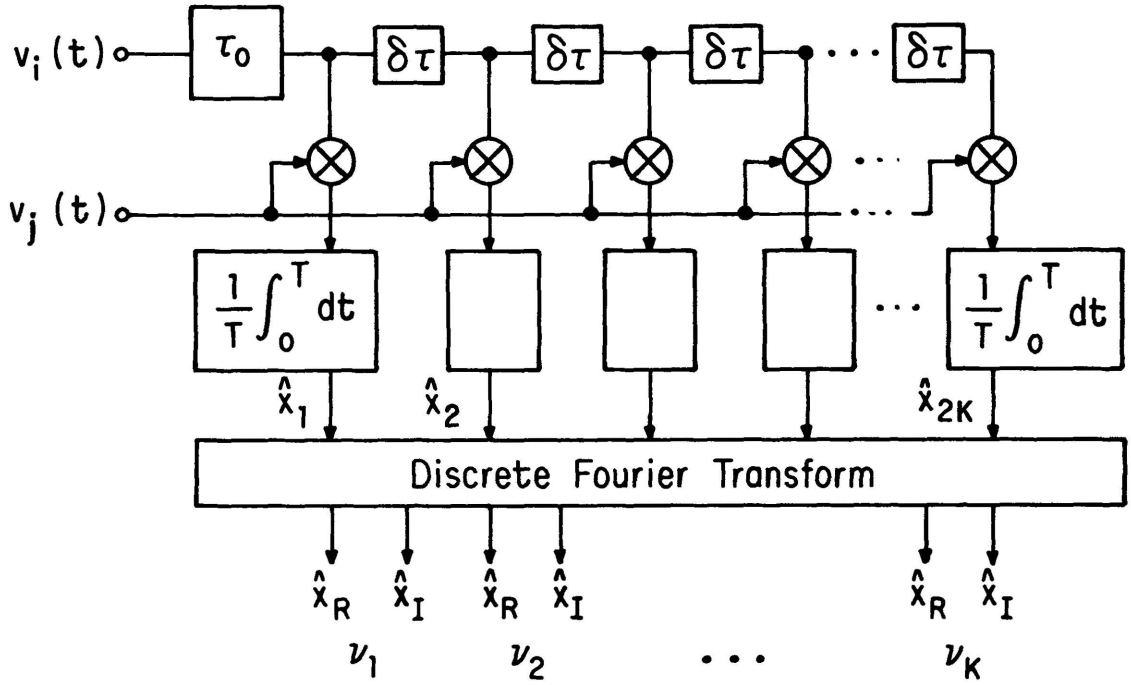


Figure 3–5. A spectroscopic correlator with frequency analysis after correlation.

formalism through which the relationships of various types of complex correlators can be explored.

Two correlators that perform equivalent computations, although different in detail, must give the same signal-to-noise ratio in their outputs, since nothing has been said about whether the inputs contain interesting information or just noise.

3. Digital Implementations

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

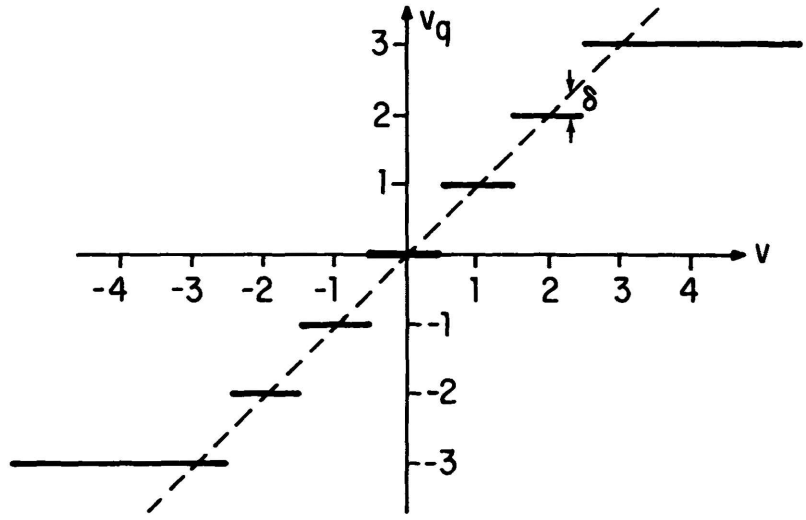


Figure 3–6. An example of a quantiser transfer function (solid lines); this quantiser has seven levels. The dashed line is the line defined by $v_q = v$, and the difference between it and the transfer function is the quantization noise, δ.

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.

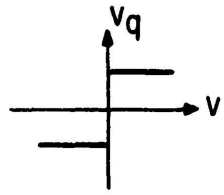
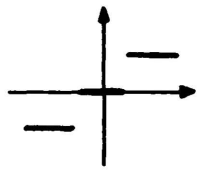
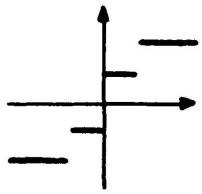
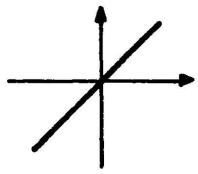
Quantization, however, does lose information. Consider Figure 3–6, which shows the transfer function of a typical quantizer. Here v is the quantizer's instantaneous input, and v_q is the corresponding output; this example shows seven distinct output states. Without loss of generality, the scale has been chosen so that v_q is the integer nearest the input value; then one 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 6).

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 one now samples 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, one is 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 3–1 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.

Table 3–1.
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.

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 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

$$x_{ij}(\tau) = \sigma_i \sigma_j \sin \frac{\pi \rho_{ij}(\tau)}{2}\,, \tag{3–9}$$

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 3–9 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 one can write

$$\rho_{ij}(\tau) = f_3(x_{ij}(\tau); \sigma_i, \sigma_j)\,, \tag{3–10}$$

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

$$x_{ij}(\tau) = f_3^{-1}(\rho_{ij}(\tau); \sigma_i, \sigma_j)\,. \tag{3–11}$$

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 3–7.

It is worth noting that relationships like Equations 3–9 and 3–11 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. 3–7). This is apparent from Equation 3–9 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.

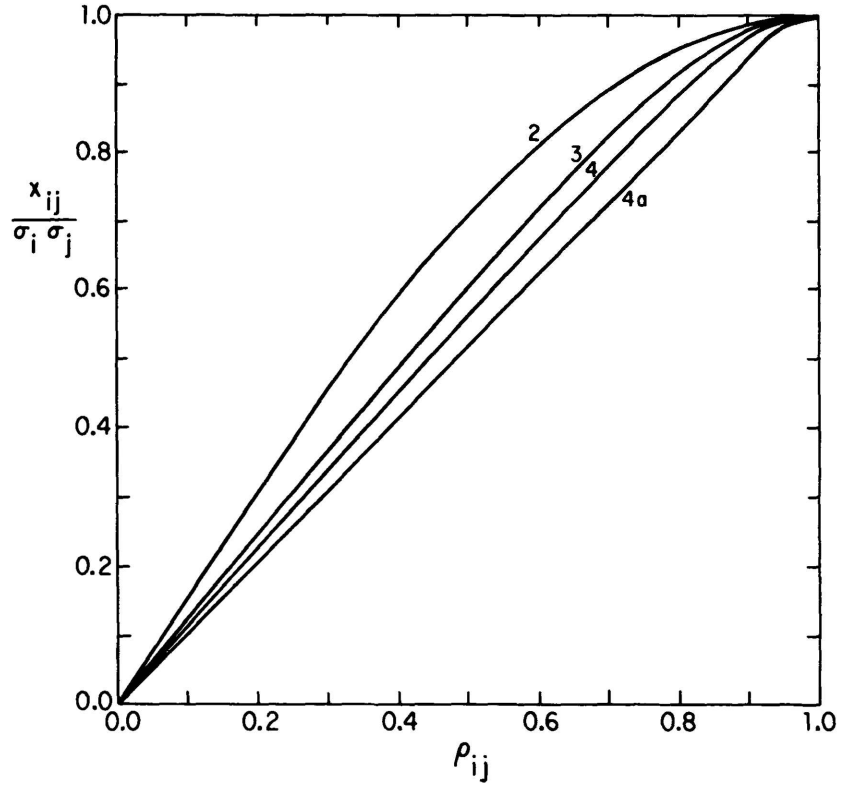


Figure 3–7. Quantization correction functions for various quantizations. In each case the signal powers are set for maximum signal-to-noise ratio. The curves are labeled according to the number of quantization levels; 4a uses a simplified multiplier (see Cooper, 1970).

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; one would rather know the cross power spectrum of the signals received at the antennas. Denoting the latter by $r'_{ij}(\nu)$, one has

$$r_{ij}(\nu) = g_i(\nu) g_j^*(\nu) r'_{ij}(\nu + \nu_{LO}) , \qquad (3\text{–}12)$$

where $g_i(\nu)$ and $g_j(\nu)$ are the complex voltage gains of the signal paths from the antennas to the correlator, and ν_{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 4 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

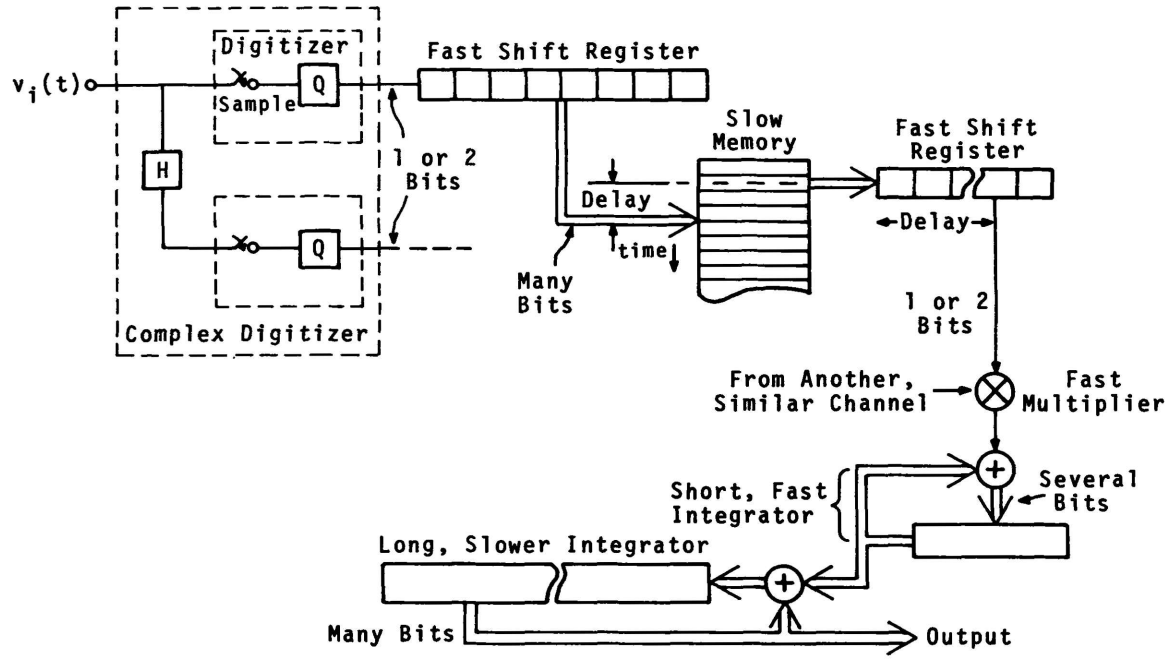


Figure 3–8. A digital implementation of a simple correlator.

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 $x_{i,j}(\tau)$ (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 3–8 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 as are needed for continuum.

4. SPECTROSCOPY

4.1. Design alternatives.

Referring back to Figures 3–4 and 3–5, recall that there are two nearly equivalent ways to implement a spectroscopic cross correlator. They differ according to whether the frequency analysis is done before or after multiplication. I want now to describe further details of the implementations, emphasizing digital circuitry.

First, note that for K frequency channels, each scheme requires $2K$ cross multipliers: two in each complex correlator of Figure 3–4, and one for each of $2K$ delays in Figure 3–5.

In Figure 3–4, with frequency analysis before cross multiplication, the filter banks could be implemented by analog circuits, using the undigitized signals. In that case, the long delay line τ_0 would also need to be analog. Such a design could be practical for a telescope requiring a relatively small number of baselines and frequency channels. Alternatively, the filters could be implemented digitally, operating on digitized signals, using length-$2K$ shift registers and fast Fourier transforms (FFTs). These would have to be capable of fast operation (an FFT every $2K$ samples), and the outputs would require more bits than the inputs by a factor of $\log_2 2K$ to avoid additional quantization noise. The correlators could be relatively slow (a factor of $2K$ below the sampling rate), but would have to handle multibit data words. There would also be losses associated with the fact that input samples not in the same $2K$-sample interval are never correlated. These tradeoffs are complicated and must be evaluated for each particular system's parameters. I will not discuss this arrangement any further, but I want to note that it has been chosen for at least one modern synthesis telescope, the millimeter wavelength array at Nobeyama Observatory, Japan.

In the other scheme (Fig. 3–5), with post-correlation frequency analysis, the multipliers must operate at the full sampling rate, but on signals having only a few possible values. The FFT has multibit input and output, but needs to be done only once per integration time (which seems like an eternity compared with a sample time, e.g., 10 sec/(10^{-6} sec) = 10^7); a floating point FFT is usually justified. This is the scheme used at the VLA, and I will concentrate on it from now on.

I should mention, however, that it is also possible to choose a design between those of Figures 3–4 and 3–5, where part of the frequency analysis is done before and part after correlation. Such a "hybrid" correlator, with an analog filter bank and digital cross correlators, is useful when the total input bandwidth is too large for processing all at once; this happens mainly at millimeter wavelengths.

Notice that, with digitized signals, the small delays $\delta\tau$ must be multiples of the sampling interval. This would seem to be no problem, because if the original signals have

get away without accurate quantization corrections.

4.3. The Gibbs phenomenon.

The effects of truncation of the cross correlation function measurement are not so avoidable, and they can have a profound effect on calibration that is quite different from the case of the autocorrelation spectrometer. To see this, note that the sampling theorem allows one to write the cross power spectrum as an infinite sum

$$r(\nu) = \int_{-\infty}^{\infty} x(\tau)\, e^{-2\pi i \nu(\tau-\tau_0)}\, d\tau \tag{3–13a}$$

$$= \sum_{k=-\infty}^{\infty} x(\tau_0 + k\delta\tau)\, e^{-2\pi i \nu k \delta\tau}\, \delta\tau\,, \tag{3–13b}$$

where the second equation holds only within the bandwidth $0 \le \nu \le \Delta\nu < 1/(2\delta\tau)$. If the sum is truncated beyond $|k| = K$, the result may be written

$$\hat{r}(\nu) = \sum_{k=-\infty}^{\infty} \sqcap(k/K)\, x(\tau_0 + k\delta\tau)\, e^{-2\pi i \nu k \delta\tau}\, \delta\tau \tag{3–14a}$$

$$= \int_{-\infty}^{\infty} \sqcap(\tau/K\delta\tau)\, \sqcup\!\sqcup(\tau/\delta\tau)\, x(\tau) e^{-2\pi i \nu(\tau-\tau_0)}\, d\tau \tag{3–14b}$$

$$= r(\nu) * \int_{-\infty}^{\infty} \sqcap(\tau/K\delta\tau)\, \sqcup\!\sqcup(\tau/\delta\tau)\, e^{-2\pi i \nu(\tau-\tau_0)}\, d\tau\,, \tag{3–14c}$$

where $\sqcap(\cdot)$ is the unit rectangle function and $\sqcup\!\sqcup(\cdot)$ 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 $K\,\mathrm{sinc}(K\nu\delta\tau)$.

Now consider the situation illustrated in Figure 3–10, where the actual and computed cross power spectra are shown for signals from a unit-flux continuum source in the reference direction; thus the interferometer's gain vs. frequency function is shown, and in this 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 3–14c; 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

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

Dividing by the apparent gain gives

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

Notice that the convolution operations prevent cancelling of the gains, as one might desire. There are better ways of estimating $V(\nu)$ than simply taking the above ratio, such as

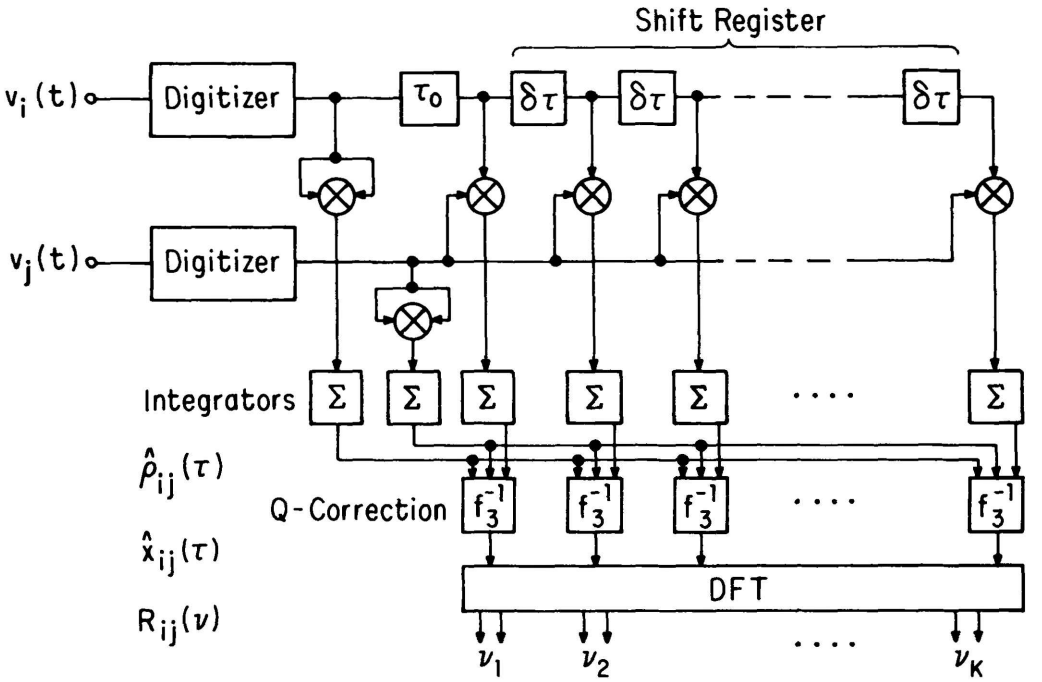


Figure 3–9. A digital cross correlation spectrometer, with self-multipliers and quantization corrections.

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 in fact the scheme of Figure 3–5 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". Secondly, 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.

Except for the quantization noise, which affects both continuum and spectroscopic digital correlators, most of the non-ideal behavior of a digital cross correlation spectrometer can be explained by the non-zero delay interval $\delta\tau$ and the finite range of delays measured. I want now to consider these effects in some detail.

4.2. Quantization corrections.

The systematic effects of the quantization on the cross power spectrum can be eliminated, in principle, by applying the quantization correction to each cross correlation measurement prior to Fourier transforming. Each measurement is then adjusted to what it would have been without quantization, except for the quantization noise. This arrangement is shown in Figure 3–9, which also illustrates the use of "self-multipliers" to determine the signal powers. 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, this effect becomes important for correlation coefficients above about 0.2. 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

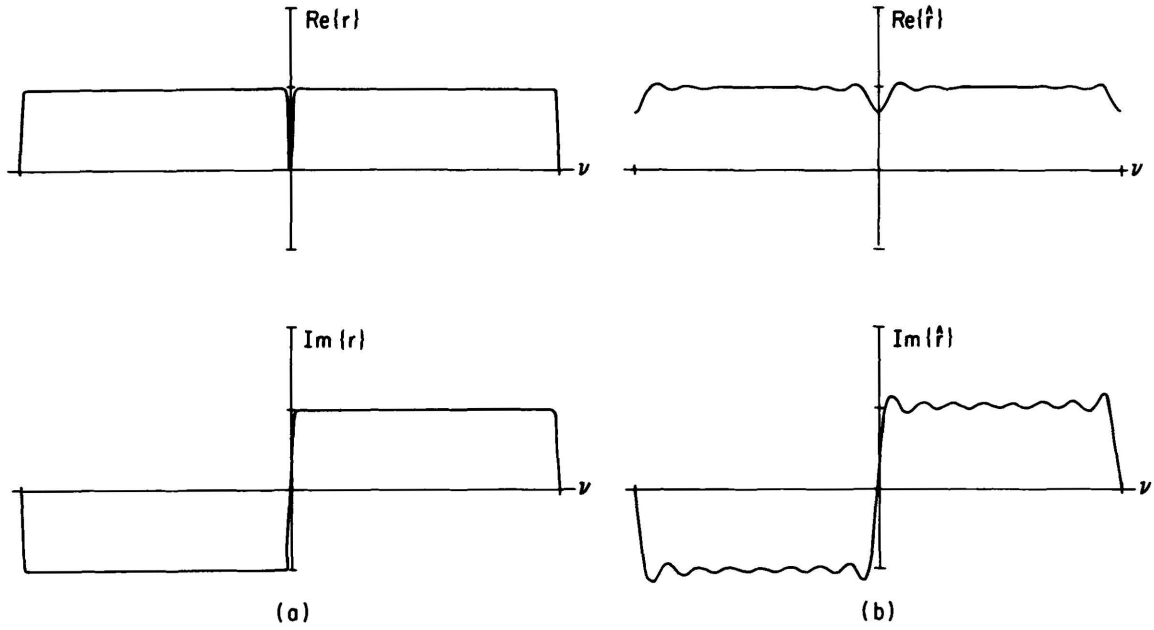


Figure 3–10. (a) The cross power spectrum resulting from a continuum source of unit flux in the reference direction: "true complex gain". Note the nonzero phase. (b) The computed cross power spectrum with 16 delays.

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 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 correlation function 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 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 $\nu_{LO}\tau_g$. This phase shift is called "fringe rotation".

In this Section, I will describe how a correlator can handle signals from receivers that do not include fringe rotation in their local oscillators. We must also consider the accuracy with which the instrumental delay must be set, and the consequences of setting it rather coarsely. I shall take up the latter question first.

Assume that the required delay can be calculated accurately and to arbitrary precision, but can only be set to discrete values spaced by $\Delta\tau$. In general 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, one 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 3–5 gives

$$R_{ij}(\tau_0 + \delta\tau) = 2\int_0^\infty r_{ij}(\nu)e^{2\pi i\nu\delta\tau}\,d\nu\,. \tag{3–15}$$

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

$$R_{ij}(\tau_0 + \delta\tau) = R_{ij}(\tau_0)\frac{1}{\Delta\nu}\int_0^{\Delta\nu} e^{2\pi i\nu\delta\tau}\,d\nu\,. \tag{3–16}$$

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 3–4 with $x(\tau + \delta\tau)$ and apply the shift theorem of Fourier transforms, obtaining

$$\hat{r}_{ij}(\nu) = r_{ij}(\nu)e^{2\pi i\nu\delta\tau}\,. \tag{3–17}$$

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. 3–5) because of the finite length of the transform, but Equation 3–17 holds fairly well for practical numbers of points.

Equations 3–16 and 3–17 are evaluated in Table 3–2 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.

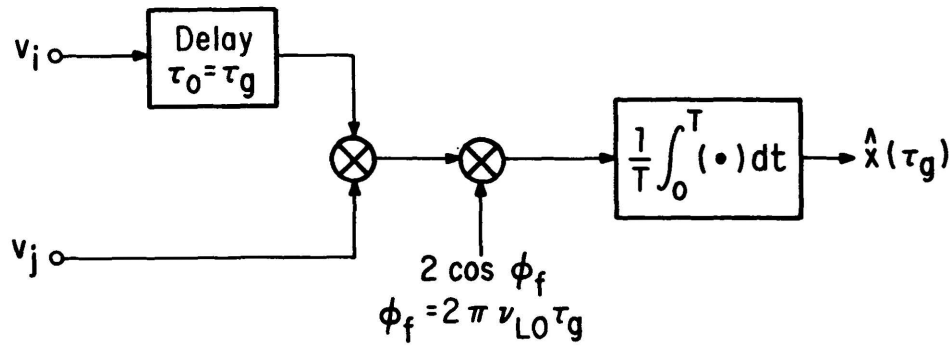


Figure 3–11. A simple correlator with fringe rotation.

Table 3–2.
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 3–2 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 (like Fig. 3–1) produces the output

$$x'(\tau_g) = x(\tau_g)\cos 2\pi\nu_{LO}\tau_g + \tilde{x}(\tau_g)\sin 2\pi\nu_{LO}\tau_g\,, \qquad (3\text{–}18)$$

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; ν_{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 only applies 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 3–11 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.

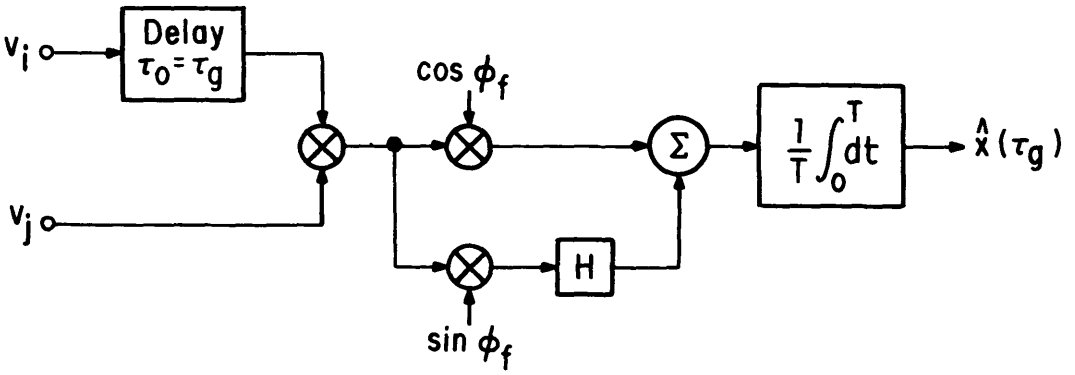


Figure 3–12. A simple correlator with "single-sideband" fringe rotation.

If the averaging time is an integral number of cycles of the fringe function, then the correlator of Figure 3–11 produces an unbiased estimate of $x(\tau_g)$. Nevertheless, this method has some disadvantages. One must integrate for at least one "fringe", and sometimes $\nu_{LO}\tau_g$ changes too slowly for this. More importantly, the signal-to-noise ratio obtained is worse than that of the Figure 3–1 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 3–12. 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 3–18, and gives a signal-to-noise ratio equal to that of Figure 3–1. 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, where the correlation function is to be measured for many closely-spaced values of τ. In that case, one can build a correlator like that of Figure 3–13. Note that the order of the fringe rotation and cross correlation multiplications has been interchanged, but that this has no effect since multiplication is associative; thus only one fringe rotator is needed for all delays. If one 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, one can show 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 6.) 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.

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. This has made sense because 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

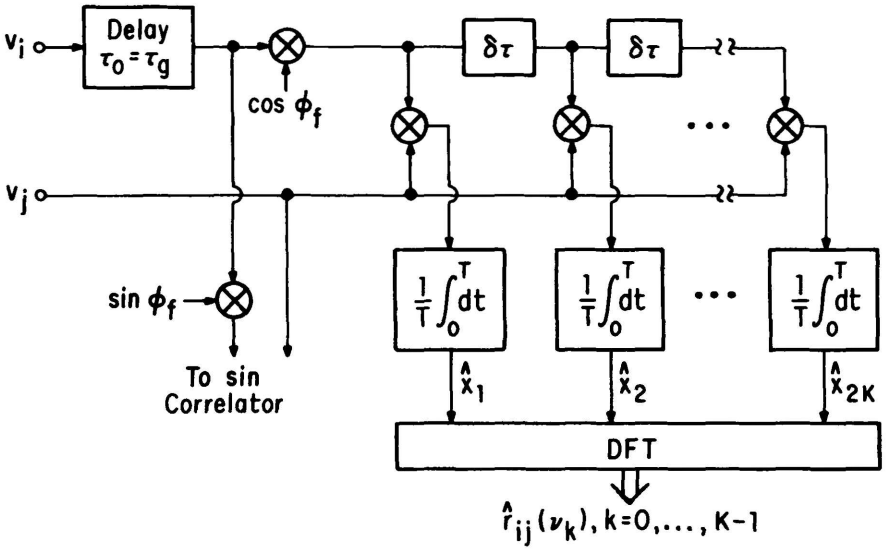


Figure 3–13. A spectroscopic correlator with fringe rotation.

designed for single-dish work. But with new, modern systems now being designed, the bandwidths and numbers of antennas are larger, vastly increasing the amount of cross correlation electronics. Nevertheless, as of this writing, the new VLBA is being designed in the "traditional" way, without LO fringe rotation.

REFERENCES

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

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

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

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 corrections of VLA correlation measurements", VLA Computer Memorandum No. 150.

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.

4. Calibration

R. Carl Bignell and Richard A. Perley

Introduction

In Lecture 1 it was shown that—after a few reasonable assumptions are made—the intensity distribution on the sky is the 2-D Fourier transform of the spatial coherence function of the radiation field (Equation 1-8). An interferometric array measures this spatial coherence function at many discrete locations specified by the projected baseline components, (u, v). In Lecture 2 it was described how the signals from each antenna are transported to a central location where, as outlined in Lecture 3, these signals are correlated and the correlations are averaged. The data from each antenna pair are then recorded; the ensemble of numbers is commonly called the *observed visibilities.*

But, before these data are recorded the radio signals must pass through the intergalactic, interstellar, and interplanetary media, and through the Earth's atmosphere. After collection by the radio antennas, the signals pass through, and are modified by, the receivers, the signal transmission system, data digitizers, and the correlator. Each medium through which the information passes modifies the data, with the result that the observed visibility often shows little resemblance to the desired quantity, the spatial coherence. Calibration is the process of determining and applying the corrections needed to produce the spatial coherence function, so that the imaging procedures, discussed in Lecture 5, can give a usable representation of the sky brightness. In this Lecture we discuss the origins and effects of various mechanisms important to interferometric data, the techniques of their determination, and the methods of correction.

1. Levels of Calibration

Calibration is the art of determining and removing the effects of corruption from the data. One can discern three levels in the calibration of interferometric instruments. These levels are distinguished by origin and timescale.

The first level calibrates effects which are unchanging, or nearly so, over long timescales. This includes antenna locations and pointing, delay constants, time reference, and receiver characteristics. Typically, the observations required for these calibrations are performed after changes in array geometry or hardware. Generally, the quantities determined are applied to the data on-line, so no further action need be taken by the observer unless the applied corrections are themselves in error. In most situations, the data may be corrected later.

The next level of calibration involves changes induced by the array electronics, such as transmission system length changes, or receiver sensitivity changes. In many cases, these can be reduced to an acceptable level by good design. Where this is not possible they can usually be monitored and corrected by on-line monitoring systems. As this level of calibration is so intimately connected with design, we pay little attention to it in this Lecture, except when the effects can be corrected by off-line calibration.

The final level of calibration, and the one of most interest to the observer, involves corrections for changes in the visibility induced by the atmosphere and the electronics, and for which acceptable on-line correction is not possible. These changes (which affect phase

much more than amplitude) can range from small values on timescales of hours or more to many radians in timescales of less than a minute. Even in these extreme cases, it is often (even usually) possible to remove completely the time-variable effects and thus to recover the spatial coherence function.

2. Sources Used for Calibration

Conceptually, the process of calibration is one of determining system constants by observing known sources of emission, and it is clearly advantageous that these known sources be as simple as possible. Thus, the ideal calibrator is an unresolved radio source with a high flux density and a well-determined position. If such a source is observed at the phase center, then the amplitude of the measured visibility must be equal to the source flux density, and the phase of the visibility must be equal to zero, independent of the baseline length. Thus, the visibility measurements obtained from observations of a calibration source yield direct estimates of the corrections which are needed to calibrate the array around the time of those observations for the direction of the calibrator. In principle, no knowledge of the mechanisms which produce the observational errors is required.

How strong, and how small, should a calibration source be? Ideally, the flux density, and hence the amplitude of the visibility, should be many times the system noise on short timescales. The flux density should also be many times the sum of the flux densities of background sources which also contribute to the visibility—since no source is truly isolated. This condition is especially important at lower frequencies, where the antenna primary beam will include large numbers of background sources. The calibrator should also be small enough that the longest baselines do not perceive a loss of visibility amplitude greater than the noise.

In practice, it is difficult, and in some cases impossible, to find sources which meet these conditions, especially with high-resolution arrays. However, it is often quite acceptable to utilize 'less-than-perfect' sources for calibration, using techniques akin to self-calibration. These techniques will be outlined in a later Section, but involve solutions for antenna gains rather than baseline gains. Since nearly all modern synthesis radio telescopes contain many more baselines than antennas, not all the available data need be used for this solution—so longer spacings, which may partially resolve a calibrator, or smaller spacings, which may be confused by background sources, can often be discarded from the solution. Indeed, as discussed in Section 5, if a reasonable initial guess of the source structure can be made, and a good idea of the total flux is at hand, independent calibration can often be dispensed with altogether.

The number of available calibrators varies widely with frequency, with resolution, and with sensitivity. Assuming sensitivities typical of modern antennas (say, with baseline noises of less than 50 mJy), there are over 500 radio sources which can be used as calibrators at frequencies between 1 and 20 GHz and on angular scales as small as approximately 0.01 arcseconds. Above this frequency range, the number of usable calibrators is reduced as receiver sensitivities are less. Below this range, the increasing primary beam size, combined with a rapidly increasing galactic background temperature, allows us to use only the strongest sources as simple calibrators. Calibration of large interferometer arrays at low frequencies will probably involve forms of self-calibration (Lecture 9). At milliarcsecond resolutions (i.e. with VLBI techniques), there may remain no sources which are sufficiently unresolved to allow straightforward calibration. See Lecture 13 for discussion of the special problems in calibration at these resolutions.

3. THE CALIBRATION FORMALISM

The calibration formula, in a reasonably general form, is

$$V'_{ij}(t,\nu) = \mathcal{G}_{ij}(t,\nu)V_{ij}(t,\nu)\,, \tag{4–1}$$

where t is the time of the observation, ν is the frequency, subscripts i and j refer to the measurement associated with antenna pair (i,j), V is the true visibility function, V' is the measured visibility function, and $\mathcal{G}$ the baseline-based gain. Each of these quantities is complex-valued. This formulation assumes that the corrections are linear for each baseline and that there is no crosstalk amongst them. These assumptions are generally well-satisfied with modern systems.

Virtually all the corruption of data takes place before correlation, so that the effects can be identified with individual antennas, rather than baselines. This observation allows the correlator gain to be factored to a product of antenna gains, so that the calibration formula can be written:

$$V'_{ij}(t,\nu) = G_iG_j^*G_{ij}V_{ij}\,, \tag{4–2}$$

where the G_i are the (complex) antenna gains, and G_{ij} is a residual, correlator-based gain. If the assumption above is perfect, then $G_{ij} = 1$. All quantities are functions of time and frequency.

In order to determine the G_i's for the N antennas from the $\mathcal{G}_{ij}$'s, the set of equations

$$\mathcal{G}_{ij} = G_iG_j^*\,, \qquad \text{for } i = 1,\ldots,N, \quad j = i+1,\ldots,N\,, \tag{4–3}$$

must be solved. Because this set of equations is invariant with respect to an arbitrary phase shift in all of the G_i, the phase part of one antenna-based gain can be set to zero. The number of complex equations is equal to $N(N-1)/2$, and the number of (real-valued) unknowns is $2N-1$; the least-squares technique is generally used to determine the G_i. The validity of the assumption that the gain corruptions are antenna-based rather than correlator-based can be checked by examining the residuals of the solutions.

In the calculation of the correction coefficients, G_i, it is convenient and also physically meaningful to deal with the amplitude and phase rather than with the real and imaginary parts. This is so because the primary effects of propagation are to rotate the phase and decrease the amplitude. Better physical insight into the effects involved is gained by examining the amplitude and phase solutions, rather than the real and imaginary parts. Hence it is convenient to separate the complex Equation 4–3 into its modulus and argument.

Continuum interferometers return no information about the shape of the spectrum within the passband supplied to the correlator. We thus can drop the explicit frequency dependence shown in Equation 4–2, so that we can write the calibration equation as

$$A'_{ij}(t)e^{i\phi'_{ij}(t)} = A_i(t)A_j(t)e^{i(\phi_i(t)-\phi_j(t))}A_{ij}(t)e^{i\phi_{ij}(t)}G_{ij}(t)\,, \tag{4–4}$$

where

$A'_{ij}(t)$ is the measured visibility amplitude,1.13truein
$\phi'_{ij}(t)$ is the measured visibility phase,
$A_{ij}(t)$ is the true visibility amplitude,
$\phi_{ij}(t)$ is the true visibility phase,

i, j	are subscripts denoting two antennas, the ith and jth,
t	is the time,
ν	is the frequency,
$A_i(t)$	is the gain correction for antenna i,
$\phi_i(t)$	is the phase correction for antenna i, and
G_{ij}	is the residual, baseline-based gain.

We discuss the phase corrections in Section 4 and amplitude calibration in Section 5. Bandpass and polarization calibration techniques are discussed in Sections 6 and 7, respectively.

4. Phase Calibration (or Focusing the Array)

Signals collected by the array elements have traversed long distances through media with different refractive indices. These variations mean that the propagation times differ from those that would have occurred, had the path been *in vacuo*. Most importantly, the signals collected by different elements have undergone different delays. Further differential delays occur due to the electronics required for conduction of the signals to the correlator. The net result of these variations is that the phase of the visibility is not that which would have been obtained by an ideal system. An analogy can be drawn with a paraboloidal surface which reflects radiation to a focal point. The geometry guarantees that signals arriving from the direction perpendicular to the plane of the antenna arrive in phase at the focus. The object of phase calibration of an array is to cause radiation from the phase tracking center to arrive at the correlator in phase. Thus, the process of phase calibration could be considered focusing the array. Because the data from each baseline are individually collected, it is not necessary to apply this calibration in real-time. Nevertheless, many of the phase-changing effects can be calculated, or monitored, in real-time, and the subsequent phase changes applied in real-time. In this Section, we discuss the major origins of phase perturbations.

4.1. Delay calibration.

As discussed in Lecture 2, signals from a celestial source must arrive at the correlator at the same time for correlation over a nonzero bandwidth. The accuracy required for coherence depends on the bandwidth. Consider two monochromatic signals, of frequency ν_0, arriving at the correlator in phase, but with delays differing by $\delta\tau$. Signals traversing the same path at frequency ν will arrive at the correlator differing in phase by an amount $\delta\phi$,

$$\delta\phi = 2\pi(\nu - \nu_0)\delta\tau\,, \tag{4-5}$$

where $\delta\tau$ is the difference in the propagation time of the signals between the wavefront and the correlator. Thus, the delay difference must satisfy the inequality $\delta\tau \ll 1/\delta\nu$ in order for the the signals across a bandwidth $\delta\nu$ to add up in phase.

There are two major components of the delay. The first is the geometric delay τ_g (Equation 2–3) which can easily be calculated from the array geometry and the location of the radio source. Several techniques are available for insertion of a variable delay in order to compensate for the geometric delay (Lecture 2, Sec. 3). Note that the geometric delay can be compensated for one direction in the sky only, so that a degree of incoherence must exist for all other directions. This is the origin of bandwidth smearing, discussed in Lectures 2 and 8. It can be held to acceptable levels only by using sufficiently narrow bandwidths.

The second component of the delay is caused by propagation time for the signals imposed by the hardware associated with each antenna. Interferometers are constructed

to make this delay equal for each signal path, and as time-invariant as possible. Through good design, the differences between antenna delays can be held constant (through internal monitoring) to better than than 0.1 nanoseconds (or 3 cm), sufficient to maintain good coherence for up to 500 MHz bandwidth. Various schemes for monitoring the internal delay are implemented on different interferometers. The time-invariant (or slowly time-variant) part of the delay for each antenna-IF is usually determined from calibrator observations. For a square bandpass of frequency width $\delta\nu$, the measured amplitude of the visibility function as a function of delay error $\delta\tau$ is given by

$$A'(\delta\tau) = \frac{\sin \pi\delta\nu\delta\tau}{\pi\delta\nu\delta\tau}. \tag{4-6}$$

The relative delay between the antennas is stepped in units of $1/\delta\nu$ until a maximum of A' is found. The geometric delay must be compensated for during this measurement. This technique of delay fitting is used in VLBI (Lecture 13) to determine both the positions of radio sources and the antenna locations.

4.2. Calibration of baselines.

If the true baseline differs from the presumed baseline by an amount $\Delta\mathbf{b}$, a phase error

$$\delta\phi = 2\pi\Delta\mathbf{b}\cdot\mathbf{s}\,, \tag{4-7}$$

results, which has a characteristic sinusoidal dependence on source declination and hour angle. Explicitly, if the antenna locations are expressed in a coordinate system with the x-axis pointing toward $\delta = 0°$, $h = 0^h$, the y-axis toward $\delta = 0°$, $h = -6^h$, and the z-axis toward $\delta = 90°$, and if Δb_x, Δb_y, and Δb_z are the components of the baseline errors expressed in this reference frame, then

$$\delta\phi = 2\pi((\Delta b_x \cos h - \Delta b_y \sin h)\cos\delta + \Delta b_z \sin\delta)\,. \tag{4-8}$$

The above equation holds for identical antennas. For non-identical antennas, the true baseline will be a function of antenna pointing position, and extra terms describing this will be necessary. Calibration of the baselines is straightforward in principle. Observations of many calibrators of well-determined positions are taken, preferably under conditions of good atmospheric stability. The coefficients of Equation 4–8 can then be determined.

The array lies on a moving object, the Earth. Since we must determine the antenna locations with respect to a fixed reference frame, precise knowledge of the motion of the Earth is required. The important contributions are:

(1) the Earth's rotation (i.e., the time),
(2) the precession and nutation of the Earth's axis of rotation,
(3) the wandering of the pole of the Earth, and
(4) the effects of Earth tides.

The non-uniformity of the Earth's rotation can be predicted to an accuracy of a few milliseconds of time. If higher accuracy is required, post-observing corrections can be made (see Sec. 4.3). Precession of the Earth's pole is approximately $20''$ per year. The largest contribution to nutation has a 19-year period and $9''$ amplitude. There are other, smaller terms of shorter period. Correction for these effects is required if data from different observations are to be combined. Besides precession and nutation, the Earth's pole undergoes an erratic wandering amounting to approximately $0''.01$. Predictions of this motion can be made.

Earth tides, caused by the Sun and Moon, cause a crustal displacement of approximately 30 cm, or 0.″01 in apparent position of an astronomical source.

Two other effects, not related to baselines, need to be mentioned, since their effects must be corrected for in order to use interferometric data. Aberration occurs due to the Earth's revolution about the Sun, and to its rotation about its axis. The maximum angular shifts are 20″ and 0.″26 respectively. Bending of rays due to the gravitational field of the Sun causes an angular displacement away from the Sun of 1.″75 at the solar limb, decreasing nearly linearly to 0.″004 at 90° from the Sun.

All other effects are believed to contribute less than 0.″002 error in the measurement of position of radio sources. Note that calibration of data through observation of nearby sources will reduce the errors of most of these effects. However, observations intended for astrometry must include all these corrections, if milliarcsecond accuracy is desired.

4.3. Correction of time errors.

Occasionally errors occur in the clock used as the time-reference for the array. This introduces a phase error given by:

$$\delta\phi = -2\pi\omega\delta t(b_x \sin h + b_y \cos h)\cos\delta\,. \tag{4–9}$$

where ω is the angular rotational velocity of the Earth, 7.29×10^{-5} rad s^{-1}. There is a fundamental limit to correcting the phase errors so introduced, because it is practical to set the clock only to an accuracy of a few milliseconds, the limit of time-keeping systems. The seriousness of time-keeping errors is much reduced however, by calibration through observations of nearby sources, since the phase error will be nearly the same for both calibrator and source.

4.4. Atmospheric phase errors.

The wavefront from a distant radio source is distorted in its journey from the source to the array, so that the phase measured by the correlator differs from that characterizing the coherence function. At centimeter wavelengths, the most important distortion occurs in the neutral atmosphere, where both the dry and wet components of the troposphere slow the propagation of the signals, while at meter wavelengths the dominant effect is due to the ionosphere. Recent reviews on the astronomical importance of these effects is given in Meeks (1976), and in Chapter 13 of Thompson, Moran, and Swenson (1986). The latter reference is especially recommended.

The extra path introduced by propagation through the atmosphere is characterized by the excess path length, $\delta L = c\delta t$, where $\delta t = \frac{1}{c}\int(n-1)dx$ is the extra propagation time introduced by the atmosphere, and n is the index of refraction. What is noted by the interferometer is not the introduced delay *per se*, but the **difference** in delay between the two antennas comprising the interferometer. Thus, the phase error introduced by the atmosphere is

$$\phi = 2\pi(\delta L_1 - \delta L_2)/\lambda\,, \tag{4–10}$$

where the subscripts refer to the two antennas. Two origins for the excess path length can generally be distinguished—the first due to the large-scale, global structure, and the other to small-scale turbulence. The large-scale structure can often be estimated and corrected for, using estimates or measures of atmospheric structure. The effects of turbulence are in general not calibratable, except through use of self-calibration.

The depth of the troposphere is about 6 km, and the decrease in the speed of propagation—usually called refraction—is about 1 part in 3000, producing an additional path length, at the zenith, of about 2.3 m. At zenith angles z less than about 80°, the effect

of Earth curvature can be neglected, and a reasonable approximation to the excess path (in cm) can be written $L = L_0 \sec z$, where $L_0 = 0.228 P_0 + e_0$. Here, P_0 and e_0 are the total pressure and water vapor partial pressure in millibars. Note that there is no dependence on frequency.

A plane–parallel atmosphere has no effect on interferometer phase (since both rays take the same time traversing it), so that the only correction needed is a small adjustment for refractive bending (typically an arcminute). However, due to Earth curvature, widely separated antennas require a phase correction, since they view the source at different elevations. A good approximation to the differential excess path is

$$\delta L = \sec z \left(\delta L_0 + \frac{c\tau L_0}{r_0} \sec z \right) , \tag{4–11}$$

where δL_0 is the difference in vertical extra path, due to the difference in heights of the antenna, τ is the geometric delay, and r_0 is the Earth's radius. For baselines of tens of kilometers, the typical correction amounts to a few centimeters.

The troposphere is characterized by micro-turbulence at size scales from meters to kilometers which has little effect on ground-based weather measurements. It is believed that the irregularities in the distribution of water vapor dominate the observed phase irregularities, which occur with characteristic timescales varying between seconds and hours. These phase fluctuations can be described by a random process with an r.m.s. that is a power law of antenna separation $|\mathbf{b}|$,

$$\phi = \phi_0 |\mathbf{b}|^a \tag{4–12}$$

where the baseline is in kilometers, and ϕ_0 is the r.m.s. phase on a 1 km baseline. Numerous test measurements at the VLA yield values of the coefficients of the root Allan variance[1] at 1 km baseline for an 8 minute timescale; median values, expressed in millimeters of excess path length, are given in Table 4–1.

Table 4–1.
Median Coefficients for Root Allan Variance

	ϕ_0	
	Day	Night
April–September	2.2	1.0
October–March	1.1	0.6

The index a of the power law is predicted by the Kolmogorov theory of turbulence to be 0.83. The observed index varies from zero to about 0.9 for antenna separations ranging from a few tens of meters to a few kilometers. The observed low values for the index indicate that the interferometer is sensitive to a regime of the power spectrum of fluctuations near the outer scale of turbulence at about one kilometer baseline and 8 minute timescales. The median index is about 0.3.

At higher frequencies on long baselines, these tropospheric effects severely limit the ability to construct coherent images. Work on monitoring the variations in the wet component (believed to be the more important part) through monitoring of atmospheric emission near the water vapor resonance line at 22.2 GHz has been done. It appears that corrections

[1]The Allan variance, for phase fluctations of characteristic timescale τ, is given by $\sigma^2(\tau) = \frac{1}{2\tau^2}\langle(\phi(t-\tau) - 2\phi(t) + \phi(t+\tau))^2\rangle$, where $\phi(t)$ denotes the phase as a function of time, and where the angle brackets denote the expectation value (averaging over time).

to the excess path can be made, accurate to better than 1 cm. However, at millimeter wavelengths, considerably better accuracy is needed before this technique can be used. It should also be remembered that due to the short correlation scale of water vapor irregularities (< 1 km), each antenna of an array must be outfitted with a radiometer—resulting in a very significant cost for multi-element arrays.

These fluctuations generally limit the accuracy to which positions of astronomical sources can be measured (and, obviously, limit the determination of baseline parameters as well).[1]

Limited success may be obtained by attempting to calibrate the effect of turbulence by observations of nearby calibrator sources. The more frequent the calibrator observations and the closer the calibrators are to the program source, the more likely it is that the calibration will adequately compensate the small-scale variations in tropospheric refraction over each antenna. It is found that calibrator-source separations less than 10°, and timescales of less than 10 minutes, are required to reduce the effects of tropospheric turbulence.

In Lecture 9 the self-calibration algorithm is discussed. This calibration technique uses the radio source itself (provided that it is sufficiently strong) as the test signal for determining the antenna-based phase errors, the bulk of which are produced by differential refraction above the antennas. This phase calibration technique is far superior to calibration by a nearby source, but it cannot be used when the source flux density is comparable to the noise (per baseline). Unfortunately, the fraction of astronomical observations for which this powerful technique can be applied is a decreasing function of frequency, since the flux densities of most sources, the number of background sources in the antenna beam, and the antenna sensitivities all typically decrease with increasing frequency.

4.5. Ionospheric phase errors.

The ionosphere is a magneto-active plasma mainly confined to a region 60–2000 km above the surface of the Earth, with the most important effects on radio astronomical observing caused by the region 300–500 km in height. The propagation paths of radio waves passing through this medium are affected because the index of refraction is a function of both the electron density N_e and the magnetic field strength (the latter dependence is important to propagation of polarized radiation). Given a profile of the electron density along the radio ray, the excess path can be calculated. A typical value is $L_0 = -4 \times 10^6 \nu^{-2}$ meters, where the frequency ν is in MHz. Note that the excess path is negative, meaning that the phase is advanced relative to vacuum (the physically relevant group delay is positive), and that there is a ν^{-2} dependence, so that ionospheric effects are dominant at low frequencies. Atmospheric and ionospheric effects are typically about equal near 4 GHz, but effects of the ionosphere can occasionally be seen as high as 8 GHz. Temporal changes in the phase of the radio signal passing through the ionosphere result from temporal changes in the electron density. There is a large diurnal effect, due to solar heating, in which N_e changes by as much as a factor of 10. There are also anomalous variations, with timescales of minutes or less.

The large-scale, or spherical, component of the ionosphere can be estimated and removed using models based on either (a) past history, or (b) the total electron content of the ionosphere, as measured by an ionosonde or by satellite transmissions. Given the vertical excess path L_0, the differential path at zenith angle z can be written

$$\delta L = \frac{L_0 c\tau}{r_0 \cos^2 z + 2h}, \qquad (4\text{–}13)$$

[1] For short observations at the VLA in the A configuration at 6 cm, source positions are accurate to about 0″.1, while observations over many hours may be accurate to a few times 0″.01.

where τ is the geometric delay, h the height of the ionosphere, and r_0 is the Earth's radius. As is the case for the neutral atmosphere, the ionosphere contains an important turbulent structure. It might be possible to predict the effects of the larger size-scale, slowly varying phenomena if total electron content measurements could be made at the time of the astronomical observations. However, the smaller-scale and more rapidly varying anomalies (which, unfortunately, are as important as the large-scale effects) would require reasonably continuous total electron content measurements over each of the antennas, in the direction of the source of interest. It does not appear that this kind of information will generally be available. A better prospect is self-calibration, since the flux density per antenna beam (or, more strictly, per isoplanatic patch[2]), due to background sources alone, appears to considerably exceed the baseline-based noise at frequencies below 500 MHz; so that if a suitable model is provided and if the array contains a sufficient number of antennas, then self-calibration can be expected to succeed.

4.6. Final phase monitoring.

Observations of calibration sources frequently interspersed with the program sources are used to correct the fast temporal changes of the phase that are due mostly to the atmosphere, and to correct for those residual phase errors which are not removed by other means. Calibrator observations are not needed, or are needed only infrequently, for compact arrays, especially at lower frequencies, where the effects of atmospheric turbulence are minor, and for objects where self-calibration can be expected to succeed.

Generally the closest suitable calibrator should be used, to improve the chances of removing the effects of atmospheric turbulence. Once the observations are complete, the observed calibrator phase can be used to predict the phase corrections which need to be applied to the program sources, using a suitable interpolation function. The type of interpolation function and the convolution interval which should be used depend, to some extent, on the phase behavior and on signal-to-noise considerations. For nearby strong calibrators simple two-point interpolation (straight-line interpolation) is reasonable, whereas for much larger separations or weak calibrators a boxcar or Gaussian average of some suitable length (say 2 hours) may be desirable. If the calibrator phase changes are very large, then care must be taken in the type of interpolation used, to ensure that there is no degradation of the interpolated values. For large changes, self-calibration (if applicable) may be the only recourse.

Note that antenna-based calibration cannot remove baseline-based gain errors. In modern, well-designed synthesis arrays, these errors are generally less than 1° in phase, and 2% in amplitude, so that their effects on images is at a very low level (see Lecture 11 for a more complete discussion). These errors are due to a host of effects, most importantly, delay errors and errors in the phase-shifting networks used to obtain the real and imaginary parts of the observed visibility function. High-dynamic range imaging will be limited by these residual errors, and it now appears possible to remove their effects through calibration. This is further discussed in Lecture 11.

5. Amplitude Calibration

Lecture 6 discusses the sensitivity of radio interferometers. For calibration, the characteristic of most interest is the timescale for changes in antenna sensitivity—that is, for system gain changes. These can be caused by many effects, including:

[2]The isoplanatic patch is the angular distance from some direction over which the atmospheric phase perturbation changes by more than some amount

5.1. Receivers.

Most modern receivers are stable over time periods of at least a few hours. Through on-line monitoring of injected calibration signals, gain fluctuations can be reduced to much less than 1%.

5.2. Antennas.

The antenna gain may be a function of the direction in which the antenna is pointing, because of gravitational deformations of the antenna structure and the surface. For azimuth-elevation antennas, this effect should depend only on elevation. Correction depends upon determining the gain curve, usually measured through observations of strong, unresolved sources.

The temporal behavior of the gain of the antenna is also affected by the sidelobe structure, since this moves with the antenna, and the antenna therefore collects changing amounts of radiation from the ground and sky. The effects are usually more important at lower frequencies, since it is more difficult to design efficient feeds at these frequencies. It is obviously important to minimize these effects through good design. Correction for these changing effects involves monitoring the total system power.

The sensitivity of an antenna is a function of position with respect to the antenna pointing axis. This function (normalized) is called the primary beam pattern $\mathcal{A}(\mathbf{s})$ (see Lectures 1 and 2). This pattern reduces the apparent intensity distribution at all points off the pointing axis, so that the measured visibility function is the inverse Fourier transform of the product $\mathcal{A}(\mathbf{s})I(\mathbf{s})$. Calibration is generally accomplished by dividing the final image by the beamshape. Inaccuracies in this procedure will occur if the beamshape varies over the observation period because of

(1) motion of the primary beam pattern away from the tracking position,
(2) rotation of a non- axisymmetric primary beam on the sky, or
(3) different primary beam shapes among the antennas.

Errors of the first type are generally known as pointing errors, and can be understood and corrected for, given a model of the behavior of the antenna as a function of azimuth and elevation. Generally, a functional dependence of the antenna pointing on these quantities is assumed, and the coefficients are determined by observations of calibrator sources of known position. For example, for azimuth-elevation antennas, a general representation is

$$\begin{aligned} \delta E &= f_1 \cos A + f_2 \sin A + f_3 + f(E) + f_5 \cos E\,, \\ \delta A &= g_1 \sin A \tan E + g_2 \cos A \tan E + g_3 \sec E + g_4\,, \end{aligned} \tag{4–14}$$

where $f_1 = g_1$ is the rotation of the antenna azimuth axis along the meridian from the vertical, $f_2 = -g_2$ is the rotation of this axis perpendicular to the meridian, f_3 is the elevation offset, or encoder error, $f(E)$ is a function describing the refraction, f_5 a sag coefficient describing the effect of gravity on the feed support system, g_3 is the azimuth collimation error (the elevation collimation error is absorbed into f_3), and g_4 is the azimuth offset, or encoder error. These equations are sufficient provided that the pointing errors and the applied corrections are small. The coefficients can be estimated by measurement of the δE and δA offsets for a large number of sources, as functions of azimuth and elevation.

However, after these systematic effects are understood and corrected for, there inevitably are residuals which will be important for some kinds of observations, especially at higher frequencies. There is no simple way to correct for these more random errors, and the best way to handle them is through good antenna design. That is, the tolerable level must be decided upon, and the antennas designed to meet that level. The most important

contributor to these residuals is generally solar heating, and though significant improvement has been made through insulation of antenna support structures, or thermal control, it is certainly preferable to avoid use of these methods—through good design. Active corrections (which could be considered calibration or advanced design) involving measurement by sensitive tiltmeters, combined with computer controlled adjustment, have been contemplated for many arrays, but to the authors' and the editors' knowledge, not implemented anywhere.

The harmful effects of pointing errors are greatly worsened if the antenna beam pattern varies significantly over the object or field being imaged, since in this case the error is both spatially and temporally variable. A common criterion of adequate pointing accuracy is that there be $< \frac{1}{10}$ full-width half-power (FWHP) error in the position, or change of position, during the observation. Observations at high frequencies are especially susceptible to pointing problems, since here the angular pointing errors are often comparable to the antenna primary beam.

The second error is important for azimuth-elevation (az-el) mounted antennas only, since, for these, the beam pattern rotates on the sky through the period of observation. This effect cannot be easily calibrated, so minimization of this problem requires good beam pattern circularity, obtained through good feed design. The effects of this error type are important for sources or fields of view with angular size comparable to the antenna beam. Observations at low frequencies—where there are many sources in the primary beam—require good beam circularity, because of confusion.

The third type of error is a problem only if:

(1) the interferometer comprises more than two elements, and
(2) the radio source, or the field of view being imaged, is comparable to or larger than the FWHP of the largest antenna.

If the array consists of just two elements, the effect can be calibrated by multiplication by the reciprocal of the geometric mean of the antenna power patterns. In most modern arrays, the differences in primary beam shapes are sufficiently small that they can be ignored for all but the very largest sources or fields of view. The effects of non-identical beams are obviously worsened if there are also significant pointing errors

5.3. Atmospheric emission and absorption.

The constituents of the troposphere emit radio noise and absorb incoming radio signals. In the radio wavelength regime, the most important sources of atmospheric attenuation are water vapor and molecular oxygen. The latter constituent dominates near the spectral line transitions at 60 and 118 GHz. At the center of these transitions, the atmosphere is completely opaque. Away from these frequencies, water vapor is the dominant absorber, with absorption maxima near 22 and 185 GHz. The former transition is the only important absorber for frequencies below approximately 50 GHz. Zenith opacity at the center of the line rarely exceeds 1 dB (corresponding to a brightness temperature of approximately 40 K), and for the commonly used 20 and 6 cm bands water vapor absorption is of order 1%. Important transitions due to other molecular species are found at frequencies exceeding 100 GHz. For a discussion, see Chapter 2.3 in Meeks (1976). The antenna temperature due to the source and sky emission can be written (assuming a simple slab model of uniform density for the atmosphere),

$$T_s e^{-\tau_\nu \sec z} + T_{\rm atm}\left(1 - e^{-\tau_\nu \sec z}\right), \qquad (4\text{–}15)$$

where τ_ν is the opacity at frequency ν. Here, the source signal T_s is attenuated, and we see that the atmosphere at temperature $T_{\rm atm}$ can contribute significantly to the antenna

temperature. At 22 GHz it is possible for atmospheric emission to be comparable to the receiver temperature T_{rx}, for very low-noise systems.

Below 10 GHz, tropospheric absorption can generally be neglected. Ionospheric absorption is appreciable only at very low frequencies, and is rarely of importance above 50 MHz.

5.4. Correlation noise.

Lecture 3 considered the loss of sensitivity involved with digital correlation techniques. Of especial importance in calibration is the variation of this degradation with degree of correlation (known commonly as the 'Van Vleck correction'). The change in correlation can be corrected for, on-line (as discussed in Lecture 3), although the loss of sensitivity cannot be.[1]

5.5. Techniques of calibration.

Many of these elevation- and time-dependent effects, except for atmospheric absorption, can be calibrated by measuring the system temperature directly and, preferably, by continuously monitoring it. The usual procedure involves injecting the signal from a stable noise source into the front end of a receiver and synchronously detecting it later on. The use of system temperature corrections can adequately correct for atmospheric absorption only when the equivalent source temperature is much greater than the receiver temperature.

In order to correct for temporal variations which are not removed by system temperature corrections it is necessary to use calibrator observations which are interspersed among the program source observations. The sources used for this purpose are usually the phase calibrators.

The flux densities of most good phase calibrators vary with time and cannot be used for absolute flux density calibration. Therefore a small number of "non-variable" calibrators is used to calculate the flux densities, first of the phase calibrators and then of the program sources.[2]

Self-calibration is effective in adjusting the relative gains of the antennas, as discussed in Lectures 9 and 11, for observations of strong radio sources. The absolute gains cannot be recovered unless the total flux of the source is known *a priori*.

6. Spectral Line Calibration

Assuming that the calibrations outlined in Sections 4 and 5 have been completed, and re-inserting the frequency dependence in the calibration equation, we find

$$A'_{ij}(t,\nu)e^{i\phi'(t,\nu)} = b_i(\nu)b_j(\nu)e^{i(\beta_i(\nu)-\beta_j(\nu))}A_{ij}(t,\nu)e^{i\phi_{ij}(t,\nu)}\,, \qquad (4\text{–}16)$$

where $b_i(\nu)$ denotes the bandpass amplitude gain correction for antenna i and $\beta_i(\nu)$ the bandpass phase correction for antenna i.

6.1. Bandpass calibration.

The channel-to-channel gain variations, both in amplitude and phase, are caused by filters used to limit the bandpasses and by instrumental effects, such as reflections in the waveguide and between subreflectors and receivers. These variations can be calibrated through observations of strong sources.

[1] At the VLA, these corrections are not yet implemented.

[2] At the VLA, the flux density scale is tied to the source 3C 286, whose flux density is assumed to be equal to that given by Baars *et al.*, 1977.

The normal procedure for determining these corrections is to observe, over the same frequency band as the program source, a strong calibrator that has no spectral line emission or absorption within the observing band. This correction is very important when small line-to-continuum ratios are to be observed and high channel-to-channel dynamic range is needed (e.g., in recombination line observations). The procedure usually requires that the bandpass be stable for the duration of the observing run.[3]

To correct for amplitude variations it is possible, using the program source itself, to obtain an estimate of $b_i(\nu)$ (only) from measurements of the autocorrelation function of each IF signal. Such bandpass normalization works well when there are no strong lines (strong relative to the system temperature), either in absorption or emission, within the passband. This method does not correct for phase variations across the band, and it is inadequate when high dynamic range is required.

7. Polarization Calibration

7.1. Polarization mixing.

Recall from Lecture 1 that the complete state of the radiation field is the superposition of two orthogonal vector quantities. Polarimetry measurements require two orthogonally polarized feeds. In an ideal antenna, these feeds respond solely to the two orthogonal propagation modes. There are four combinations, or correlations, which can be formed from the signals, and these combinations can be described in terms of the four (real) Stokes parameters; I, describing the total intensity, V, describing the circularly polarized intensity, and Q and U, describing the linearly polarized intensity. These quantities are obtained through combinations of these correlations. That is, the measured spatial coherence functions transform to the following combinations of Stokes' parameters:

(1) For circularly polarized signals,

$$\begin{pmatrix} V_{RR} \\ V_{RL} \\ V_{LR} \\ V_{LL} \end{pmatrix} = \begin{pmatrix} e^{-i(\chi_1-\chi_2)} & 0 & 0 & e^{-i(\chi_1-\chi_2)} \\ 0 & e^{-i(\chi_1+\chi_2)} & ie^{-i(\chi_1+\chi_2)} & 0 \\ 0 & e^{i(\chi_1+\chi_2)} & -ie^{i(\chi_1+\chi_2)} & 0 \\ e^{i(\chi_1-\chi_2)} & 0 & 0 & -e^{i(\chi_1-\chi_2)} \end{pmatrix} \begin{pmatrix} V_I \\ V_Q \\ V_U \\ V_V \end{pmatrix}, \quad (4\text{–}17)$$

where χ_1 and χ_2 denote the parallactic angles of the feeds ($\chi_1 \approx \chi_2$ for identically mounted feeds on closely spaced elements) and where $V_I = \mathsf{F}^{-1}I$, $V_Q = \mathsf{F}^{-1}Q$, etc. The antenna parallactic angle is related to latitude ϕ, source hour angle h, and declination δ, by $\tan\chi = \cos\phi\sin h/(\sin\phi\cos\delta - \cos\phi\sin\delta\cos h)$.

Assuming equal parallactic angles, $\chi_1 = \chi_2 = \chi$, Equation 4–17 simplifies to

$$\begin{pmatrix} V_{RR} \\ V_{RL} \\ V_{LR} \\ V_{LL} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & e^{-2i\chi} & ie^{-2i\chi} & 0 \\ 0 & e^{2i\chi} & -ie^{2i\chi} & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} V_I \\ V_Q \\ V_U \\ V_V \end{pmatrix}. \quad (4\text{–}18)$$

[3]At the VLA, this is quite often the case (but not always!) for time periods not exceeding 6–8 hours.

(2) For linearly polarized signals,

$$\begin{pmatrix} V_{VV} \\ V_{VH} \\ V_{HV} \\ V_{HH} \end{pmatrix} = \begin{pmatrix} \cos(\chi_1 - \chi_2) & \cos(\chi_1 + \chi_2) & \sin(\chi_1 + \chi_2) & -i\sin(\chi_1 - \chi_2) \\ -\sin(\chi_1 - \chi_2) & \sin(\chi_1 + \chi_2) & -\cos(\chi_1 + \chi_2) & -i\cos(\chi_1 - \chi_2) \\ \sin(\chi_1 - \chi_2) & \sin(\chi_1 + \chi_2) & -\cos(\chi_1 + \chi_2) & i\cos(\chi_1 - \chi_2) \\ \cos(\chi_1 - \chi_2) & -\cos(\chi_1 + \chi_2) & -\sin(\chi_1 + \chi_2) & -i\sin(\chi_1 - \chi_2) \end{pmatrix} \begin{pmatrix} V_I \\ V_Q \\ V_U \\ V_V \end{pmatrix}. \tag{4-19}$$

Assuming equal parallactic angles, this simplifies to

$$\begin{pmatrix} V_{VV} \\ V_{VH} \\ V_{HV} \\ V_{HH} \end{pmatrix} = \begin{pmatrix} 1 & \cos 2\chi & \sin 2\chi & 0 \\ 0 & \sin 2\chi & -\cos 2\chi & -i \\ 0 & \sin 2\chi & -\cos 2\chi & i \\ 1 & -\cos 2\chi & -\sin 2\chi & 0 \end{pmatrix} \begin{pmatrix} V_I \\ V_Q \\ V_U \\ V_V \end{pmatrix}. \tag{4-20}$$

Unfortunately, an antenna and feed do not respond solely to a single propagation mode. By diverse means, some signal from one mode contaminates the other, so that the polarization matrices become more complicated. The 'crosstalk' is generally described by D, describing the fraction of one polarization mode which leaks into another. Consider first circularly polarized feeds. If E_R and E_L are the circularly polarized signals which would be measured with an ideal system, the actual signals, v_R and v_L, are $v_R = E_L e^{-i\chi} + D_R E_L e^{i\chi}$ and $v_L = E_L e^{i\chi} + D_L E_R e^{-i\chi}$ (Bignell, 1977). For the linearly polarized case, we have $v_H = E_H \cos(\chi + \theta) - E_V \sin(\chi + \theta) + D_H(E_H \sin(\chi + \theta) + E_V \cos(\chi + \theta))$, and $v_V = E_H \sin(\chi + \theta) + E_V \cos(\chi + \theta) + D_V(E_H \cos(\chi + \theta) - E_V \sin(\chi + \theta))$, where θ is the position angle of the vertical feed. Since Stokes' parameter I is generally very much greater than Q, U, or V, and the leakage terms are also small, only products between Q, U, V, and the D's with I need be retained.

The cross-handed responses, with the above approximations, assuming equal parallactic angles, and assuming the antenna-based calibration has been performed, become:

(1) For circularly polarized feeds,

$$\begin{aligned} V_{RL} &= e^{-2i\chi}(V_Q + iV_U) + (D_{R1} + D_{L2}^*)V_I, \\ V_{LR} &= e^{2i\chi}(V_Q - iV_U) + (D_{L1} + D_{R2}^*)V_I. \end{aligned} \tag{4-21}$$

(2) For linearly polarized feeds,

$$\begin{aligned} V_{VH} &= V_Q \sin 2\chi - V_U \cos 2\chi - iV_V + (D_{V1} + D_{H2}^*)V_I, \\ V_{HV} &= V_Q \sin 2\chi - V_U \cos 2\chi + iV_V + (D_{H1} + D_{V2}^*)V_I. \end{aligned} \tag{4-22}$$

For a more explicit derivation, see Bignell (1977, 1986). Note that in all the above equations, observation of an unresolved source at the phase-tracking center allows replacement of V_I, V_Q, V_U, and V_V with I, Q, U, and V, respectively.

7.2. Calibration of the leakage terms.

The leakage terms can be calibrated by observations of unresolved calibrators. There are two approaches:

The first is to observed a source of known polarization. It is important that the polarized flux be high, so that the leakage terms can be accurately determined. This requirement, when added to the need for an unresolved source, greatly limits the choice of sources. Of these, the preferred sources are 3C 286 and 3C 138. The measurement of the D's is straightforward.

The second technique applies only to altitude-azimuth mounted antennas. For these, the antenna beam rotates on the sky during the course of the observations. This rotation causes the phase of the source polarization to vary, while that due to the antenna polarizations remain constant. Observations over a suitable range in parallactic angle allow a straightforward separation of the two contributions. In this technique, a polarized calibrator is not required. However, a high total flux density is desirable—the signal in the cross-hand channels is augmented by the total flux multiplied by the crosstalk term, as it allows a more accurate determination of these terms. This technique does not calibrate the position angle of the polarized flux density (corresponding to the phase difference between the orthogonally polarized feeds). To do this requires a short observation of a polarized source of known position angle.

After determination of the D's, the visibility data may be corrected through by applying the above equations. For these techniques to be effective, it is desirable that the change of the polarization constants be kept to a minimum.[1] This allows degrees of polarization of order 0.1% to be determined. A serious limitation for polarimetry of extended sources is that the instrumental polarization varies significantly over the primary beam, and, for azimuth-elevation antennas, both this pattern and the antenna primary pattern rotate on the sky over the observing period. (For equatorially mounted antennas, the effect is spatially constant and, if the antennas are all described by the same D's, can be removed in the image plane).

7.3. Faraday rotation.

The presence of a magnetic field in a plasma causes the plasma's index of refraction to be different for right- and left-circularly polarized radio waves, with the result that a linearly polarized wave has its plane of linear polarization rotated as it propagates through the ionosphere. The amount of rotation at wavelength λ is given by

$$E\lambda^2 \int_0^L N_e(l) H_\parallel(l)\, dl = \text{R.M.}\, \lambda^2 , \qquad (4\text{–}23)$$

where $N_e(l)$ denotes the electron density, $H_\parallel(l)$ denotes the component of the magnetic field parallel to the line of sight, E is a constant equal to 2.62×10^{-17} in c.g.s. units, and L the depth of the ionosphere, measured along the line of sight. The quantity R.M. is called the rotation measure. Typically the ionospheric rotation measure is about 1 rad m^{-2}, with values reaching as high as 15 to 20 rad m^{-2} during solar maximum. Corrections for Faraday effects are usually required for observations at wavelengths longer than 18 cm, but can occasionally be important at wavelengths as short as 10 cm. The use of models and measurements of the total electron content can correct for moderate-size rotations caused by the slowly varying diurnal component, but it cannot correct for anomalies. These may be calibratable through observations of a nearby polarized calibrator, or, perhaps, through polarization self-calibration, presuming the source has sufficient polarized flux density that variations in the apparent position angle can be monitored.

[1] At the VLA, it is found that the variations, over an 8-hour period, are less than 0.5%.

7.4. Limitations of polarization calibration.

In most modern radio astronomy antennas the D's remain constant to within 0.5% over an eight-hour period. These variations tend to increase with frequency.

The D's are not constant over the beam. The spatial variation in D is less than 0.5% only over an area of the beam, centered on the primary lobe of the antenna, with radius about 10% the FWHP. As long as the extended emission is within this region of the antenna beam, the accuracy of the polarization calibration is reasonably good. Polarized emission extending beyond this point will be progressively less accurately calibrated, with uncertainties possibly as large as a few percent near the half-power point of the beam.

The other important limitation is that the phase difference between the orthogonal modes[1] must remain constant in time. Changes in this instrumental quantity will alter the observed source position angle of polarized flux density, and lower the polarized flux density. This difference can be monitored through observations of strong calibrators.

8. Data Editing

The final step in data calibration is to identify and delete data which are irreversibly corrupted. This process requires human judgement, which can only be gained with experience. We will summarize some causes, and effects on the data, and give some guidelines for identification of affected data.

8.1. Interference.

Communications signals and radars associated with satellites, aircraft, and ground-based transmitters, as well as signals generated by the local oscillator system, can increase the system noise or cause erratic behavior in the measured visibility amplitudes and phases. This is especially true at low frequencies (< 2 GHz) and on short baselines. Spectral line observations using narrow bandwidths are particularly susceptible to interference, compared to observations using wider bandwidths, due to the smaller amount of "dilution" (dilution of the interfering signal by uncontaminated signal in the rest of the band). On the other hand, observing in spectral line mode will allow efficient removal of interfering signals if they occur in a small number of channels (and if the interesting signal does not also occupy these channels). Interference will show up in images in various ways, commonly as stripes across the image. Efficient techniques of removal are available—see Lectures 10 and 11 for details.

8.2. Shadowing and crosstalk.

When the antennas are close together one may "look" into the back of another.[2] A related problem, notable on short baselines in general, but which is especially severe under conditions of shadowing, occurs when signals radiated by one antenna (say, by the local oscillator) are picked up by another. This causes a false correlation to occur, and is generally known as 'crosstalk'.

Shadowing changes the baselines of the antennas involved, reduces the antenna gain, and distorts the primary beam of the antenna that is shadowed. The reduction of the antenna gain can be corrected by a factor based on the geometrical blockage. However, it must be emphasized that this factor is only correct for the center of the antenna beam (and only at high frequencies, where diffraction effects can be neglected). Application of this correction only makes sense if the region of interest is small compared to the antenna primary power pattern. The other effects (beam distortion, and baseline offset) can not be

[1] at the VLA, the so-called "A–C or B–D phase difference"

[2] This is common in the C and D configurations of the VLA.

simply corrected, and are most important to imaging of large fields. Given that observations affected by shadowing are almost always observations of large objects, the safest procedure is to delete all affected data. Crosstalk effects are not correctable, and afflicted data must be deleted.

8.3. Strong sources in the sidelobes of the antennas.

The presence of very strong sources in the sidelobes of the antenna beam can significantly affect the observations. This effect is most notable in observations taken near the 'Big Three', the Sun, Cassiopeia A, and Cygnus A, especially at lower frequencies. The use of wide bandwidths is effective in suppressing this type of interference—however, spectral line observations taken in daytime at wavelengths longer than 10 cm will nearly always show the effect of solar contamination.

8.4. Identification and deletion of bad data.

The principal difficulty is in identifying the data which should be deleted. This is best accomplished by examining the record-to-record consistency of the visibility amplitudes, since jumps on this timescale cannot occur in good data, unless the source is time-variable on this scale (e.g., solar and stellar flares). Various schemes have been devised to list data for quick perusal; their effectual use is a matter of experience. A useful way to spot problems is to examine the r.m.s. statistics of individual correlators for each scan, and from scan to scan. A single discrepant value will greatly increase the r.m.s. value. Perusal of matrix listings of this quantity helps to quickly identify questionable correlators, whose data can then be listed for detailed editing. Another commonly used method is the baseline-time display, showing visibility data on a TV monitor. Unusual values can quickly be identified.

Further details on calibration techniques can be found in Lectures 9 and 11.

REFERENCES

Baars, J. W. M., Genzel, R., Pauliny-Toth, I. I. K., and Witzel, A. (1977), "The absolute spectrum of Cas A; an accurate flux density scale and a set of secondary calibrators", Astron. Astrophys., **61**, 99–106.

Bignell, R. C. (1977), "Proposal for gain and polarization calibration", VLA Computer Memorandum No. 136.

Bignell, R. C. (1986), "Linear polarization observations using the VLA", in preparation.

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

Meeks, M. L. (1976), *Methods of Experimental Physics, Part B: Radio Telescopes*, Volume 12, Academic Pr., New York.

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

Wade, C. M. (1976), "Effect of the general relativity deflection on the apparent position of an object", VLA Scientific Memorandum No. 122.

5. 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. 2–27),

$$\mathcal{A}(l,m)I(l,m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{5–1}$$

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 4), 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 5–1.

Equation 5–1 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 14). 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 DFT's.

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 5–1, 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)

$$I^D(l,m) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u,v)V'(u,v)e^{2\pi i(ul+vm)}\,du\,dv\,. \tag{5–2}$$

Here, S denotes the u-v sampling function and V' the observed visibility; the prime indicates that the visibility data are noise-corrupted measurements. (For conciseness, I^D has been left unprimed, but it too is noise-corrupted whenever V is.)

[1] 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 4.

1.1. The 'direct Fourier transform' and the FFT.

Either of two methods is commonly used to numerically approximate the Fourier transform in Equation 5–2. The first, called the 'direct Fourier transform' method,[2] approximates $I^D(l,m)$ by brute-force evaluation of the sum

$$\frac{1}{M}\sum_{k=1}^{M} V'(u_k,v_k)e^{2\pi i(u_k l+v_k m)}\,. \tag{5–3}$$

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 $O(N^4)$, 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 $O(M)$. 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 $O(N^4)$! This saves much computing time for large databases, and large N especially, if an economical method of interpolation is used. However, for making small images (i.e., for N small) from small databases (M small), the 'direct Fourier transform' may be faster than the combination of gridding and FFT.

In the following Sections we first discuss weighting and selection of u-v data and how it affects the resulting images. This applies no matter how the Fourier transform is approximated. Then we touch upon the problems that are introduced by gridding the data to permit use of the FFT—the problems of aliasing and correction for gridding.

2. The Sampling Function, and Weighting the Visibility Data

The sampling function S and its Fourier transform, the synthesized beam B, were introduced in Lecture 1. In practice, the data are variously weighted, according to their reliability and to control the shape of the synthesized beam.

2.1. The sampling function.

S is a 'generalized function', or 'distribution', which may be expressed in terms of the two-dimensional Dirac delta function, or 'δ-distribution',

$$S(u,v) = \sum_{k=1}^{M} \delta(u-u_k, v-v_k)\,. \tag{5–4}$$

---

[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.

It is useful to introduce a second generalized function, called the *sampled visibility function* or, alternatively, the *u-v measurement distribution*,[1]

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

That is, $V^S = SV'$. Let F denote the Fourier transform operator. Equation 5–2 can be rewritten

$$I^D = \mathsf{F}V^S = \mathsf{F}(SV') \,. \tag{5-6}$$

By the *convolution theorem*, which says that the Fourier transform of a product of functions is the convolution of their FT's (see, e.g., Bracewell 1978),

$$I^D = \mathsf{F}S * \mathsf{F}V' \,, \tag{5-7}$$

where $*$ denotes convolution. For a point source of unit strength, centered at position (l_0, m_0), $|V'(u,v)| \equiv 1$, and $\mathsf{F}V'$ 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 $B = \mathsf{F}S * \delta = \mathsf{F}S$. Equation 5–7 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 5–3, is *exactly* I^D. That is to say, that—assuming δ-function sampling— $I^D(l,m)$, as defined by Equation 5–2, is given exactly by a discrete summation, Expression 5–3, and that Equation 5–7 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.

2.2. Weighting functions for control of the beam shape.

In analogy to Equation 5–4, a *weighted sampling function*, or *weighted sampling distribution*, can be written as

$$W(u,v) = \sum_{k=1}^{M} R_k T_k D_k \delta(u-u_k, v-v_k) \,. \tag{5-8}$$

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

$$V^W(u,v) = \sum_{k=1}^{M} R_k T_k D_k \delta(u-u_k, v-v_k) V'(u_k, v_k) \,. \tag{5-9}$$

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 kth visibility

[1]Note that the visibility measurements are not, in actuality, point samples of the inverse Fourier transform of the modified sky brightness AI, 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 8 and 16.

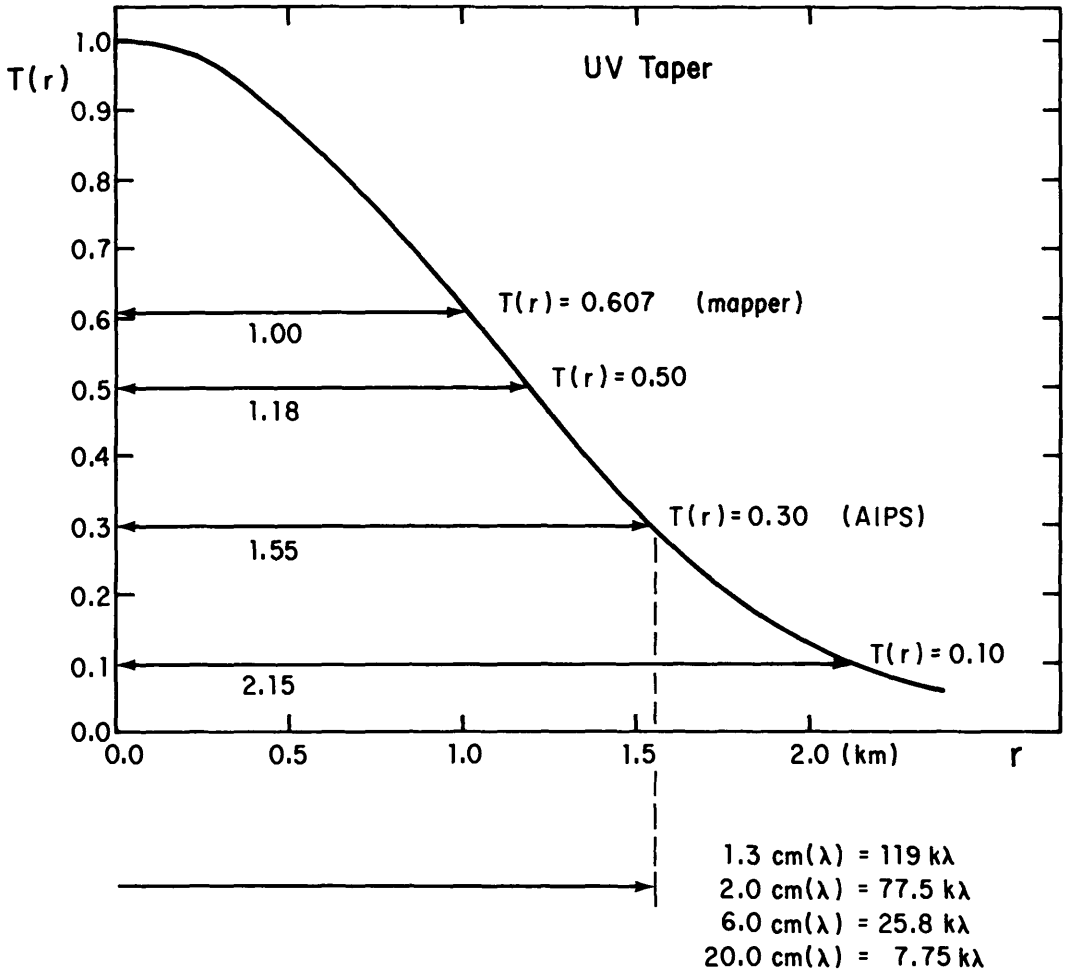


Figure 5–1. A Gaussian u-v taper with dispersion $\sigma = 1$ km.

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.,

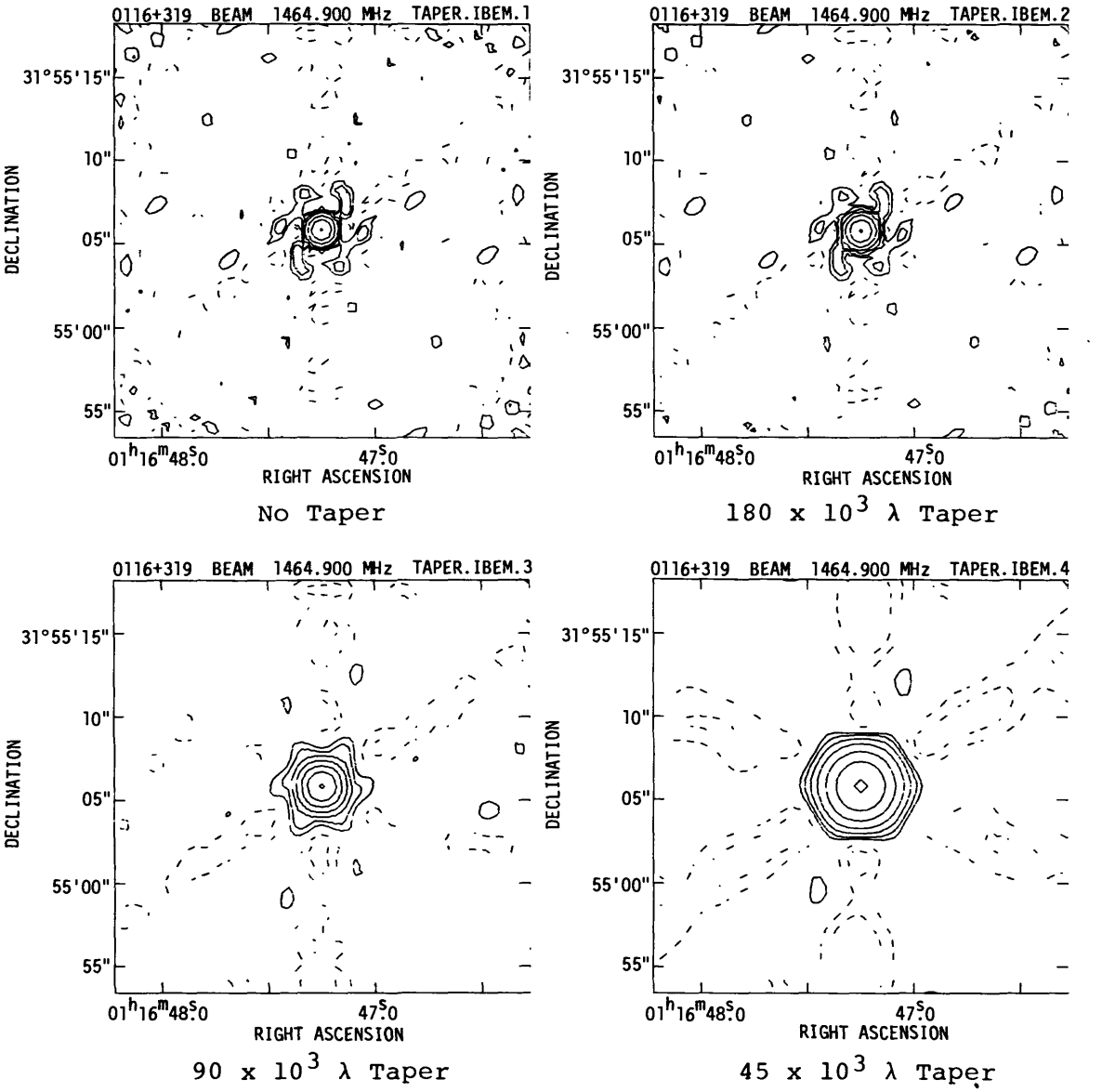


Figure 5–2. The effect of a Gaussian taper on the point source response of a VLA snapshot in the A configuration at 20 cm wavelength. As a narrower Gaussian taper (i.e., a heavier tapering) is applied, the half-power width of the point spread function increases and the inner sidelobes are reduced.

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. 5–1).

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 $\theta_{\rm HPBW} = 0.37/\sigma$ with θ in radians and σ in wavelengths. Translated into common units, $\theta_{\rm HPBW} = 0.77\lambda({\rm cm})/\sigma({\rm km})$ 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 approxima-

tion, for which $\theta_{\rm HPBW} = 2.0\lambda({\rm cm})/a({\rm km})$ arc-seconds, where a is the radius of the aperture. Real-life observational geometries and u-v coverages often produce larger $\theta_{\rm HPBW}$ and, frequently, elongated beams. Examples of the VLA point source response with different u-v tapers are shown in Figure 5–2.

Instead of de-emphasizing data near the outer boundary of the u-v coverage, it is sometimes desirable to downweight the data near $u = v = 0$. An undersampled large scale emission region may introduce large undulations in image intensity that are hard to remove. These can present a problem for detecting a weak point source embedded within a region containing extended emission. Minimum u-v limits and other forms of downweighting are often used to diminish the effect of these low spatial frequency data points.

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:

$$D_k = 1\,, \qquad \text{called } \textit{natural weighting},$$

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

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 kth 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 kth datum (gridding is discussed later).[1] This produces a beam specified largely by the tapering function T.

Sometimes, especially in the VLA "snapshot" mode of observing (see Lecture 16), 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 5–3.

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

[1] In the AIPS implementation, these blocks are called "uniform weight boxes".

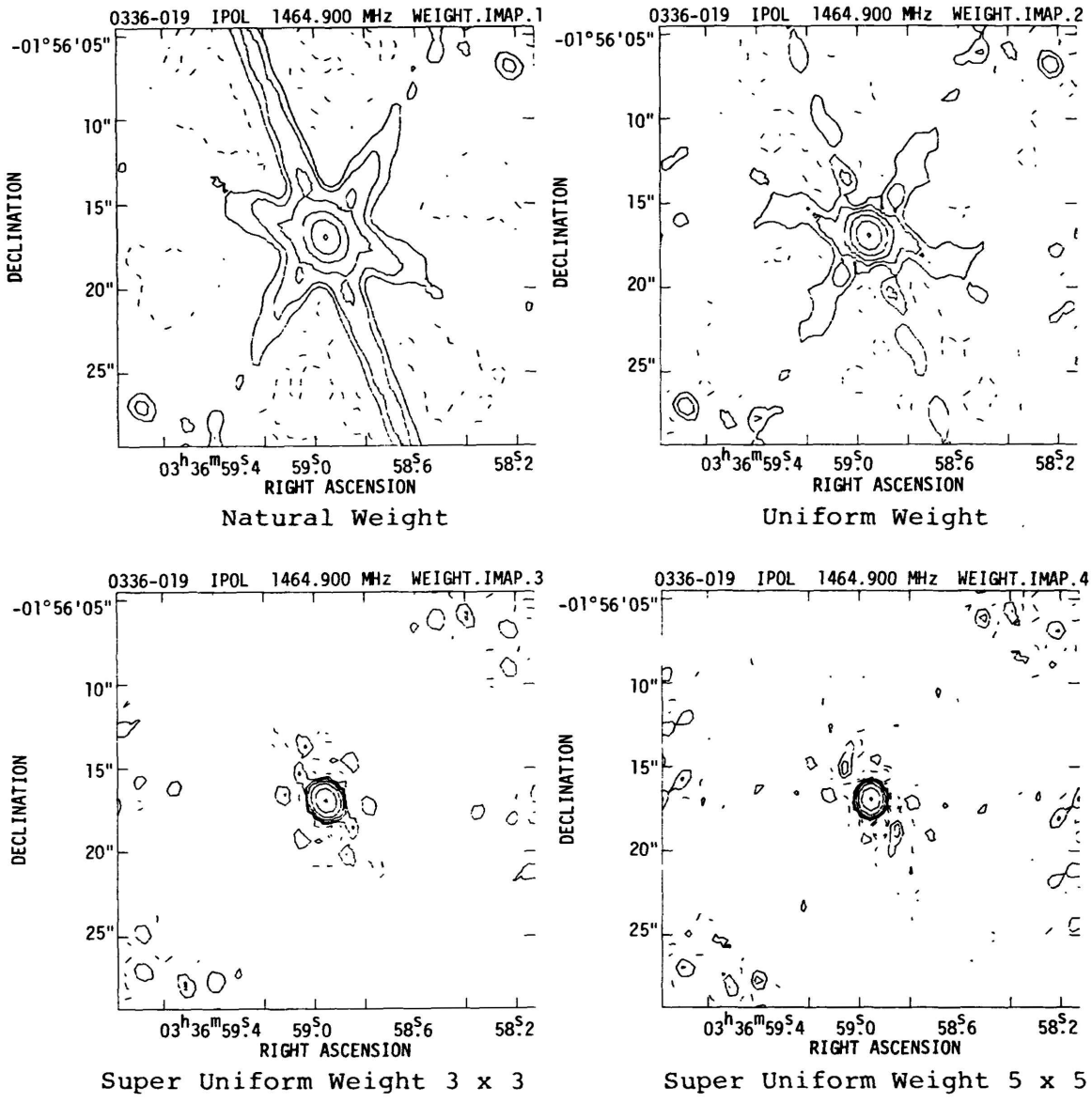


Figure 5–3. The effect of different weighting functions on a VLA "snapshot" image of a point source.

visibility values at the grid points, based on the observed values.[2] 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.

[2]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.

3.1. Gridding by convolution.

The idea is to convolve the weighted, sampled measurement distribution V^W with some suitably chosen function C, and to sample this convolution at the 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

$$\sum_{k=1}^{M} C(u_c - u_k, v_c - v_k) V^W(u_k, v_k) . \tag{5-10}$$

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 5–10 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

$$V^R = R\left(C*V^W\right) = R(C*(WV')) , \tag{5-11}$$

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 ⨿) by

$$R(u,v) = \text{⨿}(u/\Delta u, v/\Delta v) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \delta(j - u/\Delta u, k - v/\Delta v) . \tag{5-12}$$

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

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

Applying the convolution theorem to Equation 5–11, $\tilde{I}^D$ is given by

$$\tilde{I}^D = \mathsf{F}R * \left[(\mathsf{F}C)\left(\mathsf{F}V^W\right)\right] = \mathsf{F}R * \left[(\mathsf{F}C)\left(\mathsf{F}W * \mathsf{F}V'\right)\right] . \tag{5-13}$$

(Please refer now to Fig. 5–4 for a graphical interpretation of Eq. 5–13 and for an illustration of the operations that are described in the remainder of this Section.) ⨿ is its own Fourier transform; R behaves similarly—by the dilation property of the FT (see Sec. 4.1),

$$(\mathsf{F}R)(l,m) = \Delta u \Delta v \text{⨿}(l\Delta u, m\Delta v) = \Delta u \Delta v \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \delta(j - l\Delta u, k - m\Delta v) . \tag{5-14}$$

One effect of the resampling is to make $\tilde{I}^D$ 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) $\tilde{I}^D$. 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 $\tilde{I}^D$. Let P denote the region over which $\tilde{I}^D$ 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.

Figure 5–4 *(pp. 76–77).* 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 5–4. 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 5–5. 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.

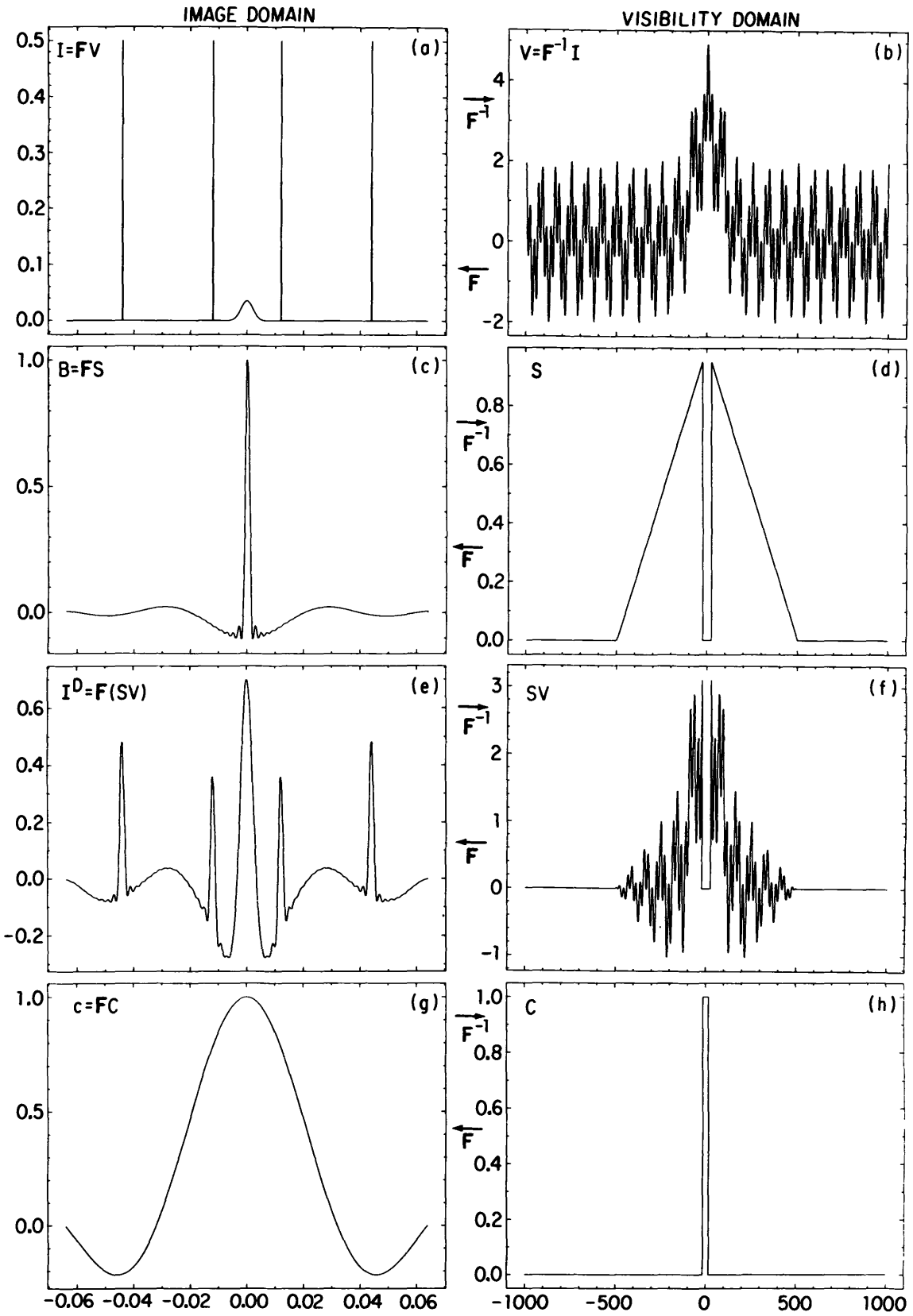


Figure 5–4 *(Caption is on p. 75). (Continued on next page.)*

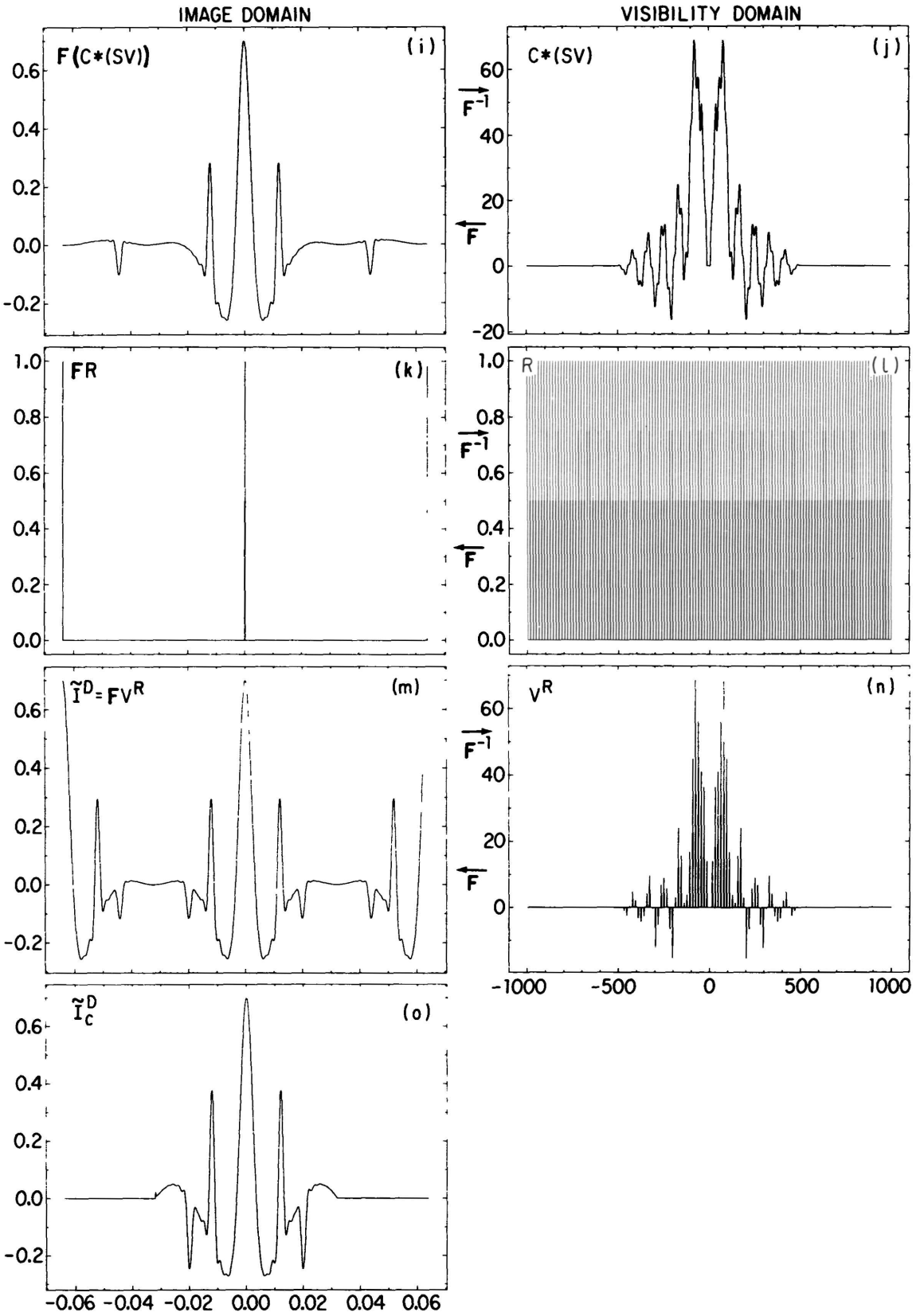


Figure 5–4 *(Continued).*

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 $\tilde{B}^D$.

Normally, $\tilde{I}^D$ and $\tilde{B}^D$ are corrected for the effect of the gridding convolution by pointwise division by c: The so-called "grid-corrected" image is given by

$$\tilde{I}_c^D(l,m) = \frac{\tilde{I}^D(l,m)}{c(l,m)}, \tag{5–15}$$

and the "grid-corrected" beam by

$$\tilde{B}_c^D(l,m) = \frac{\tilde{B}^D(l,m)}{c(l,m)}. \tag{5–16}$$

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 5–11 and $\mathsf{F}R$ in Equation 5–13. 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.[1]

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

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 $\tilde{I}_c^D$ is a good approximation to I^D—that is, Equation 5–13 becomes

$$\tilde{I}_c^D = \mathsf{F}W * \mathsf{F}V', \tag{5–17}$$

—and then the usual convolution relation between I^D, B, and I is approximately valid with $\tilde{I}_c^D$ and $\tilde{B}_c^D$ substituted for I^D and B, respectively. Note, however, that $\tilde{B}_c^D$ is usually computed only over a region of the same dimensions as the image $\tilde{I}_c^D$. For this reason, the deconvolution algorithms (described in Lecture 7) 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 5–13 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

[1]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").

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 2.

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 $O(Mm^2)$ arithmetic operations are required for gridding. These functions are described below; for more discussion see Schwab (1978):

(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.* $C(u) = \frac{\sin(\pi u/w\Delta u)}{\pi u/w\Delta u}$. 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.* $C(u) = \exp\left(-\left(\frac{|u|}{w_1 \Delta u}\right)^{\alpha}\right) \frac{\sin(\pi u / w_2 \Delta u)}{\pi u / w_2 \Delta u}$. Typically[1] $m = 6$, $w_1 = 2.52$, $w_2 = 1.55$, $\alpha = 2$; 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* concentration ratio: for given α, $\int_P w(l)|c(l)|^2 \, dl \Big/ \int_{-\infty}^{\infty} w(l)|c(l)|^2 \, dl$ is maximized, where $w(l) = |1 - 2l\Delta u|^{\alpha}$. Choosing $\alpha > 0$ gives higher alias rejection near the center of the image, at the expense of alias rejection near the edges. ψ_{00} is its own FT, in the sense that if you truncate it as done here, and then take the FT, what you get back is ψ_{00}. Similarly, the other $\psi_{\alpha 0}$ are finite Fourier self-transforms, in the sense that if you so truncate one, weight it, and transform it, what you get back is $\psi_{\alpha 0}$. $\psi_{\alpha 0}$ is used at the VLA, with $m = 6$ and $\alpha = 1$ being typical. See Schwab (1984) for further discussion and additional references.

Figure 5–5 shows the Fourier transforms of two typical gridding convolution functions, normalized to unity at $l = 0$. The abscissa on this plot is in units of image half-widths, $\eta = 2l\Delta u$, so that $\eta = \pm 1$ at the image edges. The image response is suppressed at the edge for both functions, however the exp × sinc function is flatter inside P, and drops much faster past the image edge. The aliased response can, of course, be negative, producing an apparent "hole" in the image. The plots in Figure 5–6 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 × sinc or the spheroidal. The exp × sinc has somewhat smaller corrections and, thus smaller errors (due to round-off noise and to violation of the assumptions that make the grid correction valid), near the image edges, while the spheroidal has better rejection beyond the image edge (Schwab 1984).

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. 5–7). 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.

[1] 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}.

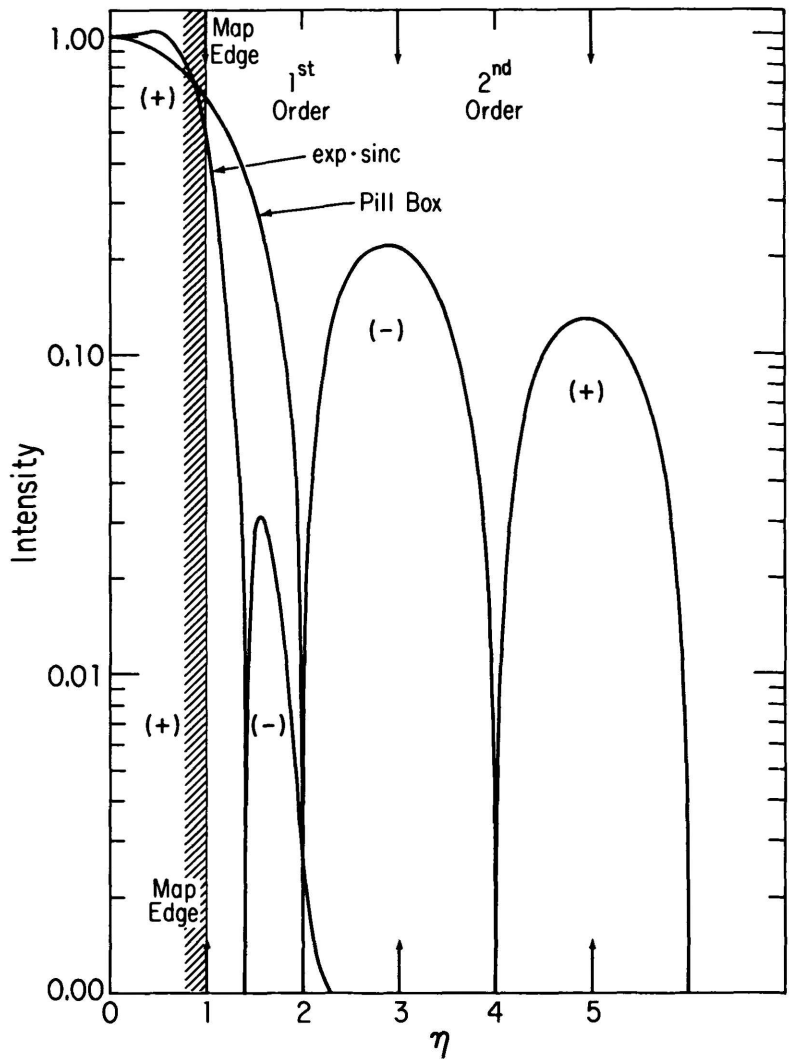


Figure 5–5. The response to a source, as a function of distance from the image center, for two typical u-v convolving functions.

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

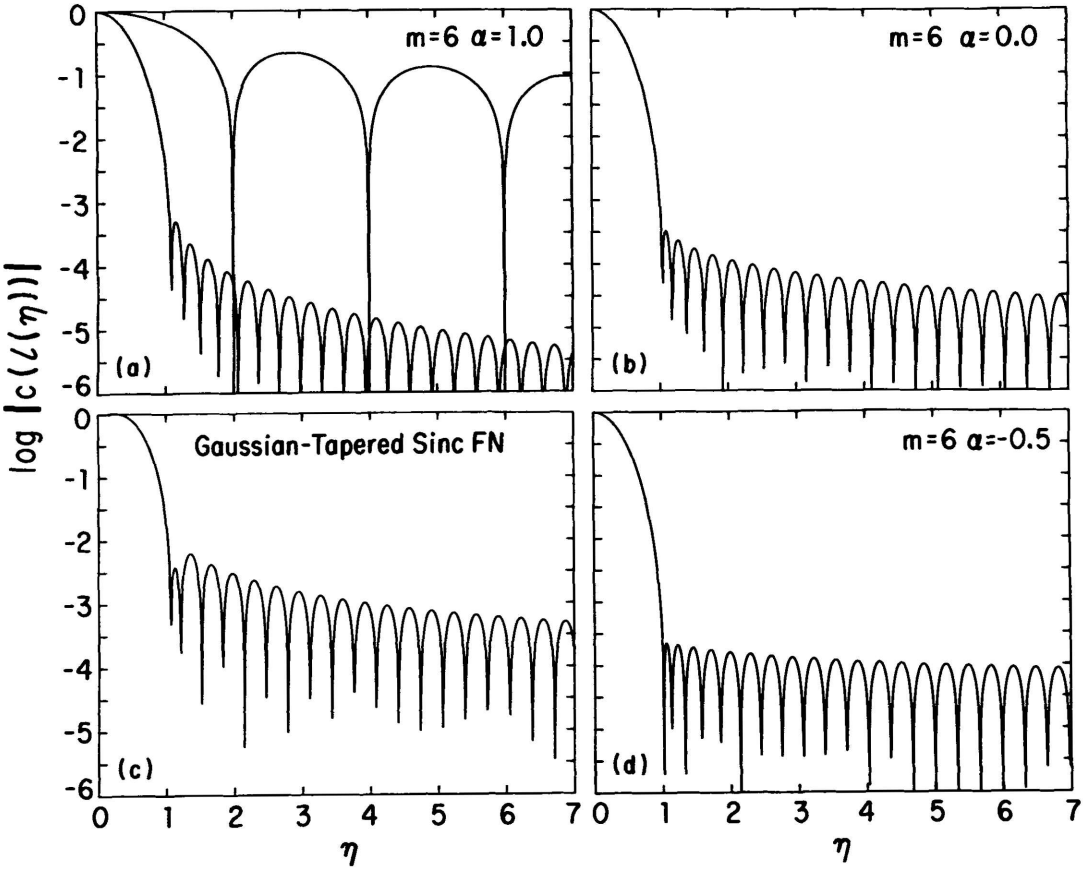


Figure 5–6. For some typical gridding convolution functions C, plots of the absolute value of the Fourier transform of C. **(a)** The spheroidal function ψ_{10}, for $m = 6$, compared with the pillbox function ($m = 1$); **(b)** the "prolate spheroidal wave function" ψ_{00}, $m = 6$; **(c)** an optimised Gaussian-tapered sinc function, $m = 6$; **(d)** the spheroidal function $\psi_{-\frac{1}{2},0}$, $m = 6$. Adapted from Schwab (1984).

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 7) 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

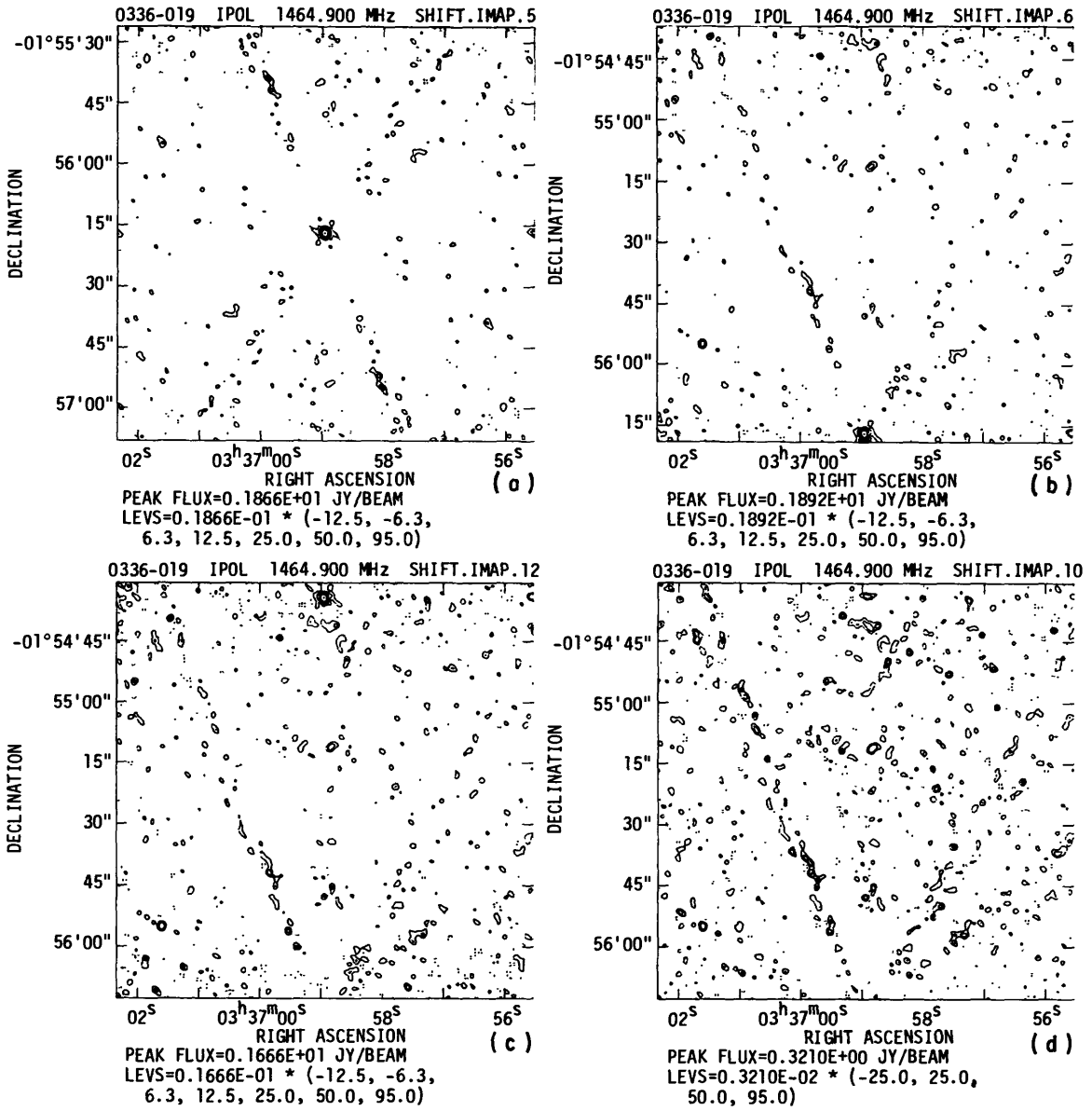


Figure 5–7. The effects of aliasing: (a) a point source at the field center; (b) the same source near the image edge; (c) the source below the lower image edge appears as an aliased image at the upper image edge, with pillbox convolution; (d) with exp × sinc convolution, the aliased source is greatly attenuated, but the sidelobe response remains the same.

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 5–8, 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 FT, which, if inserted into Equation 5–13, 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 FR, but one with a very long period, $100/\Delta u$. The grid correction given by Equation 5–15 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 (1984) gives cheap and accurate rational approximations to the spheroidal functions; the step function approximation is unnecessary.)

4.3. Non-coplanar baselines.

In Equation 5–1 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 5–1 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

$$V(u,v,w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\mathcal{A}(l,m)I(l,m)}{\sqrt{1-l^2-m^2}} e^{-2\pi i\left(ul+vm+w\left(\sqrt{1-l^2-m^2}-1\right)\right)}\, dl\, dm\,. \qquad (5\text{–}18)$$

This can be rewritten as

$$V(u,v,w)e^{-2\pi iw} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\mathcal{A}(l,m)I(l,m)}{\sqrt{1-l^2-m^2}}\, \delta(n-\sqrt{1-l^2-m^2}) e^{-2\pi i(ul+vm+wn)}\, dl\, dm\, dn\,. \qquad (5\text{–}19)$$

Now, by sampling V, weighting by $e^{-2\pi iw}$ and by the Fourier kernel, and integrating over (u, v, w), one obtains an analog of Equation 5–2,

$$I^{3D}(l,m,n) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u,v,w)V(u,v,w)e^{-2\pi iw}e^{2\pi i(ul+vm+wn)}\, du\, dv\, dw\,, \qquad (5\text{–}20)$$

which (*cf.* Eq. 5–19) is equal to a three-dimensional convolution—the convolution of

$$I^{3}(l,m,n) \equiv \frac{\mathcal{A}(l,m)I(l,m)}{\sqrt{1-l^2-m^2}}\, \delta(n-\sqrt{1-l^2-m^2})\,, \qquad (5\text{–}21)$$

with

$$B^{3D}(l,m,n) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} S(u,v,w)e^{2\pi i(ul+vm+wn)}\, du\, dv\, dw\,. \qquad (5\text{–}22)$$

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 5–20. In applying the 'direct Fourier transform' method, one simply uses a discrete summation, in analog to Equation 5–3. In the FFT method, w-terms need

to be inserted into Expression 5–10, defining the gridding operation; a 3-D FFT yields a three-dimensional discretely sampled image[1]; 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.

One additional approach to this problem, involving a combination of mosaicing and deconvolution, is mentioned below.

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, $\tilde{I}_c^D$. 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 $\tilde{I}_c^D = \tilde{B}_c^D * I$, is approximately valid, this first step provides a starting point for the deconvolution (i.e., sidelobe removal) process described in Lecture 7. 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 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 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.

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 5–4.

References

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

[1] In the FFT method, one normally would want a shift of origin, in order to get the plane tangent to the celestial sphere at $(0,0,1)$ shifted to the origin of the third coordinate axis of the grid. This involves multiplying the data by $e^{2\pi i w}$, which cancels the multiplication by $e^{-2\pi i w}$ in Equation 5–20.

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. (1980), "Optimal gridding", VLA Scientific Memorandum No. 132, NRAO.

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

6. 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 6–1 shows a schematic diagram for a two-element, single-multiplier, correlation interferometer. All of the electronics from the output of the feed horn up 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]

$$P = k_B \Delta\nu T\,, \tag{6–1}$$

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

[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 (1966).

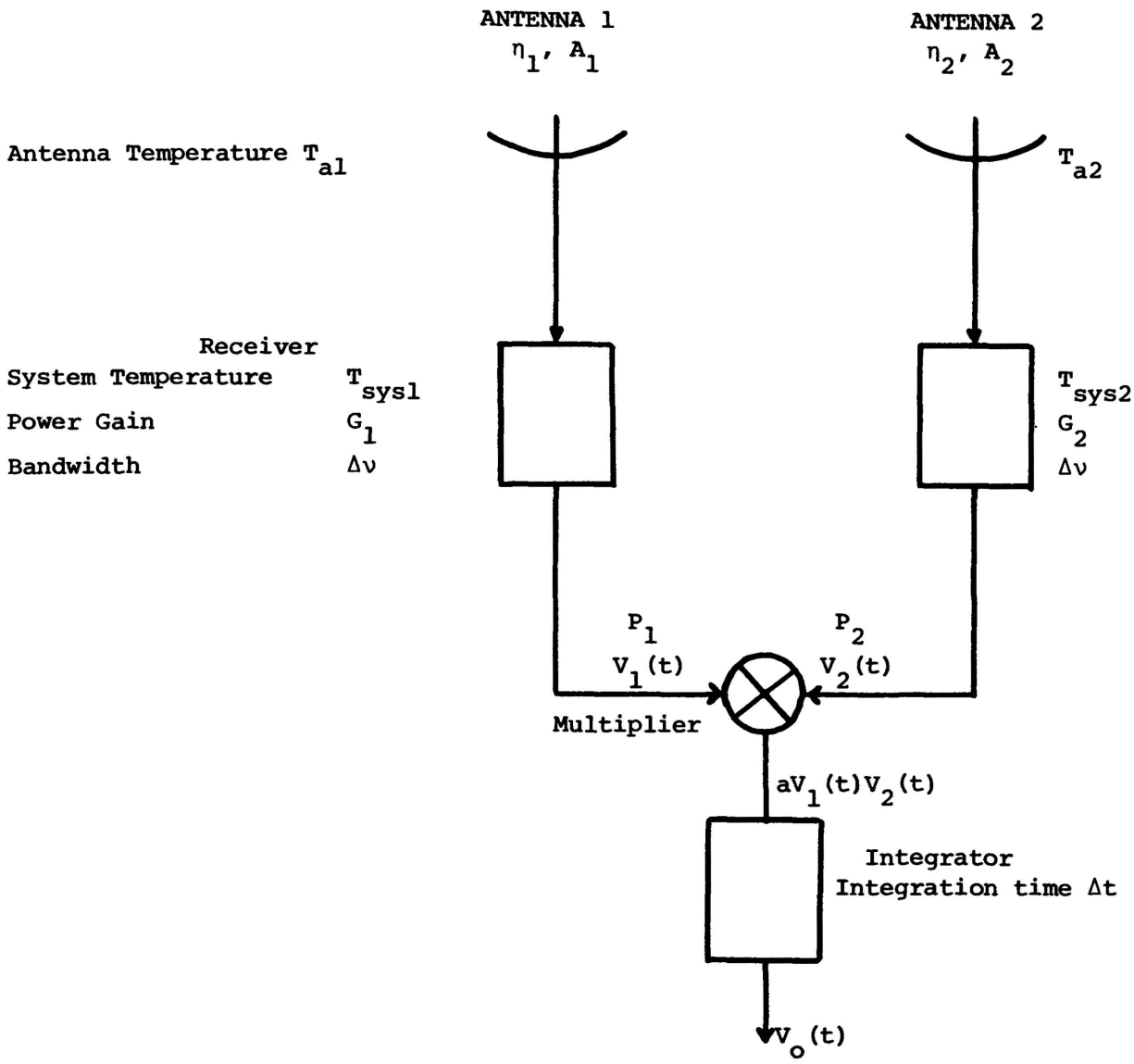


Figure 6–1. Simple block diagram of a two-antenna, single-multiplier, correlation interferometer.

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_{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

$$k_B T_{sys} \Delta\nu G = P_N \,. \tag{6–2}$$

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

$$k_B T_a \Delta\nu G = P_a \,. \tag{6-3}$$

Note that we have defined T_{sys} to include only the effects of additive noise. Sometimes T_{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_{sys}. In most practical situations this point is not important because $T_a \ll T_{sys}$. It is interesting to break T_{sys} up into its component parts:

$$T_{sys} = T_{bg} + T_{sky} + T_{spill} + T_{loss} + T_{cal} + T_{rx} \,, \tag{6-4}$$

where

T_{sys} = Total system temperature excluding noise contribution from the observed radio source,

T_{bg} = Noise contribution from microwave and galactic backgrounds,

T_{sky} = Noise contributed from atmospheric emission,

T_{spill} = Noise contribution due to ground radiation scattering into the feed from the subreflector edge, feed legs, dish edge, etc.,

T_{loss} = Noise contribution due to losses in the feed and input waveguide,

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

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

Table 6–1 gives typical values at the zenith for these noise contributions for the six VLA receivers.

Table 6–1.
Noise Contributions in the VLA Receivers

Wavelength Band	T_{bg} (K)	T_{sky} (K)	T_{spill} (K)	T_{loss} (K)	T_{cal} (K)	T_{rx} (K)	T_{sys} (K)
92 cm	3	25	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	3	3	2	2	32	45
2 cm	3	8	6	13	6	80	116
1.3 cm	3	17	6	21	7	296	350

For the VLA, the T_{spill} and T_{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. At 1.3 cm wavelength the T_{rx} contribution should decrease to approximately 100 K when the new cooled preamplifiers are installed.

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_{spill} will change as the orientation of the antenna with respect to the ground varies. More significantly, at high frequencies the contribution T_{sky} from atmospheric emission varies with time and antenna position. Assuming a plane-parallel atmosphere with temperature T_{atm}, the dependence of T_{sky} on position is given by

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

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

$$T_a = T_{a_0} e^{-\tau \csc E}.$$

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

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

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 (1969), Crane (1982), Rogers (1968, 1976), Staelin (1974), and Tiuri (1964, 1966). The following derivation follows that of Crane (1982).

Consider the case in which the interferometer shown in Figure 6–1 is observing an unpolarized point source of flux density S (Janskys, 1 Jy $= 10^{-26}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$). The antenna temperature of antenna 1 due to the source is given by

$$T_{a_1} = \frac{\eta_1 A_1 S}{2k_B} = K_1 S, \qquad (6\text{–}5)$$

where η_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 6–5 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 6–5, is given by

$$K = \frac{\eta A}{2k_B}.$$

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

Table 6–2.
Flux-Collecting Ability of VLA Antennas

Wavelength Band	η (%)	$K = T_a/S$ ($\mathrm{K\,Jy^{-1}}$)
92 cm	40	0.071
20 cm	51	0.091
6 cm	65	0.116
3.6 cm	65	0.116
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

$$\begin{aligned} V_1(t) &= S_1(t) + n_1(t)\,, \\ V_2(t) &= S_2(t) + n_2(t)\,, \end{aligned} \tag{6-6}$$

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

$$\phi_{s_1}(\tau) = \langle S_1 S_1' \rangle\,, \tag{6-7}$$

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

$$S_2(t) = \sqrt{\frac{K_1}{K_2}}\, S_1(t)\,. \tag{6-8}$$

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

$$\phi_{s_1}(0) = G k_B T_{a_1} \Delta\nu\,. \tag{6-9}$$

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

$$\phi_{n_1}(\tau) = \langle n_1 n_1' \rangle\,, \tag{6-10}$$

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

$$\phi_{n_1}(0) = G k_B T_{\mathrm{sys}_1} \Delta\nu\,, \tag{6-11}$$

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

$$\langle n_2 n_2' \rangle = \frac{T_{\mathrm{sys}_2}}{T_{\mathrm{sys}_1}} \phi_{n_1}(\tau)\,. \tag{6-12}$$

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

$$\phi_m(\tau) = a^2 \langle V_1(t)V_2(t)V_1(t+\tau)V_2(t+\tau)\rangle . \tag{6-13}$$

Using Equation 6–6

$$\phi_m(\tau) = a^2 \langle (S_1+n_1)(S_2+n_2)(S_1'+n_1')(S_2'+n_2')\rangle , \tag{6-14}$$

where a is a constant.

Equation 6–14 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)

$$\begin{aligned}\phi_m(\tau) = a^2[&\langle (S_1+n_1)(S_2+n_2)\rangle \cdot \langle (S_1'+n_1')(S_2'+n_2')\rangle \\ &+\langle (S_1+n_1)(S_1'+n_1')\rangle \cdot \langle (S_2+n_2)(S_2'+n_2')\rangle \\ &+\langle (S_1+n_1)(S_2'+n_2')\rangle \cdot \langle (S_2+n_2)(S_1'+n_1')\rangle] .\end{aligned} \tag{6-15}$$

The required multiplications in Equation 6–15 can now be carried out simply. Setting the averages of the products of all uncorrelated voltages to zero, and using Equations 6–7, 6–8, 6–10 and 6–12, Equation 6–15 becomes

$$\phi_m(\tau) = a^2 \left(\frac{K_2}{K_1}\phi_{s_1}^2(0) + 2\frac{K_2}{K_1}\phi_{s_1}^2(\tau) + \left(\frac{K_2}{K_1} + \frac{T_{sys_2}}{T_{sys_1}}\right)\phi_{s_1}(\tau)\phi_{n_1}(\tau) + \frac{T_{sys_2}}{T_{sys_1}}\phi_{n_1}^2(\tau)\right) . \tag{6-16}$$

Applying the Wiener–Khinchine theorem to Equation 6–16 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

$$\begin{aligned}\Phi_m(\nu) = a^2 \Bigg(&\frac{K_2}{K_1}\phi_{s_1}^2(0)\delta(\nu) + 2\frac{K_2}{K_1}\int_{-\infty}^{\infty}\Phi_{s_1}(\alpha)\Phi_{s_1}(\nu-\alpha)\,d\alpha \\ &+\left(\frac{K_2}{K_1} + \frac{T_{sys_2}}{T_{sys_1}}\right)\int_{-\infty}^{\infty}\Phi_{s_1}(\alpha)\Phi_{n_1}(\nu-\alpha)\,d\alpha + \frac{T_{sys_2}}{T_{sys_1}}\int_{-\infty}^{\infty}\Phi_{n_1}(\alpha)\Phi_{n_1}(\nu-\alpha)\,d\alpha\Bigg) ,\end{aligned} \tag{6-17}$$

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

The power spectral components of the multiplier output are shown in Figure 6–2, 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 6–2 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. 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

$$P_{\text{noise}} = a^2\frac{K_2}{K_1}\phi_s^2(0) .$$

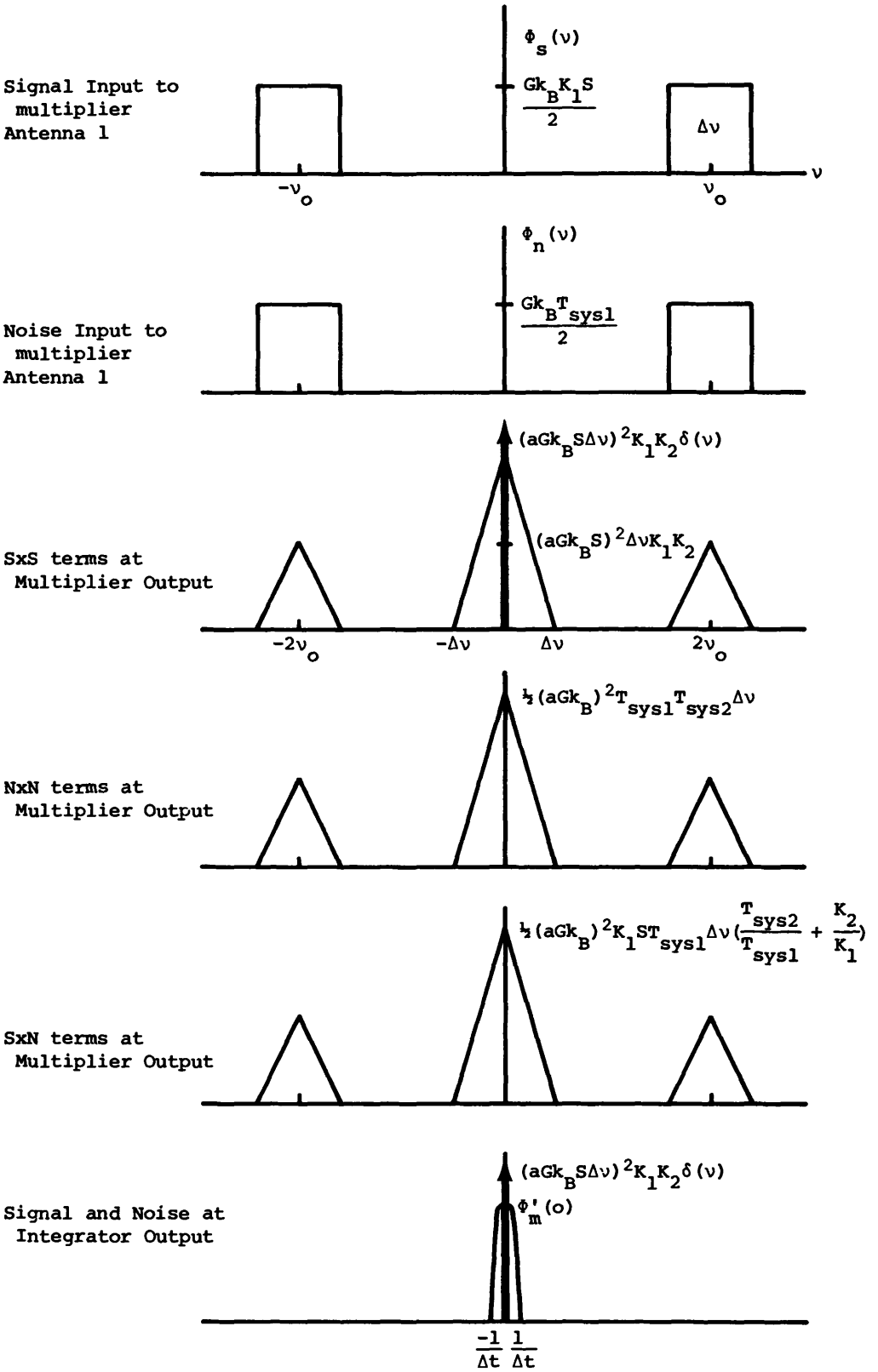


Figure 6–2. Power spectra for a two-antenna, single-multiplier, correlation interferometer.

Thus, from Equations 6–9 and 6–5 the DC voltage is

$$\langle V_0 \rangle = aGk_B S \Delta\nu \sqrt{K_1 K_2}\,. \tag{6–18}$$

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

$$\langle V_0^2(t) \rangle = \int_{-\infty}^{\infty} \Phi'_m(\nu) |H(\nu)|^2 \, d\nu\,, \tag{6–19}$$

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 6–2. For an ideal integrator the impulse response is

$$h(t) = \begin{cases} \frac{1}{\Delta t}\,, & 0 < t < \Delta t\,, \\ 0\,, & \text{elsewhere}\,, \end{cases} \tag{6–20}$$

and the output noise power P_{noise} is, from Equation 6–19,

$$P_{\text{noise}} = \Phi'_m(0) \int_{-\infty}^{\infty} |H(\nu)|^2 \, d\nu\,. \tag{6–21}$$

Using Parseval's theorem

$$\begin{aligned} P_{\text{noise}} &= \Phi'_m(0) \int_{-\infty}^{\infty} h^2(t)\, dt \\ &= \frac{\Phi'_m(0)}{\Delta t} \qquad (6\text{–}22) \\ &= (aGk_B)^2 \frac{\Delta\nu}{\Delta t} \left(K_1 K_2 S^2 + \frac{K_1 S T_{\text{sys}_1}}{2} \left(\frac{T_{\text{sys}_2}}{T_{\text{sys}_1}} + \frac{K_2}{K_1} \right) + \frac{T_{\text{sys}_1} T_{\text{sys}_2}}{2} \right). \qquad (6\text{–}23) \end{aligned}$$

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

$$\Delta S = \frac{\text{r.m.s. noise}}{\text{correlator scaling factor (V Jy}^{-1})} = \sqrt{P_{\text{noise}}} \Big/ \frac{\partial V_0}{\partial S}\,. \tag{6–24}$$

Thus from Equations 6–24, 6–23, and 6–18,

$$\Delta S = \frac{1}{\sqrt{\Delta t \Delta\nu}} \sqrt{S^2 + \frac{S}{2} \left(\frac{T_{\text{sys}_1}}{K_1} + \frac{T_{\text{sys}_2}}{K_2} \right) + \frac{T_{\text{sys}_1} T_{\text{sys}_2}}{2 K_1 K_2}}\,. \tag{6–25}$$

In the case of identical antennas and receivers, Equation 6–25 simplifies to

$$\Delta S = \frac{1}{\sqrt{\Delta t \Delta\nu}} \sqrt{S^2 + \frac{S T_{\text{sys}}}{K} + \frac{T_{\text{sys}}^2}{2K^2}}\,. \tag{6–26}$$

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

$$\Delta S = \frac{T_{\text{sys}}}{K} \frac{1}{\sqrt{2 \Delta t \Delta\nu}} = \frac{\sqrt{2} k_B T_{\text{sys}}}{\eta A \sqrt{\Delta t \Delta\nu}}\,. \tag{6–27}$$

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 dish 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 dish separately.

In the case of a strong source, $S \gg T_{sys}/K$, Equation 6–26 becomes

$$\Delta S = \frac{S}{\sqrt{\Delta t \Delta \nu}} = \frac{2k_B T_a}{\sqrt{\Delta t \Delta \nu}\,\eta A}. \tag{6-28}$$

Equation 6–28 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 6–28 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 6–26 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

$$S_c(\nu) = A(\nu)\cos(\phi(\nu) + \phi_1(\nu) - \phi_2(\nu) + \omega\Delta\tau),$$

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

$$V_0 = ak_B\sqrt{K_1K_2}\int_0^\infty G_1(\nu)G_2(\nu)S_c(\nu)\,d\nu,$$

and the output noise power is given by

$$P_{\text{noise}} = \frac{a^2k_B^2K_1K_2}{2\Delta\tau} \times \int_0^\infty G_1^2(\nu)G_2^2(\nu)\left[S_c^2(\nu) + S^2(\nu) + S(\nu)\left(\frac{T_{sys_1}(\nu)}{K_1} + \frac{T_{sys_2}(\nu)}{K_2}\right) + \frac{T_{sys_1}(\nu)T_{sys_2}(\nu)}{K_1K_2}\right]d\nu.$$

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

$$\Delta S = \frac{\sqrt{\int_0^\infty G_1^2(\nu)G_2^2(\nu)\left[S_c^2(\nu) + S^2(\nu) + S(\nu)\left(\frac{T_{sys_1}(\nu)}{K_1} + \frac{T_{sys_2}(\nu)}{K_2}\right) + \frac{T_{sys_1}(\nu)T_{sys_2}(\nu)}{K_1K_2}\right]d\nu}}{\sqrt{2\Delta t}\int_0^\infty G_1(\nu)G_2(\nu)\,d\nu}.$$

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

$$\Delta S = \frac{S}{\sqrt{2\Delta\nu\Delta t}}.$$

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 6–27 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,

$$\eta_c = \frac{\text{sensitivity of the correlator}}{\text{sensitivity of a perfect analog correlator having the same } \Delta t}. \qquad (6\text{–}30)$$

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 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-band 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-band spectral-line mode, 0.77.

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

$$\Delta S = \frac{\sqrt{2}k_B T_{sys}}{\sqrt{\Delta t \Delta\nu} A \eta_a \eta_c}, \qquad (6\text{–}31)$$

where η_a is the antenna aperture efficiency.

Table 6–3 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 6–3.
VLA Single-Multiplier Interferometer Sensitivity

Wavelength Band	ΔS (mJy)
92 cm	73
20 cm	28
6 cm	18
3.6 cm	16
2 cm	52
1.3 cm	180*

*Approximately 80 mJy after receiver upgrade

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

The derivation of Equation 6–31 assumed that the point source was at the phase center 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 a point source of arbitrary position 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

$$S_m = \sqrt{S_R^2 + S_I^2}\,, \tag{6–32}$$

$$\phi_m = \tan^{-1}\frac{S_I}{S_R}\,. \tag{6–33}$$

Both S_R and S_I are contaminated by noise with r.m.s. value ΔS given by Equation 6–31, 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. This problem has been examined by several authors (Rogers 1968, 1976; Vinokur 1965; Hjellming and Basart 1982; Berge 1965; Moran 1973, 1976).

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

$$P(S_m) = \frac{S_m}{\Delta S^2}\, I_0\left(\frac{S_m S}{\Delta S^2}\right) \exp\frac{-(S_m^2 + S^2)}{2\Delta S^2}\,, \tag{6–34}$$

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 6–3 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).

$$\langle S_m \rangle \simeq \sqrt{\frac{\pi}{2}}\,\Delta S\left(1 + \frac{S^2}{4\Delta S^2}\right)\,, \tag{6–35}$$

and

$$\sigma_{S_m} \simeq \sqrt{2 - \frac{\pi}{2}}\,\Delta S\left(1 + \frac{S^2}{4\Delta S^2}\right)\,. \tag{6–36}$$

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

$$I_0\left(\frac{S_m S}{\Delta S^2}\right) \simeq \frac{\Delta S}{\sqrt{2\pi S_m S}} \exp\frac{S_m S}{\Delta S^2}\,,$$

and

$$P(S_m) \simeq \frac{1}{\Delta S}\sqrt{\frac{S_m}{2\pi S}} \exp\frac{-(S_m - S)^2}{2\Delta S^2}\,.$$

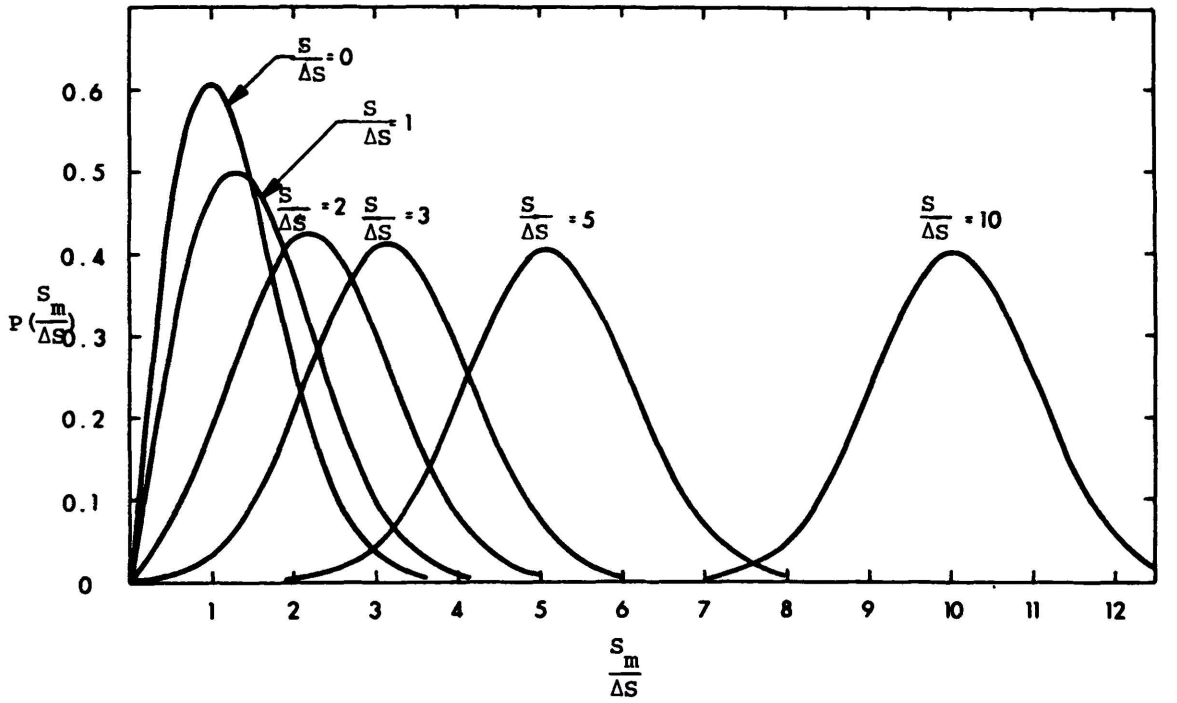


Figure 6–3. The probability distribution of the measured amplitude is plotted as a function of the apparent signal-to-noise ratio for a number of values of the true signal-to-noise ratio.

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

$$P(S_m) \simeq \frac{1}{\sqrt{2\pi}\Delta S} \exp \frac{-(S_m - S)^2}{2\Delta S^2}, \tag{6–37}$$

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),

$$P(\phi - \phi_m) = \frac{1}{2\pi} \exp\left(\frac{-S^2}{2\Delta S^2}\right)\left(1 + G\sqrt{\pi}e^{G^2}(1 + \operatorname{erf} G)\right), \tag{6–38}$$

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 6–4 for several values of $S/\Delta S$. For small $S/\Delta S$ the r.m.s. phase error is given by (Moran 1976)

$$\sigma_{\phi_m} \simeq \frac{\pi}{\sqrt{3}}\left(1 - \sqrt{\frac{9}{2\pi^3}}\frac{S}{\Delta S}\right), \tag{6–39}$$

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

$$\sigma_{\phi_m} = \frac{\Delta S}{S}. \tag{6–40}$$

Figure 6–4 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 $S/\Delta S$ of 0 and 1 is much more obvious than the difference between the associated amplitude distributions.

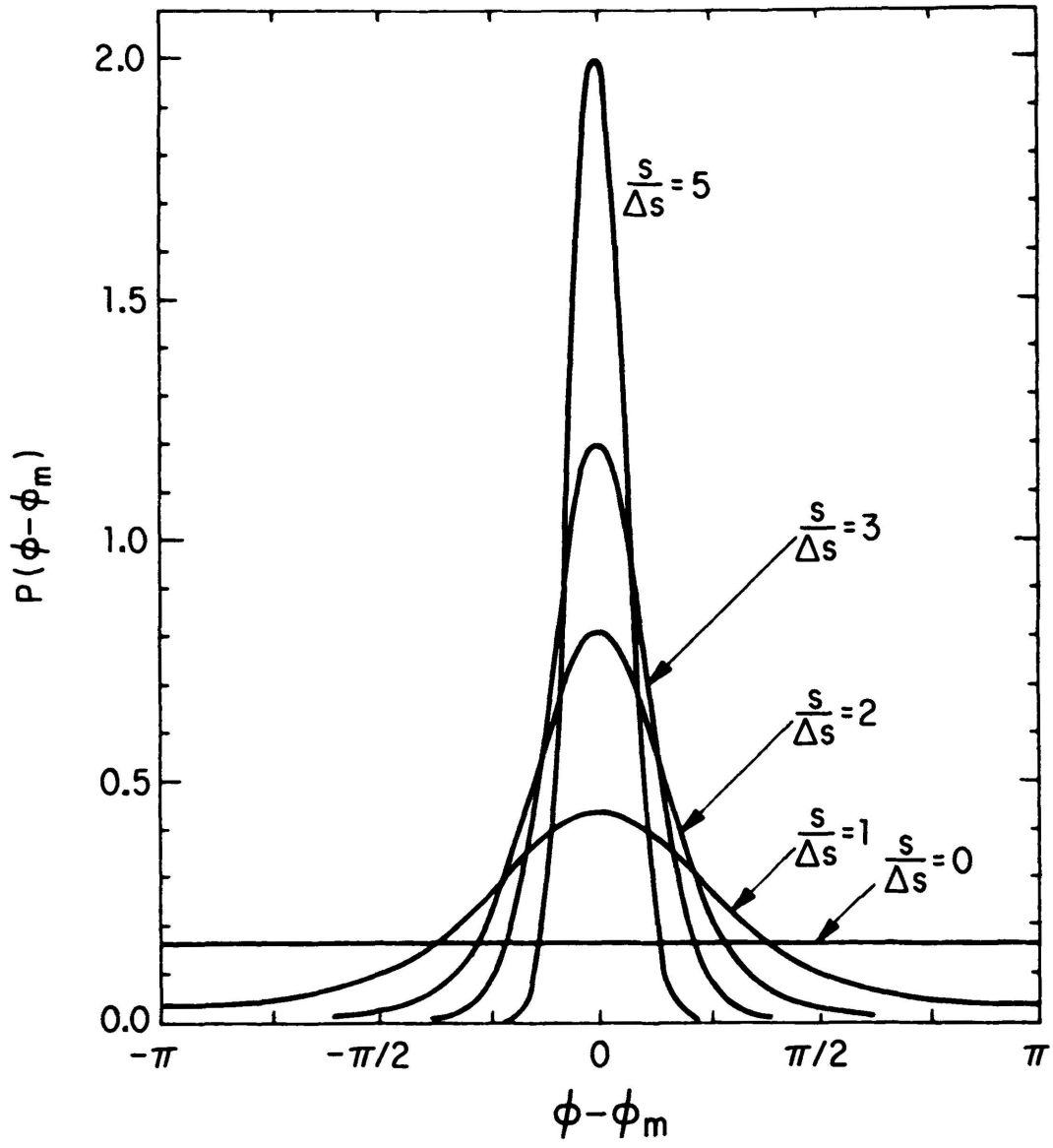


Figure 6–4. The probability distribution of the measured phase is plotted as a function of $\phi - \phi_m$ for a number of values of the true signal-to-noise ratio.

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)

$$I(l,m) = \iint V(u,v) e^{2\pi i(ul+vm)}\, du\, dv, \tag{6–41}$$

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

$$V = S_R + iS_I, \tag{6–42}$$

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, Expression 6–41 is approximated by, including the

tapering and weighting functions discussed in Lecture 5,

$$I_m(l,m) = K\sum_{\ell=0}^{2L} T_\ell W_\ell V_\ell e^{2\pi i(u_\ell l + v_\ell m)}, \tag{6–43}$$

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 6–43 can be rewritten using Equation 6–42

$$I_m(l,m) = 2K\sum_{\ell=1}^{L} T_\ell W_\ell \left(S_{R_\ell}\cos 2\pi(u_\ell l + v_\ell m) - S_{I_\ell}\sin 2\pi(u_\ell l + v_\ell m)\right), \tag{6–44}$$

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 6–44 is a 'direct Fourier transform' and the noise will be the same at all points on the image. At the image center

$$I_m(0,0) = 2K\sum_{\ell=1}^{L} T_\ell W_\ell S_{R_\ell}. \tag{6–45}$$

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

$$S_{R_\ell} = S + n_{R_\ell}, \tag{6–46}$$

where n_{R_ℓ} is the noise in the real part of the correlator output and has the properties $n_{R_\ell} = 0$ and $n_{R_\ell}^2 = \Delta S^2$, where ΔS is given by Equation 6–31. The expected value at the image center is

$$I_m(0,0) = 2KS\sum_{\ell=1}^{L} T_\ell W_\ell. \tag{6–47}$$

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

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

so that

$$I_m(0,0) = S, \tag{6–48}$$

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

$$\Delta I_m = 2K\Delta S\sqrt{\sum_{\ell=1}^{L} T_\ell^2 W_\ell^2}. \tag{6–49}$$

For a naturally weighted, untapered image, this simplifies to

$$\Delta I_m = \frac{\Delta S}{\sqrt{L}}. \tag{6–50}$$

For an array with C complex correlators, correlator integration time Δt and total observation time T, the number of measurements is

$$L = \frac{CT}{\Delta t}. \tag{6-51}$$

Combining Equations 6–31, 6–50 and 6–51 gives the desired expression for the noise in the image

$$\Delta I_m = \frac{\sqrt{2}k_B T_{sys}}{A\eta_a\eta_c\sqrt{CT\Delta\nu}}. \tag{6-52}$$

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 6–51 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, $C = N(N-1)$. As N becomes large, C approaches N^2, and ΔI_m becomes

$$\Delta I_m = \frac{\sqrt{2}k_B T_{sys}}{\eta_a\eta_c NA\sqrt{T\Delta\nu}}, \tag{6-53}$$

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 each antenna 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 6–4 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 6–4.
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	9
2 cm	30
1.3 cm	100*

*Approximately 44 μJy/beam after receiver upgrade

In natural weighting, every correlator measurement is given the same weight. This gives the highest sensitivity for detecting point sources. As described in Lecture 5 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

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

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 5, 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), 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 6–5 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 labelled "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.)

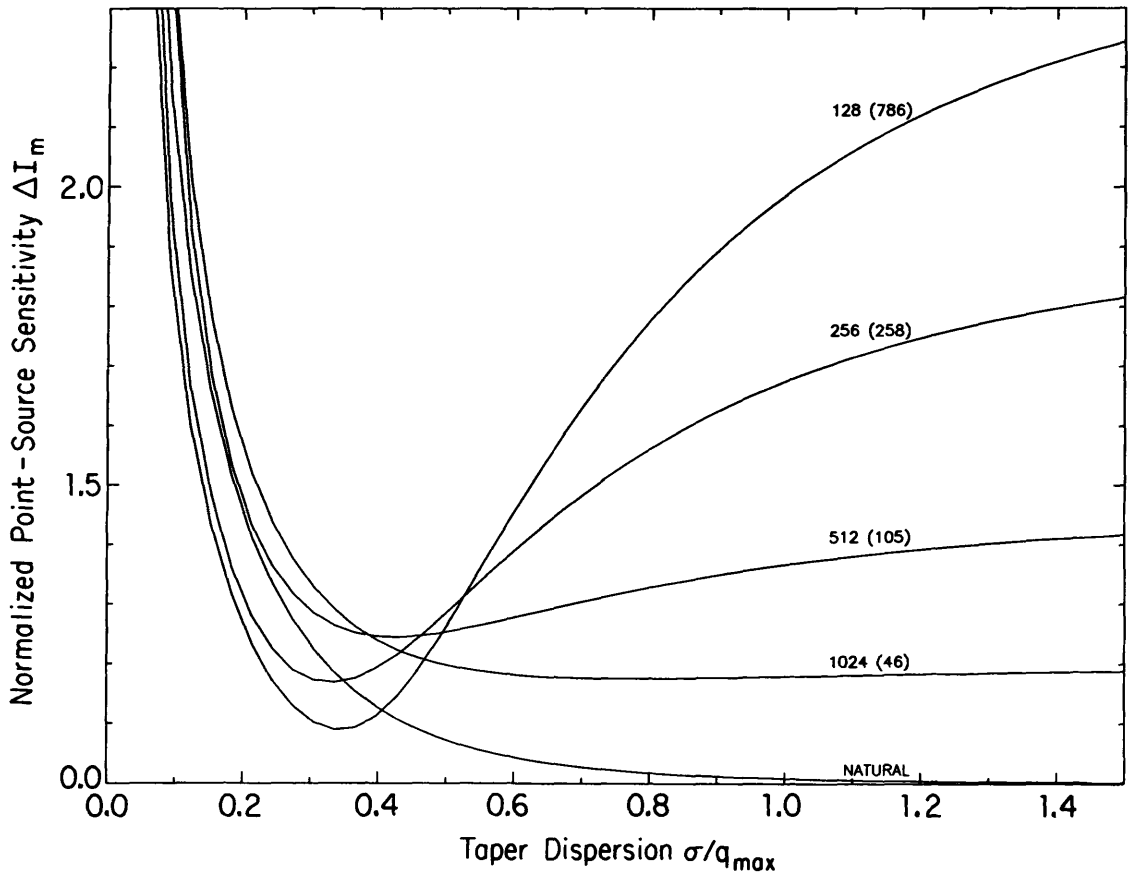


Figure 6–5. The effects of applying a Gaussian tapering function on the point-source sensitivity of the VLA, shown for natural weighting and for uniform weighting with four different numbers of cells spanning the u-v plane. The calculations are for a source at the North Celestial Pole, observed for twelve hours with an integration time of 100 seconds, and have been normalized to the result for untapered natural weighting.

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 an 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. Fomalont and Wright (1973) give further discussion of this point. In general, the resolution of an array increases linearly with the size of the array, 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 array

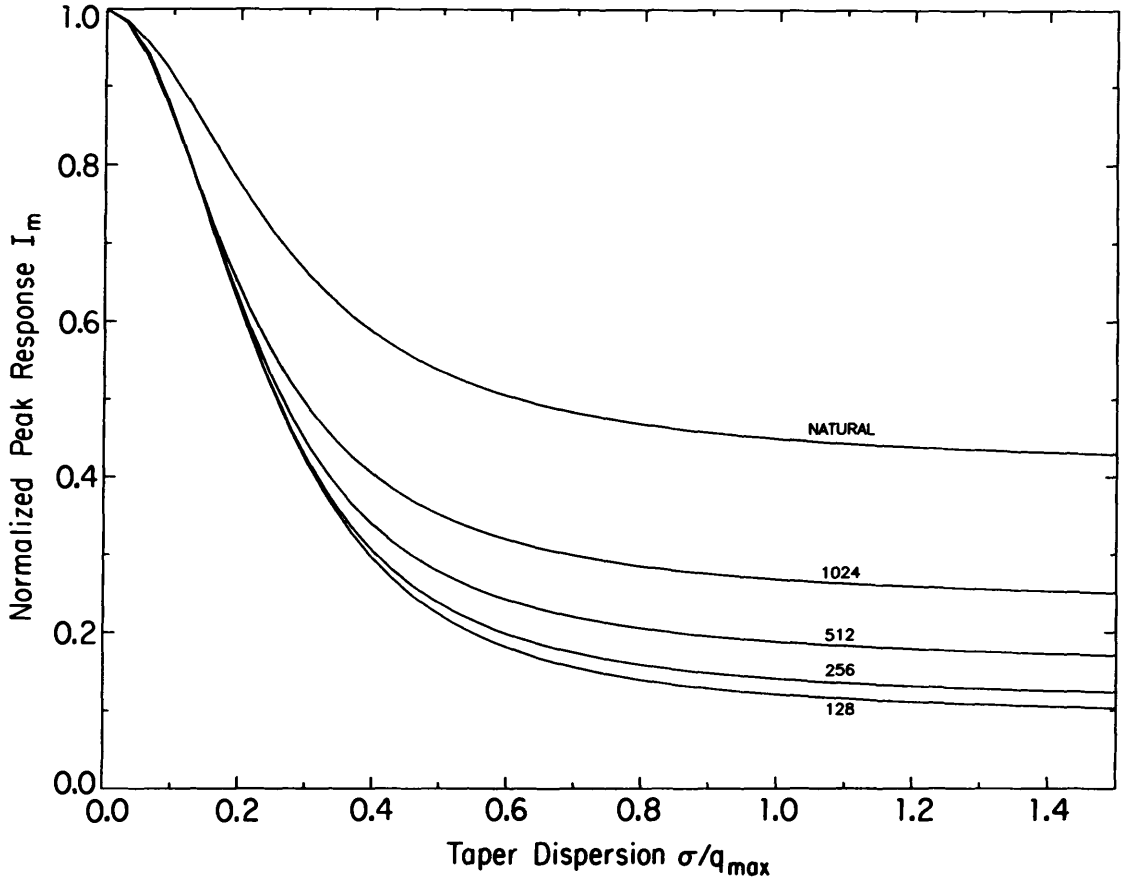


Figure 6–6. The effects of applying a Gaussian tapering function on the peak response of the VLA to a circular Gaussian source with θ of $2.08\lambda/q_{max}$, shown for natural weighting and for uniform weighting with four different numbers of cells spanning the u-v plane. The calculations are for a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

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 16).

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 5. 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 6–5, ΔI_m remains constant or even decreases for a wide range of σ. So for $\sigma \approx 0.3$–$0.4q_{max}$, the improvement in sensitivity to an extended source may be a factor of 4–10 over 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 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

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

As with the point-source sensitivity, the peak response of the VLA to a circular Gaussian source depends upon many variables; Figure 6–6 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 6–5. The

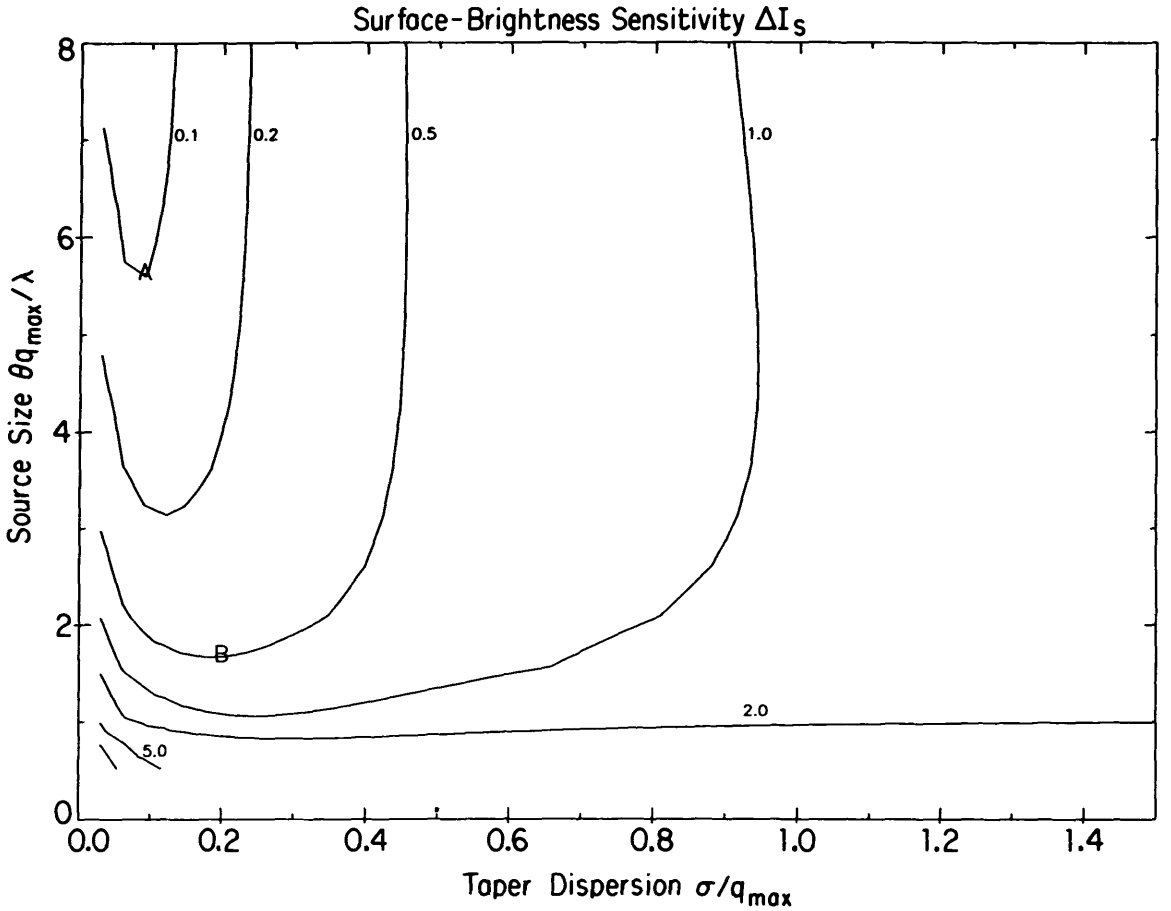


Figure 6–7. The effects of tapering and source size on the surface-brightness sensitivity of the VLA, shown for uniform weighting over 512 cells, and assuming a Gaussian tapering function of dispersion σ. The calculations are for a source at the North Celestial Pole, observed for twelve hours with an integration time of 100 seconds. The contour levels are 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, and 10.0 in arbitrary units which scale as $3.285^{2(1-n)}$ where n is 1, 2, 3, or 4 for the **A**, **B**, **C**, or **D** configurations, respectively. (Note that the calculations do not extend to zero size or taper.) Points **A** and **B** indicate the tapers which provide the optimum sensitivities for observing the same source ($\theta = 5.6\lambda/q_{\mathbf{max},A} = 1.7\lambda/q_{\mathbf{max},B}$) in the **A** and **B** configurations, respectively; for further discussion see the text.

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

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

As illustrated in Figures 6–5 and 6–6, this function will depend upon the details of the observing strategy used, the parameters chosen for the image, and many other variables. Figure 6–7 shows ΔI_s for the case of 512 u-v cells already shown in Figures 6–5 and 6–6. Figure 6–7 can be used, for example, to compare the surface-brightness sensitivities of the **A** and **B** configurations. Point **A** shows that the optimum sensitivity for a source with a θ of $5.6\lambda/q_{\mathbf{max},A}$ ($1.7\lambda/q_{\mathbf{max},B}$), observed in the **A** configuration, is 0.1 for $\sigma \approx 0.1 q_{\mathbf{max},A}$. Point **B** shows that the optimum sensitivity for the same source observed in the **B** configuration is $0.5(3.285)^{-2}$, or 0.046, for $\sigma \approx 0.2 q_{\mathbf{max},B}$. As expected, the **B** configuration is more sensitive to extended structure than the **A** configuration, although, in this example, not by the factor often suggested by the simple arguments used above. One reason is that the tapering function can be adjusted to obtain the optimum sensitivity in each configuration.

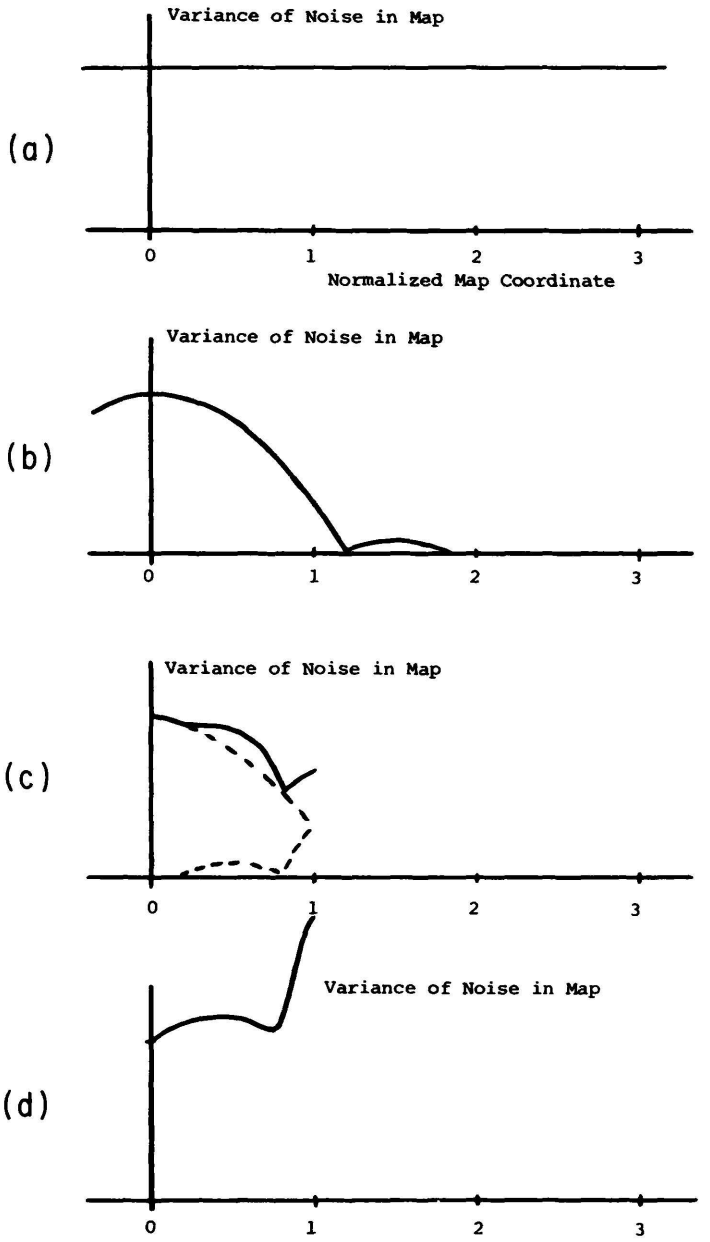


Figure 6–8. The effects of convolution and gridding in the u-v plane on the noise in an image: (a) 'Direct Fourier transform', no convolution or gridding. (b) 'Direct Fourier transform' after convolution. (c) Fast Fourier transform after convolution and gridding. (d) Fast Fourier transform after convolution and gridding, followed by division by the Fourier transform of the convolving function.

The lesson to be learned from this discussion is that the optimum taper for a particular observation can only be determined by trial and error.

7. The Effects of Convolution and Gridding on Sensitivity

The sensitivity analyses in Section 5 are appropriate when images are made using a '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 6–8a (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 6–8b, 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 6–8c. 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 6–8d. 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 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 effect 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 array, we may express the measured brightness image $I_m(l,m)$ as

$$I_m(l,m) = I(l,m)P(l,m) + N(l,m)\,, \tag{6–54}$$

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

$$I_m(l,m)/P(l,m) = I(l,m) + N(l,m)/P(l,m)\,. \tag{6–55}$$

In this case Equation 6–55 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 6–3 and 6–4.

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, N. N. and Högbom, J. A. (1969), *Radiotelescopes*, Cambridge University Press, London.

Clark, B. G. (1976), "Gridding and Signal-to-Noise Ratios", NRAO, VLA Scientific Memo. No. 124.

Cooper, B. (1970), "Correlators With Two-Bit Quantization", *Aust. J. Phys.*, **23**, 521–527.

Crane, P. C. (1982), "Signal Analysis of a Correlation Interferometer", NRAO, VLA Scientific Memo. No. 140.

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", NRAO, VLA Technical Report No. 39.

Fomalont, E. B. and Wright, M. C. H. (1973), "Interferometry and Aperture Synthesis", in *Galactic and Extragalactic Radio Astronomy*, Springer–Verlag, New York.

Greisen, E. W. (1976), "On the Effects of Convolution in the u-v Plane", NRAO, VLA Scientific Memo. No. 123.

Greisen, E. W. (1979), "The Effects of Various Convolving Functions on Aliasing and Relative Signal-to-Noise Ratios", NRAO, VLA Scientific Memo. No. 131.

Hjellming, R. M. and Basart, J. P. (1982), "The Theory of the Instrument", in *An Introduction to the NRAO Very Large Array*, NRAO.

Kraus, J. D. (1966), *Radio Astronomy*, McGraw–Hill, New York.

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, Academic Press, New York.

Rogers, A. E. E. (1968), "Spectral Line Interferometry and Interferometer Noise Analysis", MIT Lincoln Labs., Technical Report No. 441.

Rogers, A. E. E. (1976), "Theory of Two-Element Interferometers", in *Methods of Experimental Physics*, Vol. 12C, Academic Press, New York.

Spangler, S. R. (1982), "Correction of VLA K-Band Amplitudes for Atmospheric Attenuation", NRAO, VLA Scientific Memorandum No. 143.

Staelin, D. H. (1974), "The Detection and Measurement of Radio Astronomical Signals", MIT Dept. Elec. Engrng., Lecture Notes.

Tiuri, M. E. (1964), "Radio Astronomy Receivers", *IEEE Trans.*, **AP-12**, 930–938.

Tiuri, M. E. (1966), "Radio-Telescope Receivers", in Kraus, J. D., *Radio Astronomy*, McGraw–Hill, New York.

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), "Optimisation in the Search for a Sinusoid of Known Period in the Presence of Noise. Application to Radio Astronomy", *Ann. D'Ap.*, **28**, 412–445.

7. Deconvolution

TIM CORNWELL

1. DECONVOLUTION

This Lecture describes how the visibility samples collected by an interferometer 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 5 will have defects due to the limited sampling of the u-v plane. Non-linear deconvolution is required to correct these defects.

As described in Lectures 1 and 2, an interferometer 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, Section 4.2 and Lecture 2, Section 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:

$$V(u,v) = \iint_S I(l,m) e^{-2\pi i(ul+vm)}\, dl\, dm\,, \tag{7-1}$$

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 $\hat{I}$ of the source intensity that is capable of representing all the visibility data consists of a two-dimensional grid of δ-functions with strengths $\hat{I}(p\Delta l, q\Delta m)$ where Δl and Δm are the separations of the grid elements in the two orthogonal sky coordinates. The visibility $\hat{V}$ predicted by this model is given by:

$$\hat{V}(u,v) = \sum_{p=1}^{N_l} \sum_{q=1}^{N_m} \hat{I}(p\Delta l, q\Delta m) e^{-2\pi i(pu\Delta l + qu\Delta m)}\,. \tag{7-2}$$

For simplicity I will henceforth denote the discrete form $\hat{I}(p\Delta l, q\Delta m)$ by the notation $\hat{I}_{p,q}$. Assuming reasonably uniform sampling of a region of the u-v plane, one can expect to estimate source features with widths ranging from $O(1/\max(u,v))$ up to $O(1/\min(u,v))$. 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:

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

This model has $N_l N_m$ free parameters, namely the cell flux densities $\hat{I}_{p,q}$. The measurements constrain the model such that at the sampled u-v points:

$$V(u_r, v_r) = \hat{V}(u_r, v_r) + \epsilon(u_r, v_r)\,, \tag{7-3}$$

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 7–3 as a multiplicative relation:

$$V(u,v) = W(u,v)(\hat{V}(u,v) + \epsilon(u,v)) , \tag{7–4}$$

where $W(u,v)$ is a weighted sampling function (see Lecture 5, Equation 5–8) which is non-zero only for sampled points of the u-v plane:

$$W(u,v) = \sum_r W_r \delta(u - u_r, v - v_r) . \tag{7–5}$$

By the convolution theorem, this translates into a convolution relation in the image plane:

$$I^D_{p,q} = \sum_{p',q'} B_{p-p',q-q'} \hat{I}_{p',q'} + E_{p,q} , \tag{7–6}$$

where:

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

$$B_{p,q} = \sum_r W(u_r, v_r) \mathrm{Re}(e^{2\pi i (p u_r \Delta l + q v_r \Delta m)}) . \tag{7–8}$$

$E_{p,q}$ in Equation 7–6 is the noise image obtained by replacing V in Equation 7–7 by $\epsilon(u_r, v_r)$. Note that the $B_{p,q}$ given by Equation 7–8 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, I will not signify this gridding and grid correction explicitly). The Hermitian nature of the visibility has been used in this rearrangement.

Equation 7–4 represents the constraint that the model $\hat{I}_{p,q}$, 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 5).

I 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 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 7–6 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 non-linear 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 only extends 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 7–7 so that the bias in weight due to the vagaries of sampling are corrected. For each grid point the visibility samples are summed with appropriate weights, and the total weight normalized to unity. In such circumstances, known as uniform weighting, the principal solution is thus equal to the dirty image and is given by the convolution of the true brightness distribution with the dirty beam. For most synthesis arrays currently in use, the dirty beam has sidelobes in the range 1 to 10%. Sidelobes represent an unavoidable confusion over the true distribution of any emission in the dirty image, which can be resolved only either by making further observations or by introducing *a priori* information such as the limits in extent of the source. For example, consider uniformly weighted observations of a point source: the dirty image is just the dirty beam centered on the point source position. Without *a priori* information we cannot tell whether the source is a point or is shaped like the dirty beam. Of course we know that Stokes parameter I must be positive and that usually radio sources do not resemble dirty beams (in particular they do not have sidelobe patterns extending to infinity) and so we could use this information as an extra clue. One further unsatisfactory aspect of the principal solution, besides its implausibility, is that it changes (sometimes drastically) as more visibility data are added. A better estimator would possess greater stability.

A priori information is thus the key; in the rest of this Lecture I 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.

I now discuss 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.

The 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 $\hat{I}_{p,q}$ 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 (Section 3.3 below), etc.

2.2. The Clark algorithm.

Clark (1980) has developed an FFT-based 'CLEAN' algorithm. A large part of the work in 'CLEAN' is involved in shifting and scaling the dirty beam; since this is essentially a convolution it may, in some circumstances, be more efficiently performed via two-dimensional FFTs. Clark's algorithm does this, finding approximate positions and strengths of the components via 'CLEAN' using only a small patch of the dirty beam.

In detail, the Clark algorithm has two cycles, the major and minor cycles. The *minor cycle* proceeds as follows:

(1) A beam patch (a segment of the discrete representation of the beam) is selected to include the highest exterior sidelobe.
(2) Points are selected from the dirty image if they have an intensity, as a fraction of the image peak, greater than the highest exterior sidelobe of the beam.
(3) A 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 1984, 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 8).

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 Section 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. Equations 7–7 and 7–8), 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 re-defined. 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 7–7 and 7–8 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 I 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 5) then the number of independent visibility samples cannot be greater than $O(N_l N_m)$, 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 Section 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 I 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 which is sometimes successful in countering the instability. A small 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 I 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). I 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 1984). 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:

$$\mathcal{H} = -\sum_k I_k \ln\left(\frac{I_k}{M_k e}\right), \tag{7-10}$$

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:

$$\chi^2 = \sum_r \frac{|V(u_r, v_r) - \hat{V}(u_r, v_r)|^2}{\sigma^2_{V(u_r,v_r)}}. \tag{7-11}$$

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 7–10 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 interferometer data. A spike in the default can sometimes be used to indicate the presence of an unresolved source, which could otherwise cause problems (see Section 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 folk lore 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 non-linear algorithms which 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 (the convolved image makes a very good default image for another deconvolution!).

(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 non-linear 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.

Quite often, the super-resolution exhibited by MEM images is reliable and can be trusted up to an order of magnitude in solid angle.

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.

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. I 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, 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. Future Developments

Deconvolution in radio astronomy is dominated by two *non-linear* algorithms, 'CLEAN' and MEM. Other non-linear 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.

The concept of a default image can be extended to 'CLEAN' and other algorithms, and 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 only be applicable to certain telescopes, such as east–west arrays.

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. I 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 inbalance.

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.

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 stabilising the clean algorithm", *Astron. Astrophys.*, **121**, 281–285.

Cornwell, T. J. (1984), "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.*, **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**, to appear.

Schwab, F. R. (1984), "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.

8. Special Problems in Imaging

WILLIAM D. COTTON

In practical applications, one or more of the simplifying assumptions which 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 relatively practical point of view; the general nature of the problems is described, as are the conditions under which they become important. Finally, there is a discussion of techniques used to reduce the distortions and/or the error 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 which become important only when a wide field of view is imaged.

1.1. Bandwidth smearing (chromatic aberration).

The effect of finite bandwidth on a correlator was discussed in Lecture 2; this effect can be shown by expressing u and v as functions of frequency and explicitly averaging over frequency. The monochromatic Fourier transform relation between visibility and intensity (Lecture 1, Equation 1–9) can be re-expressed in terms of the *bandwidth-smeared* intensity $\tilde{I}(l,m)$, the *frequency-dependent* u's and v's and the instrumental bandpass $g(\nu)$ as:

$$\tilde{I}(l,m) = \iint \tilde{V}(u_0,v_0) e^{2\pi i(u_0 l + v_0 m)} du_0\, dv_0\,, \tag{8–1}$$

where

$$\tilde{V}(u_0,v_0) = \frac{1}{\Delta\nu} \int V\left(u_0\frac{\nu}{\nu_0}, v_0\frac{\nu}{\nu_0}\right)\left(\frac{\nu}{\nu_0}\right)^2 g(\nu) e^{2\pi i \frac{\nu-\nu_0}{\nu_0}(u_0 l + v_0 m)}\, d\nu\,, \tag{8–2}$$

and

$$\nu_0 = \text{reference frequency},$$
$$u = u_0\left(1+\frac{\nu-\nu_0}{\nu_0}\right) = u_0\frac{\nu}{\nu_0}\,,$$
$$v = v_0\left(1+\frac{\nu-\nu_0}{\nu_0}\right) = v_0\frac{\nu}{\nu_0}\,,$$
$$\text{and } \Delta\nu = \text{the observing bandwidth}.$$

In general, the effect is to smear $I(l,m)$ with a radially oriented image of the bandpass. This smearing is not a proper convolution since the smearing function is a function of (l,m). A specific example is worked out in the Appendix to this Lecture in which the radial extent of the image of the bandpass is shown to be proportional to $\sqrt{l^2+m^2}\,\Delta\nu/\nu_0$.

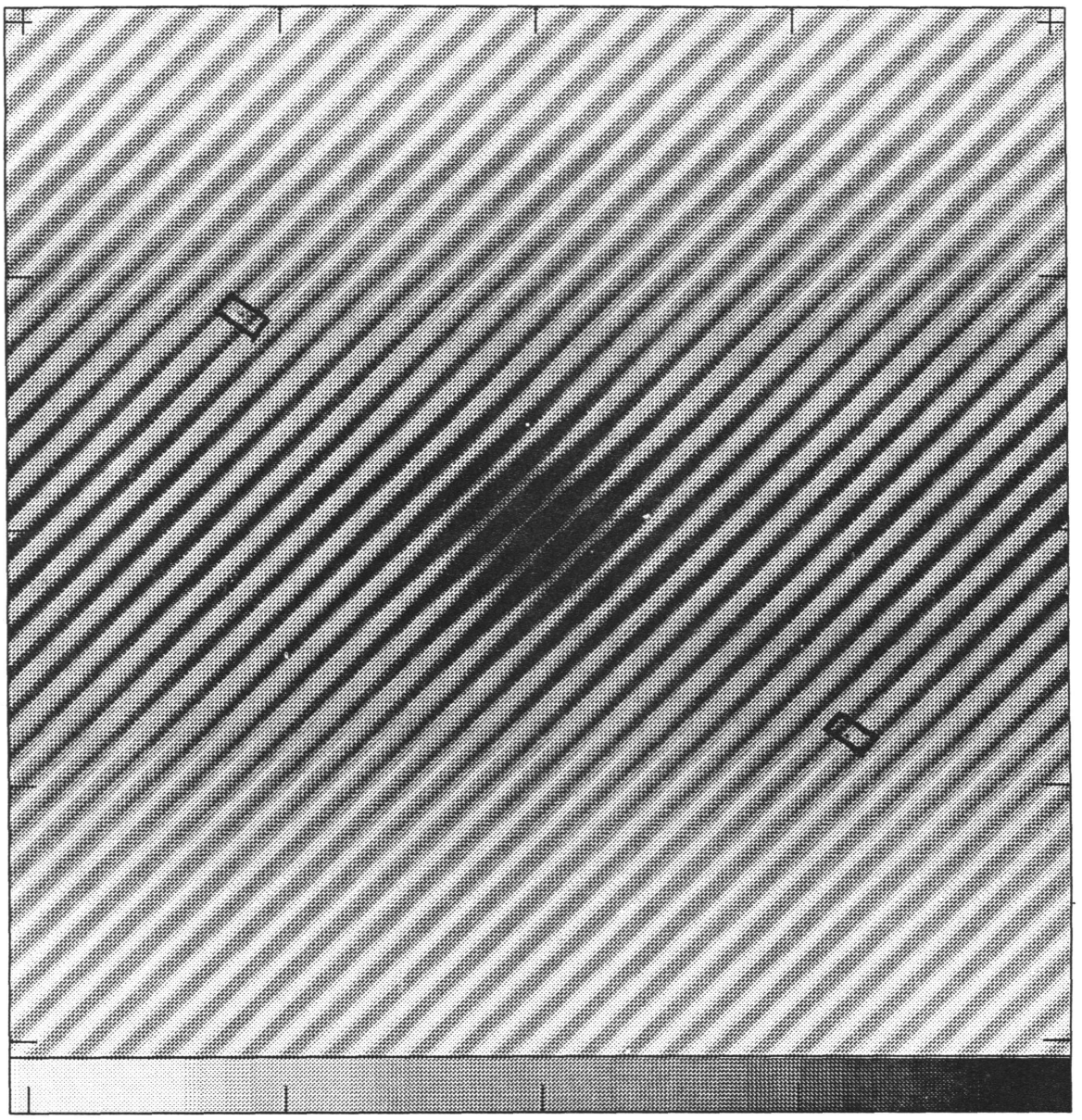


Figure 8–1. The grey scale shows the real part of the inverse Fourier transform (visibility function) of a model source brightness distribution. The boxes indicate the region over which a given data sample might be averaged; the radial extent of the box is determined by the bandwidth, and the asimuthal 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.

The effect of using a finite bandpass is to average over a finite region of the u-v plane. Smearing occurs when the visibility changes significantly in the region over which the averaging takes place, as in Figure 8–1.

Since the averaging due to the bandpass is along a radial line, the smearing in the image plane is also in the radial direction.

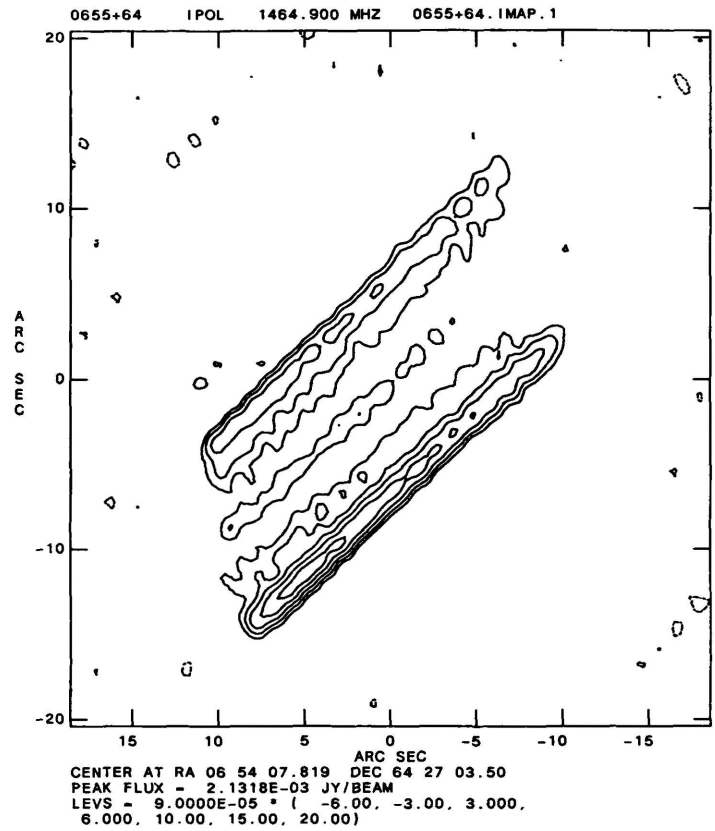


Figure 8–2. The effect of bandwidth smearing on a source 12!9 northeast of the delay tracking center. The smearing is along the radial direction.

A practical example of this effect is shown in Figure 8–2, in which the image of a bandwidth-smeared extragalactic double source is shown. This observation was made with the VLA at 1.4 GHz with a 50 MHz bandpass, and the source was 12!9 from the phase tracking center.

As described above, the width of the smeared image is proportional to the fractional bandwidth—multiplied by a function of the separation (l, m) from the phase tracking center. For sufficiently small fields of view, the smearing has less effect than the convolution with the synthesized beam and is thus relatively unimportant. For a rectangular bandpass function, the degradation of the response of an interferometer to a point source is shown in Lecture 2 and in the Appendix to be:

$$\frac{\sin \pi \frac{\Delta\nu}{\nu_0} \sqrt{u^2 + v^2}\,\theta}{\pi \frac{\Delta\nu}{\nu_0} \sqrt{u^2 + v^2}\,\theta}, \tag{8–3}$$

where θ is the angular distance from the phase center, measured in radians: $\theta = \sqrt{l^2 + m^2}$. A conservative approach is to consider the image to have been substantially distorted if the amplitude on the longest baseline is reduced by more than 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 the region that is of interest. 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. An example of the successful deconvolution of the effects of the source shown in Figure 8–2 from another field is given in Section 1.3 below.

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 up into a number of narrowband 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 relatively small ($\approx$ 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 1982). 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), then there is a significant improvement of the u-v coverage of the observation—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 (1984).

[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.

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 which will result in a distortion of the derived image.

The effects of time-average smearing are much more difficult to analyze than those of bandwidth smearing, because they depend on the time derivative of the observing geometry. Due to the complex 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 will mimic resolution, and the image of a point source away from the phase center will appear 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 will be convolved with the image of the time-averaging function, and the profile of the source will be rectangular.

Figure 8–3 shows a relatively extreme example of the effects of time-average smearing derived from model data. This Figure shows the 'CLEAN' image derived for a given model point source, with and without time-average smearing.

Lecture 2, Section 11, described how to determine whether time-average smearing is a problem compared with bandwidth smearing. The principal reasons for longer integration times 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 are not overwhelming, the simplest solution to time-average smearing problems is to use a shorter 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 which 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

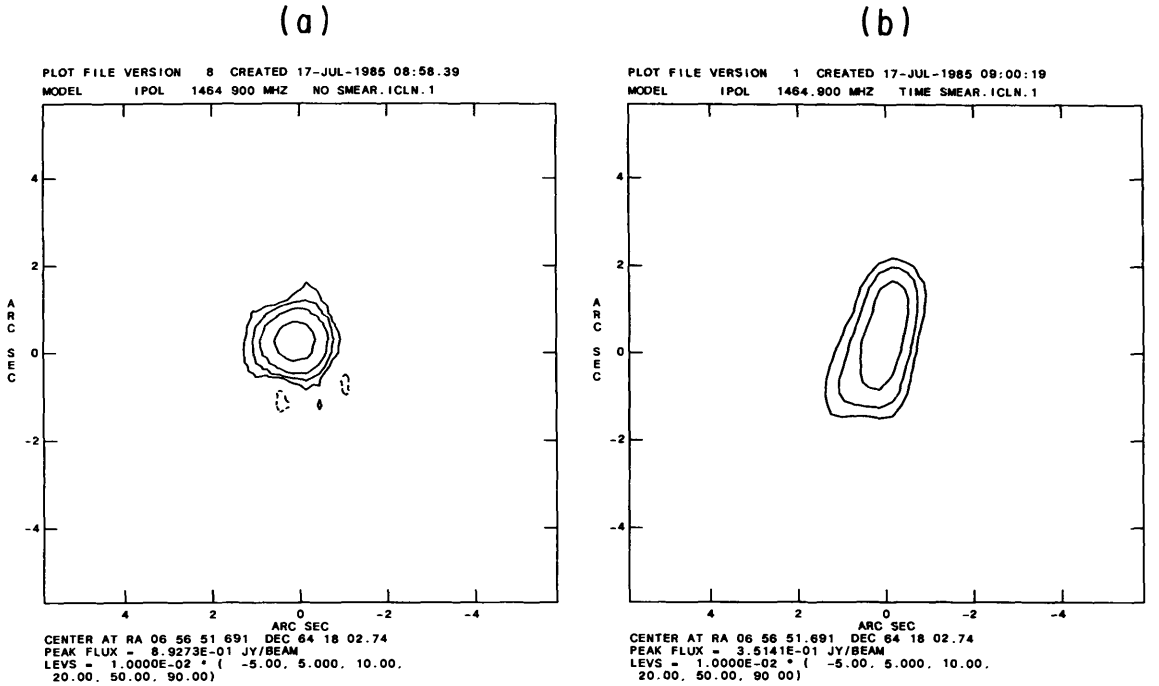


Figure 8–3. (a) shows the 'CLEAN'ed image of a point model ≈ 500 synthesized beamwidths west of the phase center without time-averaging, and (b) shows the 'CLEAN'ed response to averaged data for the same model, showing the effects of time-average smearing.

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 which 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 8–4 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 8–4a clearly shows the sidelobes of distant confusing sources (one of which is shown

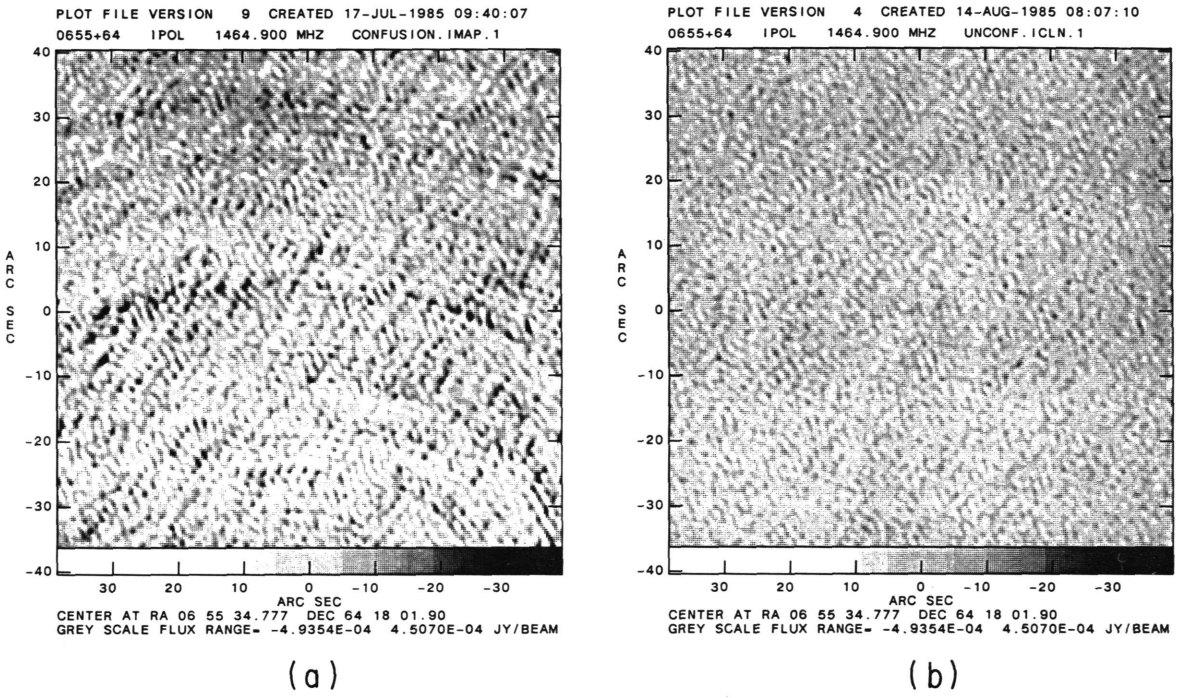


Figure 8–4. (a) The region around a pulsar observed with the VLA at 1.4 GHz, showing the sidelobes of distant, confusing sources. (b) The same region as in (a) with the effects of the confusing sources removed by 'CLEAN'.

in Figure 8–2). 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 very expensive when the image size becomes very large, as in the field shown in Figure 8–4. If most of the region to be imaged is blank, then it is more economical to process only the subregions that are of interest.

Since the sidelobes of sources in one subregion must be removed from the other subregions, the subregions must all 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 8–4b shows the result on the image shown in Figure 8–4a 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 8–4a is 109 μJy and in Figure 8–4b is 62 μJy. It is of interest to note that the bandwidth-smeared image shown in Figure 8–2 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.

In order 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 frequently much more economical to subtract the current 'CLEAN' model from the ungrid-

ded u-v data every so often, then re-grid and re-FFT the data. This approach (termed the Cotton–Schwab algorithm in Lecture 7, Section 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.

A number of 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 which 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) 'direct Fourier transform' of a linear combination of N delta functions (point components), evaluated at a given u, v and w, is given by

$$V(u,v,w) = \sum_{i=1}^{N} A_i e^{-2\pi i(l_i u + m_i v + n_i w)}, \qquad (8\text{–}4)$$

where

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

The role of the w term in Equation 8–4 is to correct 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 which 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 very 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.

In order 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 was shown in Lecture 2 the phase error (in radians) incurred by ignoring the w term is:

$$\text{error} \approx \pi w \theta^2, \qquad (8\text{–}5)$$

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 1973, Hudson 1977, Clark 1981)

For a coplanar array, w in the azimuth of the source is $\approx \sqrt{u^2 + v^2} \sin z$ where z is the instrumental zenith angle. Using this relation, Equation 8–5, and the relation "phase error (in turns, i.e., multiples of 2π)" = "position error (radians)" × "spatial frequency (wavelengths)", the apparent position shift in arcseconds is approximately given by:

$$\text{position error} \approx \frac{\theta^2}{2 \times 2.06 \times 10^5} \sin z\,. \tag{8–6}$$

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 8–6 is small compared to the synthesized beam size, then 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 8–5. This Figure shows model source data for a point 47!5 from the phase center, for VLA u-v coverage obtained at 40° declination. Figure 8–5a shows the image derived for a full track of the object, and Figure 8–5b shows the image derived for a single 30 minute subset of the data. Figure 8–5a shows a gross distortion of the image as the apparent position of the source changes during the day. Figure 8–5b 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

$$w = au + bv\,, \tag{8–7}$$

and there will be a simple geometric distortion of the image which can be corrected. This technique, which is especially useful for east–west arrays, is in use at Westerbork. 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.

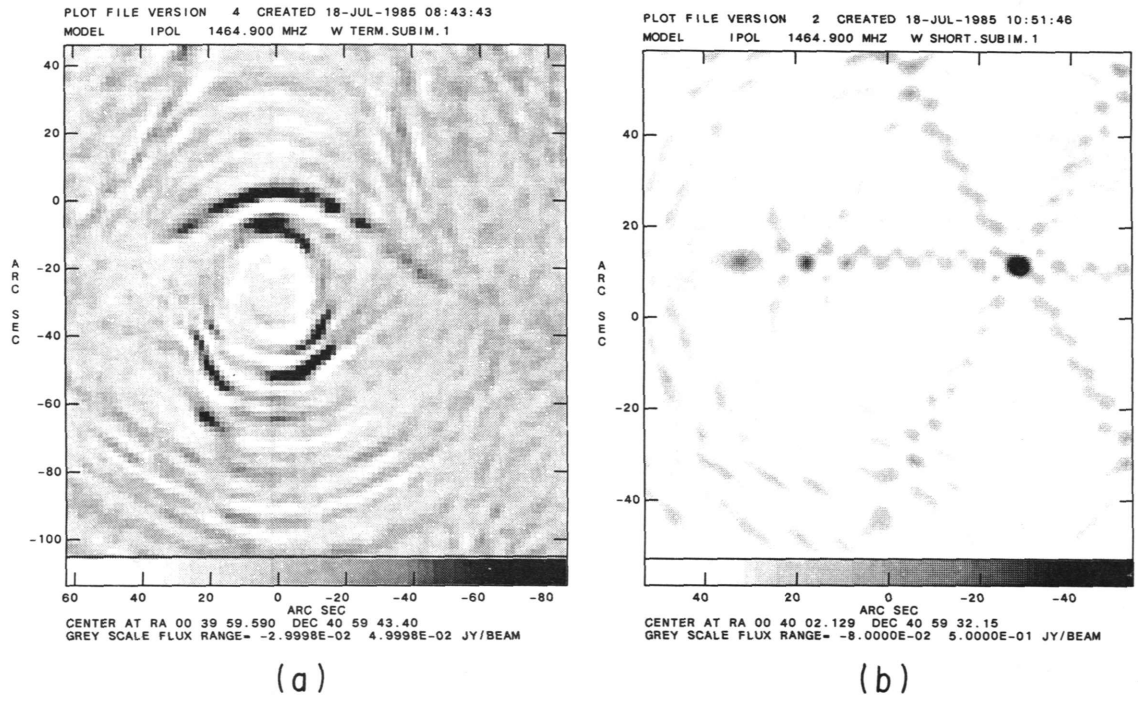


Figure 8–5. (a) The response of the VLA to a point model source 47!5 in RA from the phase center, for full coverage in the VLA B configuration at 1.4 GHs. Zeroes on the axes label the correct position of the source; the model contained 1 Jy, but the peak in the image is 0.071 Jy. (b) Similar to (a), but made using the *u-v* coverage corresponding to only 30 minutes of observation. The peak in the image is 0.948 Jy.

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.

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 (1984) 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 $A(l,m)$ 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 provide a great deal of 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 $A(l,m)$ 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. For a more detailed discussion of this technique, also known as *tesselation*, see Cornwell (1985).

2. Time-variable Effects.

There are a number of time-variable effects which 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 8–6.

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 which appear similar to jets—these are due to the arms of the VLA.

Two approaches which 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

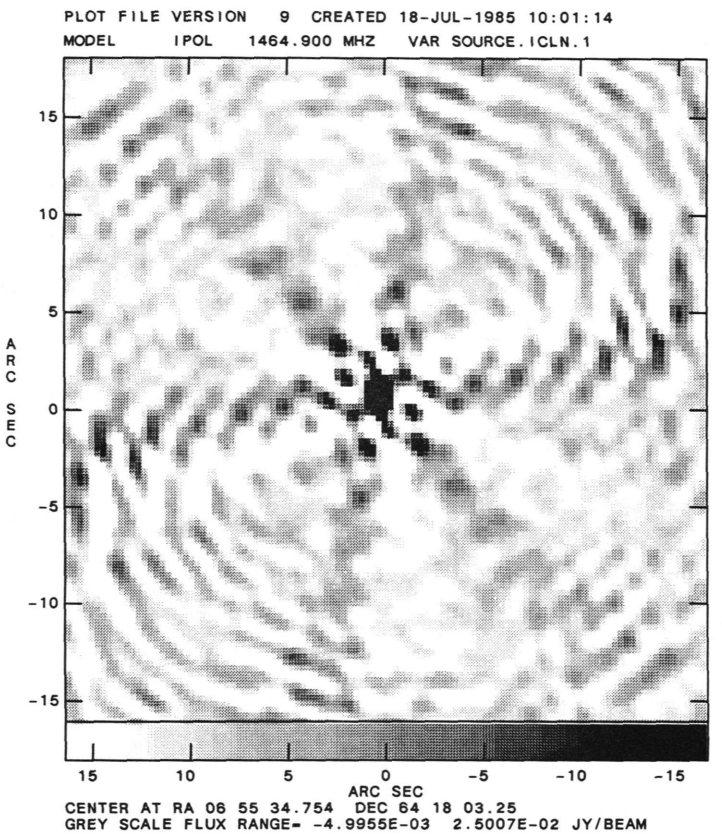


Figure 8–6. The deconvolved ('CLEAN'ed) image derived from model data for a point source with time-variable flux density. The u-v distribution used was that of a source observed with the VLA in the A configuration at 1.4 GHz.

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 be variable on the timescale of the observations. Two examples of this are observations of the sun, which can vary on short timescales, and observations of planets, which rotate. In these cases, if an image is desired, then the data must be divided into sufficiently short time intervals. This will result in relatively poor u-v coverage and correspondingly poor dynamic range. If the desired result can be described by a time-evolving model, such as for VLBI observations of the rapidly changing galactic object SS 433, then the parameters of the model can be fitted directly to the observations.

2.2. Variable sidelobes.

Antennas with altitude-azimuth mounts have the property that the antenna primary beam pattern $A(l, m)$ rotates on the sky. If there are strong confusing sources outside of the main beam of the antenna pattern, then 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. (1973), "Curvature of the sky", VLA Scientific Memorandum No. 107.

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.

Clark, B. G. (1982), "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. (1984), "Broadband mapping of sources with spatially varying spectral index", NRAO VLB Array Memorandum No. 324.

Cornwell, T. J. (1985), "Mosaicing with the mm array", NRAO Millimeter Array Memorandum No. 32.

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

APPENDIX

An Example of the Bandwidth Smearing Effect. Let us consider a specific example of the bandwidth smearing of a unit amplitude point source. Since the smearing is due to the averaging along a radial path in the u-v plane, we can consider the one-dimensional case, with no loss of generality. Using the shift theorem (Bracewell 1978) the visibility function becomes

$$V(u) = e^{-2\pi i u l_0}\,, \tag{A8–1}$$

where $l_0 \equiv$ the location of the source. Further, assume a rectangular bandpass function which is given by

$$g(\nu) = \begin{cases} 1, & \text{if } |\nu - \nu_0| < \Delta\nu/2\,, \\ 0, & \text{otherwise}\,. \end{cases} \tag{A8–2}$$

The relation between intensity and visibility can then be explicitly stated by averaging over frequency:

$$I(l) = \int_{-\infty}^{\infty} \frac{1}{\Delta\nu} \int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2} e^{2\pi i u l} e^{-2\pi i u l_0}\, d\nu\, du\,. \tag{A8–3}$$

Expressing u explicitly as a function of frequency,

$$u = u_0 \left(1 + \frac{\nu - \nu_0}{\nu_0}\right)\,,$$

and $du = \frac{\nu}{\nu_0} du_0$. Since the fractional bandpass can be assumed to be small, ν/ν_0 will be close to unity and can be ignored. Rewriting the expression for the intensity,

$$\begin{aligned} I(l) &= \int_{-\infty}^{\infty} \frac{1}{\Delta\nu} \int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2} e^{2\pi i \left(l u_0 \left(1+\frac{\nu-\nu_0}{\nu_0}\right) - l_0 u_0 \left(1+\frac{\nu-\nu_0}{\nu_0}\right)\right)}\, d\nu\, du_0 \\ &= \int_{-\infty}^{\infty} \frac{1}{\Delta\nu} e^{2\pi i u_0 (l-l_0)} \int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2} e^{2\pi i \frac{\nu-\nu_0}{\nu_0} u_0 (l-l_0)}\, d\nu\, du_0\,. \end{aligned} \tag{A8–4}$$

Let us for the moment consider the inner integral—substitute $\nu' = \nu - \nu_0, d\nu' = d\nu$. Then

$$\frac{1}{\Delta\nu}\int_{-\Delta\nu/2}^{\Delta\nu/2} e^{2\pi i\nu' u_0 \frac{l-l_0}{\nu_0}}\, d\nu' = \frac{\sin 2\pi\nu' u_0 \frac{l-l_0}{\nu_0}}{2\pi\nu' u_0 \frac{l-l_0}{\nu_0}} - \frac{i\cos 2\pi\nu' u_0 \frac{l-l_0}{\nu_0}}{2\pi\nu' u_0 \frac{l-l_0}{\nu_0}}\Bigg|_{-\Delta\nu/2}^{\Delta\nu/2}$$

$$= \frac{\sin \pi\Delta\nu u_0 \frac{l-l_0}{\nu_0}}{\pi\Delta\nu u_0 \frac{l-l_0}{\nu_0}} \qquad \text{(A8–5)}$$

$$\equiv \operatorname{sinc} \frac{\Delta\nu u_0 (l-l_0)}{\nu_0}.$$

Equations A8–5 yield the result which earlier was stated without proof (Eq. 8–3).

By the convolution theorem, Equation A8–4 can be rewritten as

$$I(l) = \int_{-\infty}^{\infty} e^{2\pi i u_0 (l-l_0)}\, du_0 * \int_{-\infty}^{\infty} \operatorname{sinc}\left(\frac{\Delta\nu u_0 (l-l_0)}{\nu_0}\right) e^{2\pi i u_0 l}\, du_0\,, \qquad \text{(A8–6)}$$

where $*$ denotes convolution. The first integral corresponds to our initial model; i.e., $\delta(l_0)$. Bracewell (1978) solves the second integral, which gives what we will call the smearing function $S(l)$. Recognizing the Fourier transform of the sinc function as a unit step function, and applying the similarity theorem (see Bracewell 1978), we get

$$S(l) = \frac{1}{\left|\frac{\Delta\nu}{\nu_0}(l-l_0)\right|}\,\Pi\left(\frac{l}{\frac{\Delta\nu}{\nu_0}(l-l_0)}\right), \qquad \text{(A8–7)}$$

where

$$\Pi(s) \equiv \begin{cases} 1\,, & \text{if } |s| < \frac{1}{2}\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Again applying the approximation $\frac{\Delta\nu}{\nu_0} \ll 1$, we find that the width of this function is $l_0 \frac{\Delta\nu}{\nu_0}$, as was asserted in the text for the two-dimensional case.

9. Self-Calibration

TIM CORNWELL

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.

Ordinary calibration (see Lecture 4) relies upon the monitoring of the variable quantities in the array by frequent observations of a calibrator source of known structure, strength and position. The relationship between the visibility $V_{ij,\mathrm{obs}}$ observed at time t on the i-j baseline and the true visibility $V_{ij,\mathrm{true}}(t)$ can be written very generally as:

$$V_{ij,\mathrm{obs}}(t) = G_i(t)G_j^*(t)G_{ij}(t)V_{ij,\mathrm{true}}(t) + a_{ij}(t) + \epsilon_{ij}(t)\,. \tag{9-1}$$

The terms $G_i(t)$ and $G_j(t)$ represent the effects of the complex gains of the array elements i and j; the term $G_{ij}(t)$ represents the non-factorable part of the gain on the i-j baseline; $a_{ij}(t)$ represents an offset term and $\epsilon_{ij}(t)$ is a pure noise term due to the thermal noise. The effects $G_{ij}(t)$ and $a_{ij}(t)$ which factor per ***baseline*** can usually be eliminated to a satisfactory degree by clever design (see Lecture 3), so I will mainly ignore their presence hereafter. Equation 9–1 can then be simplified to

$$V_{ij,\mathrm{obs}}(t) = G_i(t)G_j^*(t)V_{ij,\mathrm{true}}(t) + \epsilon_{ij}(t)\,. \tag{9-2}$$

For simplicity I have neglected the effects of time averaging and finite bandwidth, discussed in Lectures 2 and 8; 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 Lecture 4).

Given a calibrator 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 correction of 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 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 (Cornwell 1981):

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

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.

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 calibrator 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 are left. 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

I 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 be factored out as two element-related terms. 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 I will show that for large enough numbers of array elements this effect is not too important since the ratio of constraints to degrees of freedom increases.

It is clear what one can expect to lose by letting the element gains be free variables but the 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. I will examine these in turn.

2.1. Redundant calibration.

Suppose that the geometry of the interferometer array is arranged 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 the solution of 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

[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.

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 only be 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 7). Hence self-calibration is really just another method like 'CLEAN' (Lecture 7, Section 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 $\hat{I}$ of the sky intensity distribution, the Fourier transform $\hat{V}$ of which, when corrected by some complex gain factors, reproduces the observed visibilities to within the noise level. The model $\hat{I}$ 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 7). One convenient method (Schwab 1980) of obtaining such an agreement is to minimize, by adjusting both the complex element gains G_i, G_j and the model intensity distribution $\hat{I}$, the sum of squares of residuals

$$S = \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) |V_{ij,\mathrm{obs}}(t_k) - G_i(t_k) G_j^*(t_k) \hat{V}_{ij}(t_k)|^2 , \tag{9–4}$$

where the $w_{ij}(t_k)$ are weights (purely from signal-to-noise considerations these should be set to the reciprocals of the variance of the $\epsilon_{ij}(t_k)$). 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 Section 5.3).

An interesting and illuminating connection to ordinary calibration is apparent if Equation 9–4 is re-expressed as:

$$S = \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) |\hat{V}_{ij}(t_k)|^2 |X_{ij}(t_k) - G_i(t_k) G_j^*(t_k)|^2 , \tag{9–5}$$

where:

$$X_{ij}(t) = \frac{V_{ij,\mathrm{obs}}(t)}{\hat{V}_{ij}(t)} . \tag{9–6}$$

Division by the model visibilities $\hat{V}_{ij}(t)$ turns the object being imaged into a pseudo-point source, though admittedly with rather strange receiver noise, which 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 $\hat{V}_{ij}(t)$ 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 we 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

$$V_{ij,\mathrm{corr}}(t) = \frac{V_{ij,\mathrm{obs}}(t)}{G_i(t)G_j^*(t)} . \tag{9–7}$$

(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 1980) and in Lecture 7 we showed that both 'CLEAN' (Section 2) and the Maximum Entropy Method (MEM, Section 4) 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 $G_i(t)$ to maximize the image entropy

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

subject to:

$$\begin{aligned} S &= \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) |V_{ij,\mathrm{obs}}(t_k) - G_i(t_k) G_j^*(t_k) \hat{V}_{ij}(t_k)|^2 \\ &= \text{expected value} , \end{aligned} \tag{9–9}$$

and:

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

where $\hat{V}_{ij}(t)$ 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 it 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 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 I will concentrate on self-calibration. First, however, I 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:

$$\phi_{ij,\mathrm{obs}}(t) = \phi_{ij,\mathrm{true}}(t) + \theta_i(t) - \theta_j(t) + \text{noise term}\,, \tag{9–11}$$

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 $C_{ijk,\mathrm{obs}}(t)$, known as the observed *closure phase*[1], is given by:

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

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

$$\Gamma_{ijkl}(t) = \frac{A_{ij}(t)A_{kl}(t)}{A_{ik}(t)A_{jl}(t)}\,, \tag{9–13}$$

where the A's here denote the amplitudes of the complex gains. 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) 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.

[1]This terminology is similar to that of closing, or closure, errors in traversed loops, used by surveyors. — *Eds.*

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 Equation 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 optimum 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-arc-second 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, 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 non-linear 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 7 for further details).

Self-calibration fails either when the signal-to-noise ratio is sufficiently poor or when 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.

5. Practical Problems in Self-Calibration

I 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 11.

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[1]. 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

[1]See Lecture 11 for discussion of possible exceptions to this rule. — *Eds.*

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 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 $V_{ij,\mathrm{obs}}(t)$ 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 optimum averaging time, τ_{sc}, obviously depends upon the timescale for gain changes and upon the source strength. The error in the gain estimate due to the receiver noise on a nearly unresolved source is (for good signal-to-noise ratio), for amplitude and phase correction,

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

and, for phase correction,

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

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 optimum 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 that the errors in the gain estimate are 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 (1982) 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:

$$S = \sum_k \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k) \left| V_{ij,\mathrm{obs}}(t_k) - G_i(t_k) G_j^*(t_k) \hat{V}_{ij}(t_k) \right| . \tag{9-16}$$

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 should provide superior results. Averaging of the data also alleviates this problem since seriously discrepant points are downweighted in the averages $\langle V_{ij}/\hat{V}_{ij} \rangle$.

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 4 and 12).

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:

$$S = \sum_k \sum_l \sum_{\substack{i,j \\ i \neq j}} w_{ij}(t_k, \nu_l) |V_{ij,\mathrm{obs}}(t_k, \nu_l) - G_i(t_k) G_j^*(t_k) \hat{V}_{ij}(t_k, \nu_l)|^2 . \tag{9-17}$$

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, in 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 R. Linfield: 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 not 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 7) 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 be 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, 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. I shall merely summarize some of these (see Clark's memo for more 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 is comparable to the main lobe of the primary beam $A(l, m)$ of the array elements.
(4) Non-isoplanaticity of the atmosphere, i.e., different parts of the field of view to be imaged are seen through different cells in the atmosphere. Schwab (1984) has described a solution to this problem.
(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 11.

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.

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

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, W. Va.), 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.

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

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.

Schwab, F. R. (1980), "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.

Schwab, F. R. (1984), "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 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 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) $\simeq 10^6$. Any point

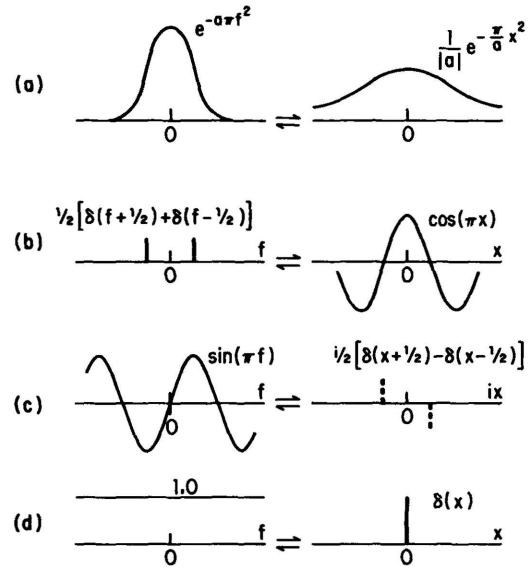


Figure 10–1. Fourier transform pairs. The broken lines indicate the imaginary domain. For many more examples see Bracewell (1978).

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

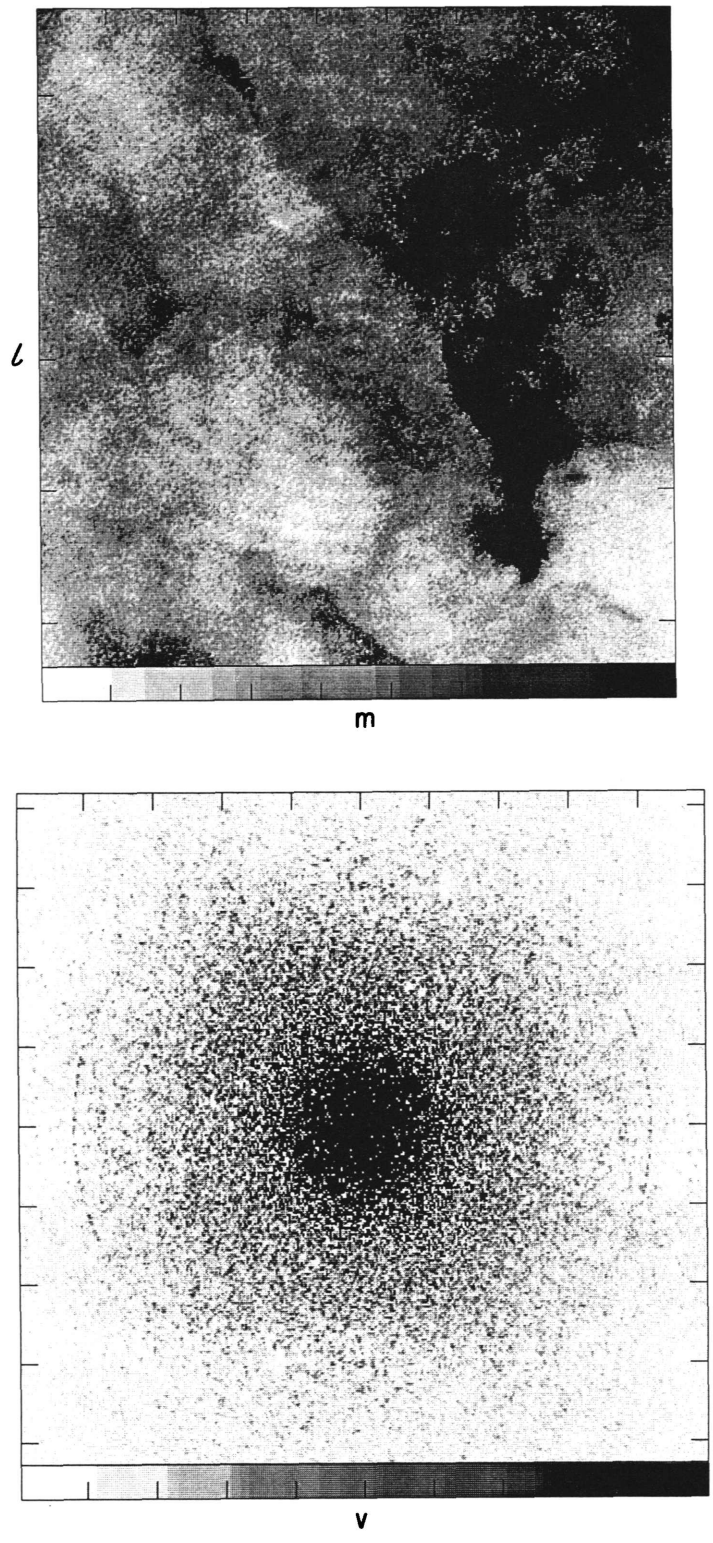


Figure 10–2. *(Top)* The inner portion of an image of Cassiopeia A, centered on the phase tracking center. A baseline-based error which persisted throughout the observations caused the concentric rings in this image. *(Bottom)* The (inverse) Fourier transform of the above image. The two curved linear features near the left and right edges of the display correspond to the locus of the error-corrupted interferometer baseline.

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).

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 5).

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 8) 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.

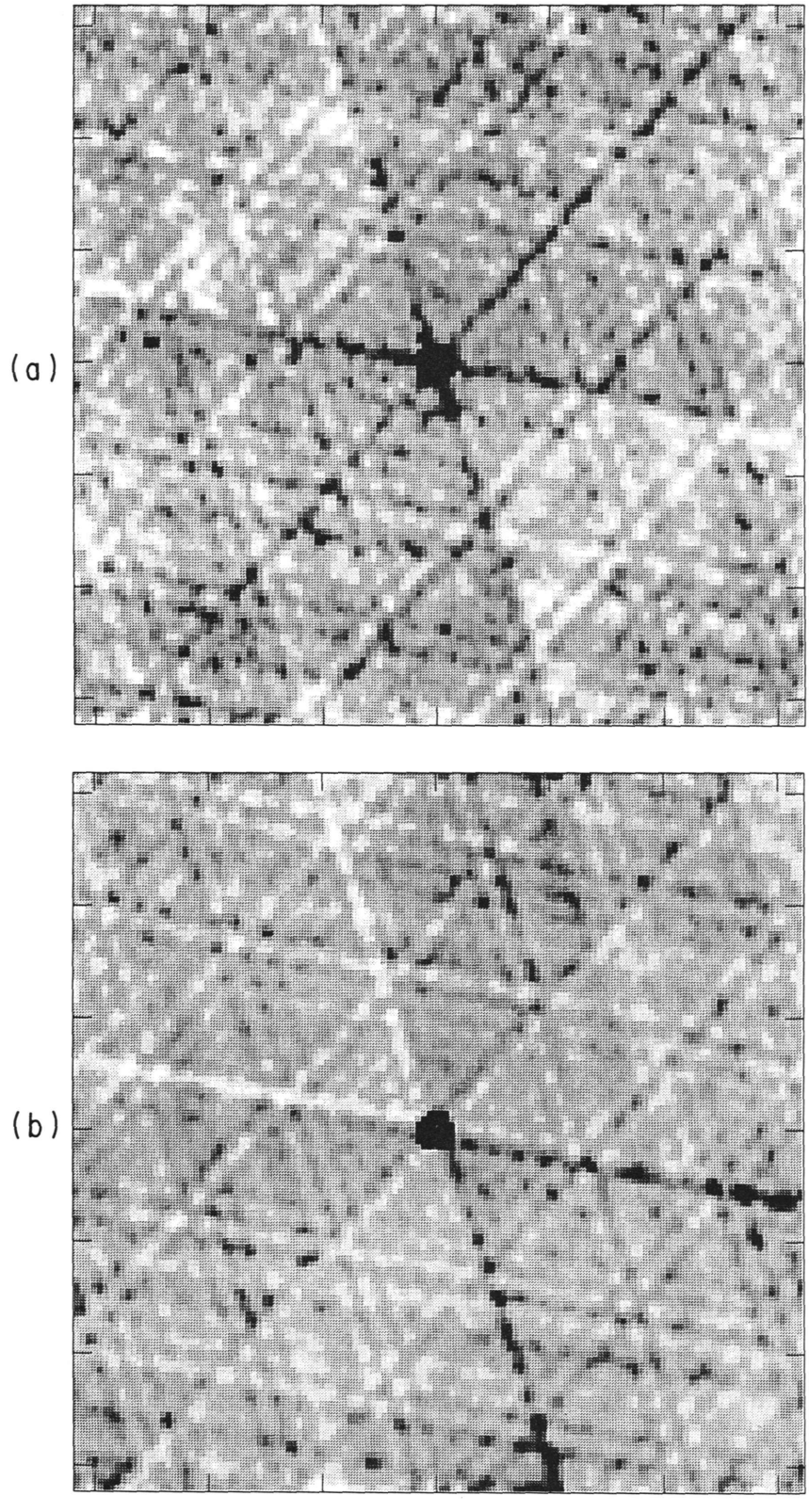


Figure 10–3. Images from a "snapshot" observation of a point source **(a)** with a 10% amplitude error on one antenna and **(b)** with a 10% phase error on one antenna.

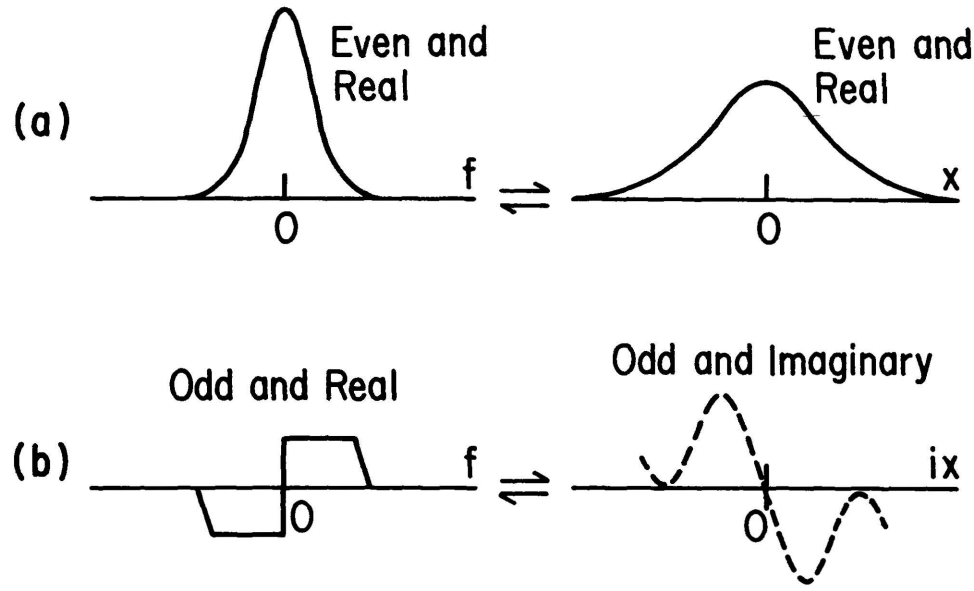


Figure 10–4. Fourier transforms of symmetric and asymmetric functions. The broken line indicates the imaginary domain.

3. EXAMPLES

3.1. Additive errors.

These are errors of the form (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 8. 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

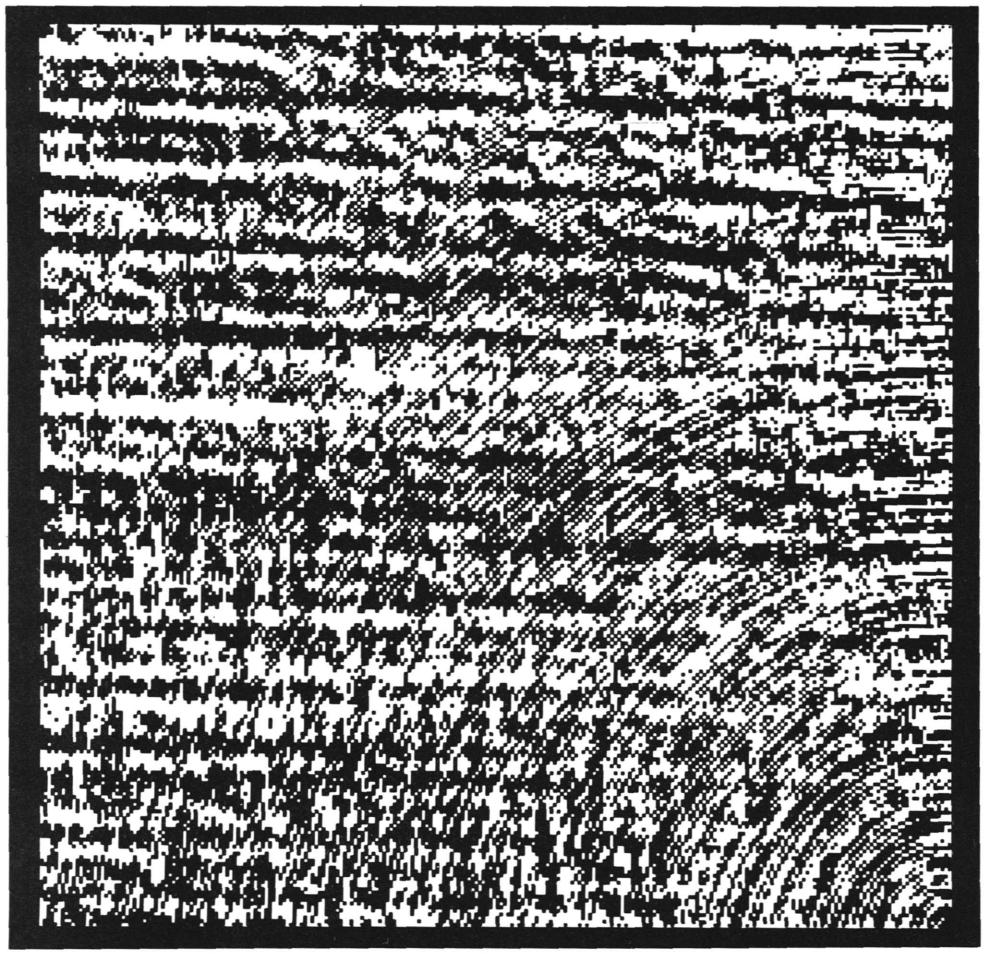


Figure 10–5. Top left quadrant of an image of a point source (bottom right corner) with continuous narrow band interference.

fringe rate is near zero. The example in Figure 10–5, taken from Thompson (1982), shows the result of a constant source of interference on an 8 hour observation. The interference has caused horizontal stripes through the image because the only baselines for which the expected fringe rate is zero are those which project to a North–South orientation.

Another way to look at this is to note that a source of interference at a *fixed* position is like a source at the North pole. Hence the pattern of horizontal stripes is just a small section of a set of rings concentric with the North pole.

3.1.3. Cross-talk. This is the same kind of effect as external interference, except that the interfering signal is generated in one antenna and transmitted to another. Since it usually occurs between close antennas, it is a more serious problem in compact arrays (such as the D configuration of the VLA), and it results in an error in the image with a very large angular scale.

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 6. 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 would as 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 (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 7, we normally attempt to correct its effects by using some form of deconvolution algorithm e.g., 'CLEAN'. How well we do depends on the size of the unsampled regions and their location relative to significant structure in the visibility function, especially near $u = v = 0$. The problem is that when you look at your raw image it is difficult to distinguish the effects caused by the missing information from effects caused by errors in the measured data. Our main clue about the nature of effects caused by the missing information is in the point spread function (dirty beam). The sidelobes of this dirty beam are the (negative) Fourier transform of the missing information. If an image has features around the sources which look just like the sidelobe pattern of the dirty beam then this is most likely to be an effect of the missing u-v spacings. If we see effects which have a very different shape then they *may* be caused by errors in the data. But beware of the following complication: When making this assessment we use the dirty beam to give us a way to gauge the effect of the missing information on a *point source*. The sidelobe pattern for an *extended source* is not the same. A very extended source is affected only by the information that is missing at *short* u-v spacings. Although this information is included in the point source response function, it may be present with such low 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 which 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

[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.*

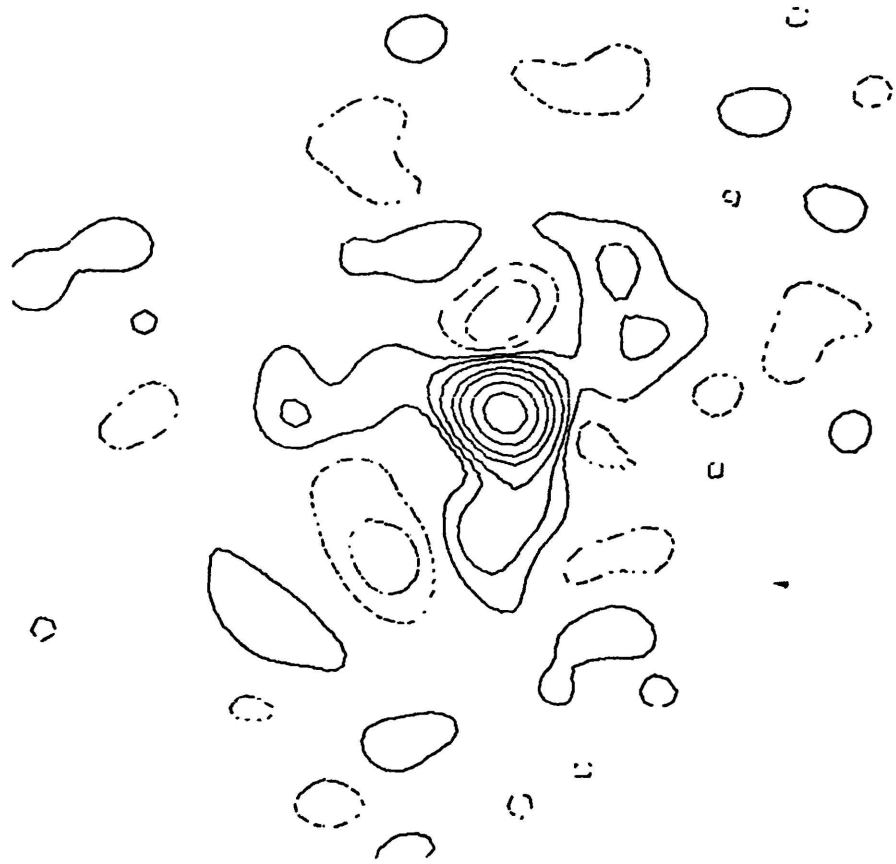


Figure 10–6. Asymmetric pattern which could be caused by atmospheric phase errors.

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 4 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

$$V(u)e^{-2\pi iu\Delta},$$

and then by the shift theorem for Fourier transforms we have

$$V(u)e^{-2\pi iu\Delta} \rightleftharpoons I(l-\Delta).$$

Hence, the effect of an atmospheric phase error is to shift the position of the source. For longer baselines, a random phase error will be introduced; this will cause the image to be convolved with an asymmetric error function (see the example in Fig. 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 (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 have been discussed extensively in Lectures 2 and 8. Their characteristics are easy to recognize in an image, since the 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 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 which can be different for each element and which 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. This type of error is discussed extensively in Lecture 8, Section 3.

3.4. Computational errors.

A number of additional errors can be introduced by the computational methods used to produce a final image. Since these have all been 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 5. The effect of having noncoplanar interferometer baselines and finite computing precision are discussed in Lecture 8. Errors may also be introduced by the image restoration algorithms (e.g., 'CLEAN') and the self-calibration technique; these are discussed in Lectures 7 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 Section 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 11).

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—to introduce a point to be discussed in Lecture 15: 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. (1982), "The response of a radio astronomy synthesis array to interfering signals", *IEEE Trans. Antennas Propagat.*, **AP-30**, 450–456.

11. High-Fidelity 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, Westerbork, and MERLIN. The production of the images is by no means automatic, for the data are invariably corrupted by a host of errors, due both to 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 noise near the theoretical limit. The complexity of these algorithms, especially in regard to their interaction with the user, requires that users be familiar with their use and effect. It is important to realize at 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. This Lecture is intended to discuss and elucidate these techniques and to display their potential. Real observations of a familiar source will be used to illustrate the results of various steps in the process of image improvement.

This Lecture title includes 'Fidelity', rather than 'Dynamic Range', and it is important to understand the distinction. Common usage has the latter term meaning the ratio between the peak brightness on the image and the r.m.s. noise in a region believed to be void of emission (such regions, fortunately, are commonplace in astronomy). 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 calculated lower limit to the error in the brightness of a non-void region. The true error distribution is non-uniform; indeed, the errors of calibration must result in effects which behave like the sidelobes in the beam. These 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 included. These are due to inadequate sampling of the u-v plane, and will result in an imaging error which varies from pixel to pixel. The estimation of these spatially dependent errors appears to be impossible, or at least impractical.

The term 'Image Fidelity' is meant to describe how close the resultant image is to the true brightness distribution. The problem with this generalization is that in the absence of knowledge of the true image, the errors cannot be calculated. Some studies have been made, using artificial sources, to assist the planning of array configurations. From these studies there arise no simple rules, other than the self-evident one that the more complete the u-v plane coverage, the better the image fidelity. Another approach, experimental in nature, is to observe an object on different days, with different u-v plane coverages, but sufficient to allow reliable deconvolution. Independent processing (by different individuals, if deemed important) will likely result in slightly different images. These can form a sort of 'ensemble average' by which errors of the sort mentioned can be estimated. Only limited work has been done in this way.

Lacking a firm method for estimating image fidelity, we must fall back upon experience and intuition. Experience with considerable quantities of VLA data has shown that self-calibration of data of simple, strong, isolated objects results in reduction of the 'noise'

in regions of no known structure, and an increase in 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—and it is therefore believed that the improvement in dynamic range which usually accompanies self-calibration actually represents an increase in image fidelity. As the process of self-calibration continues, 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. Thus, this ratio shall be employed in this Lecture as the parameter of choice in judging the fidelity of the images which result from processes discussed at length in this Lecture. The reader should 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.

It must be emphasized that the techniques described below are not for every database. It is necessary that every observer calculate the expected thermal noise, and compare this to what is attained. If the actual noise on the image is equal, or close, to the expected, and if there are no large artifacts (strictly speaking, there can't be, if the noise is as expected), then no further processing of the type to be described is required. Don't waste time in pursuit of impossible goals!

The principles behind the techniques employed and discussed in the following have been discussed in detail in the previous Lectures. In particular, the Lectures on Calibration (#4), Self-Calibration (#9), Deconvolution (#7), and Error Recognition (#10) are of particular relevance.

2. Dynamic Range—Possibilities and Realities

2.1. Definitions and origins of important errors.

Lectures 4 and 9 discussed calibration of interferometric data at length, so only a brief review will be given here. The relation between the true visibility, V'_{ij} and the observed visibility, V_{ij}, can be written

$$V_{ij} = G_i G_j^* G_{ij} V'_{ij} + a_{ij} + \epsilon_{ij} .$$

Here, G_i and G_j are the factorable antenna gains for antennas i and j, G_{ij} is the non-factorable multiplicative baseline-dependent gain ($G_{ij} - 1$ is commonly called the 'closure error'), and the remaining terms describe baseline-dependent offsets and thermal noise, respectively. All quantities can be considered to be functions of time. By proper design, the offset terms can be reduced to levels important only for extremely deep searches for weak sources—a situation in which high image fidelity will rarely be important. Conversely, the situations in which high fidelity is expected will rarely be affected by these additive offsets. Attention will thus be restricted to the gains G_i and G_{ij}.

It is useful to consider the origin of these errors. An antenna-based gain is a (complex) product of the antenna gain, normally a slowly varying quantity, and an atmospheric component, which varies on short time scales. The antenna gains in a well designed system vary slowly, so they can be calibrated *a priori*. At the VLA, the amplitudes of the gains are pre-calibrated (through periodic observations of sources with known flux density) at most bands to an accuracy of a few percent (although in units of tens of Janskys), but there is no attempt to pre-calibrate the antenna phases. The atmospheric 'gain' is almost entirely a phase effect, with absorption effects becoming important only at very low (less

than 20 MHz), and very high (greater than 20 GHz) frequencies. The time variability can be extreme, with changes of radians in minutes being possible on longer baselines.

The baseline-based errors are mainly instrumental in origin. The important exception is the effect of background sources. If the summed flux density of background sources approaches that of a calibrator source, closure errors will occur if a point source model is used. This effect will exceed the instrumental errors at wavelengths longer than about 20 cm. This limitation is not fundamental since a correct representation of the calibrator—that is, including the background sources—will allow proper calibration. The instrumental contributions to baseline-dependent errors are many. Recent studies done for the VLA (Bagri, 1986) have indicated two main contributions. The first is incorrect delay settings. For 50 MHz bandwidth, 2 nsec delay errors will result in amplitude closure errors of approximately 5%. The obvious solution is to set the delays more accurately—however, these settings drift in time, and more studies are required to determine the optimum combination of accuracy and repeatability. The second, and more fundamental, contribution comes from errors in the quadrature networks, the Hilbert transform devices described in Lecture 3 (Sec. 2). These wideband 90° phase shifting devices, used in the VLA continuum system, are analog devices with an inherent error of 1%–2% in amplitude, and 1°–2° in phase. In VLA spectral line mode the quadrature networks are not used, so this error is not present (see Fig. 3–5 of Lecture 3). Furthermore, in spectral line mode the effects of delay errors are unimportant, since these effects are inversely proportional to bandwidth and images are made for each channel. The best solution to dynamic range limitations is likely to include use of spectral line mode.

In principle, calculation of both types of gain errors can be based on observations of bright, isolated point sources. Standard calibration, as described in Lecture 4, seeks to calculate only the antenna-based gains, which, in nearly all cases, contain the largest errors. However, the residual errors, due to non-isoplanatic[1] effects and inadequate calibrator source models, will vary from a few degrees to many radians. The effect of these errors is to limit the dynamic range to values of tens to a few hundred. Since this is not adequate, self-calibration can often allow dynamic ranges of up to 20,000. Beyond this, correction of baseline-dependent errors is required. The techniques to do this are given in Section 3. Before this, it is instructive to consider the effects on an image of these errors.

2.2. Effects of calibration errors on imaging.

I here apply some simple arguments to allow a rough estimate of the best dynamic range achievable with various classes of errors. For simplicity, I consider only one dimension—expansion to two is trivial. I consider phase and amplitude errors separately, beginning with phase.

Consider a single 'snapshot' observation 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

$$V(u) = \delta(u - u_0)e^{-i\phi}$$

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

[1]Non-isoplanatic effects are important when the phase in the direction of the calibrator is appreciably different from that in the direction of the source. The magnitude of an important difference depends on the dynamic range.

integral gives a contribution of $2\cos(2\pi u_k l)$. (The factor 2 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

$$I(l) = 2\phi\sin(2\pi u_0 l) + 2\sum_{k=1}^{N(N-1)/2}\cos(2\pi u_k l)$$

while the beam, or point spread function is

$$B(l) = 2\sum_{k=1}^{N(N-1)/2}\cos(2\pi u_k l)\,.$$

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 subtraction of 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

$$D = \frac{N(N-1)}{\sqrt{2}\phi} \sim \frac{N^2}{\sqrt{2}\phi}\,,$$

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 with the substitutions

$$\phi \rightarrow \epsilon \quad \text{and} \quad \sin \rightarrow \cos\,.$$

An important conclusion from this exercise is the following:

A 10° phase error is as bad as a 20% amplitude error.

Since 10° degree phase errors are commonplace, while 20% amplitude errors are rare, it is clear that phase correction is by far the most important component of self-calibration.

Suppose now the error is antenna-based. Thus, instead of one bad baseline, there are N. Then, again assuming incoherence in the noise (which is approximately right), the dynamic range becomes

$$D = \sqrt{\frac{N}{2}}\frac{N-1}{\phi} \sim \sqrt{\frac{N}{2}}\frac{N}{\phi}\,.$$

If all baselines have random errors of this magnitude, the dynamic range is decreased from the single correlator case by a factor $\sqrt{N(N-1)/2}$, giving

$$D = \frac{\sqrt{N(N-1)}}{\phi} \sim \frac{N}{\phi}\,.$$

Extension of these results to many snapshots is straightforward, as the signal rises with the number of observations, M, and the noise, presuming incoherence, with $\sqrt{M}$. Thus, for the last case, the dynamic range becomes

$$D = \frac{\sqrt{M}\sqrt{N(N-1)}}{\phi} \sim \frac{\sqrt{M}N}{\phi}.$$

These equations, though simplistic, return reasonable results. They predict that for a single snapshot with 27 antennas, residual calibration errors of order 10° will limit the dynamic range to approximately 2500 if confined to a single baseline, 500 if due to an antenna, and 100 if equally distributed amongst all antennas. After self-calibration, the residual errors are of order 0.5°, so the limiting dynamic ranges are a factor of 20 higher. The observed limiting dynamic range of $\sim$ 10,000 is then predicted if the number of 'independent' scans is approximately 20 to 30. These considerations can also be used to estimate the required phase accuracy needed to attain the theoretical best dynamic range. Anticipating a result derived in the next section, the maximum potential dynamic range attainable for the VLA is of order 10^7, —the required phase tolerance is better than 10^{-3} degrees.

These simple considerations indicate that residual calibration errors of a few degrees are responsible for the limiting dynamic ranges attained in imaging. Before embarking on a detailed description of the techniques for removing these errors, it is useful to enumerate other sources of error which may ultimately limit the dynamic range of an image.

2.3. Other forms of errors.

The ultimate dynamic range will be that set by the thermal noise. This has been derived in Lecture 6, and can be written

$$\delta I = \frac{C}{\sqrt{N(N-1)\Delta\nu\tau}},$$

where C is a constant depending upon the antenna size and efficiency, the receiver 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 little more to be done, and the time that would have been wasted can now be used for some other purpose. Note that the potential dynamic range for some objects is enormous. For example, consider the famous quasar, 3C 273. The core flux density is approximately 35 Jy, while the theoretical noise, for the VLA with 35 hours of observing at 6 cm, is about 10 μJy. Thus, the theoretical maximum dynamic range exceeds 3 million! Even more extreme examples can be imagined (or concocted).

Before the thermal limit is reached, other factors may be important. We should consider the following, incomplete list:

2.3.1. Van Vleck correction. This time-dependent baseline-based error, discussed in Lecture 3, cannot be corrected by the techniques described here. It is important only for very strong objects such as Cygnus A (especially at lower frequencies), and for spectral line observations of masers.

2.3.2. Calculation errors. These take many forms. Errors in baseline coordinates cause spatially dependent errors in the image. Gridding errors 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 use of the Fast Fourier Transform, is often important, especially in smaller databases which have vastly less data than the number of u-v cells to be filled. Use of 16-bit integers in imaging gives an interesting, and unnecessary, limitation of approximately 65,000 in dynamic range. Round-off errors in internal calculations are expected to show up in the 1–10 million range in dynamic range.

2.3.3. 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.[1] What one recovers is a spatially filtered image, —and often a bad one at that, since the transfer function is rarely 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 is significantly larger than the reciprocal of the source angular size, the deconvolution algorithms cannot possibly reconstruct correctly.[2] 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.

3. Techniques of Error Correction

It is clear that the antenna- and correlator-based errors are the most important limitations to high dynamic range imaging. I will now discuss the techniques developed over the past few years which are employed in accurately removing these errors. I will illustrate my remarks with examples of the improvement of images of the well-known quasar 3C 273.

3.1. 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 (e.g., the 'ANTSOL' listings at the VLA). Keep in mind that background sources will cause apparent closure errors varying from in excess of 10% at 90 cm to less than 1% at 2 cm. These will not degrade dynamic range, since they will

[1] 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.

[2] 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.

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 3.3. Before flagging in this manner, be sure to calculate, using the concepts presented Section 2.2, the magnitude of the error which will be important. If the expected dynamic range is low (say, < 100), the tolerance to a closure error is very high, and correlator flagging will usually be unnecessary, and often undesirable.

3.2. Antenna-based error correction using self-calibration.

If the noise on an initial image is significantly above that expected, and 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. Very 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 3C273, as illustrated in Figure 11–1. Due to the large quantity of data, only 10% of the data are actually plotted—however, this is 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 most 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 that in this method of self-calibration, any positional information on the source position provided by the original phases is lost.

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 was number 11, at which point 33 Jy had been removed, I would have used 10 components as the model, with a u-v restriction of 120 Kλ to 650 Kλ. In this case, the safe approach is to calibrate the phases alone. After this, a new image would be created, and the subsequent deconvolution should produce more positive components than before. This initiates a second round of self-calibration, with both amplitude and phase solutions. When employing amplitude solutions, it is usual and advisable to 'float' the gains—renormalize the gain solutions so 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.

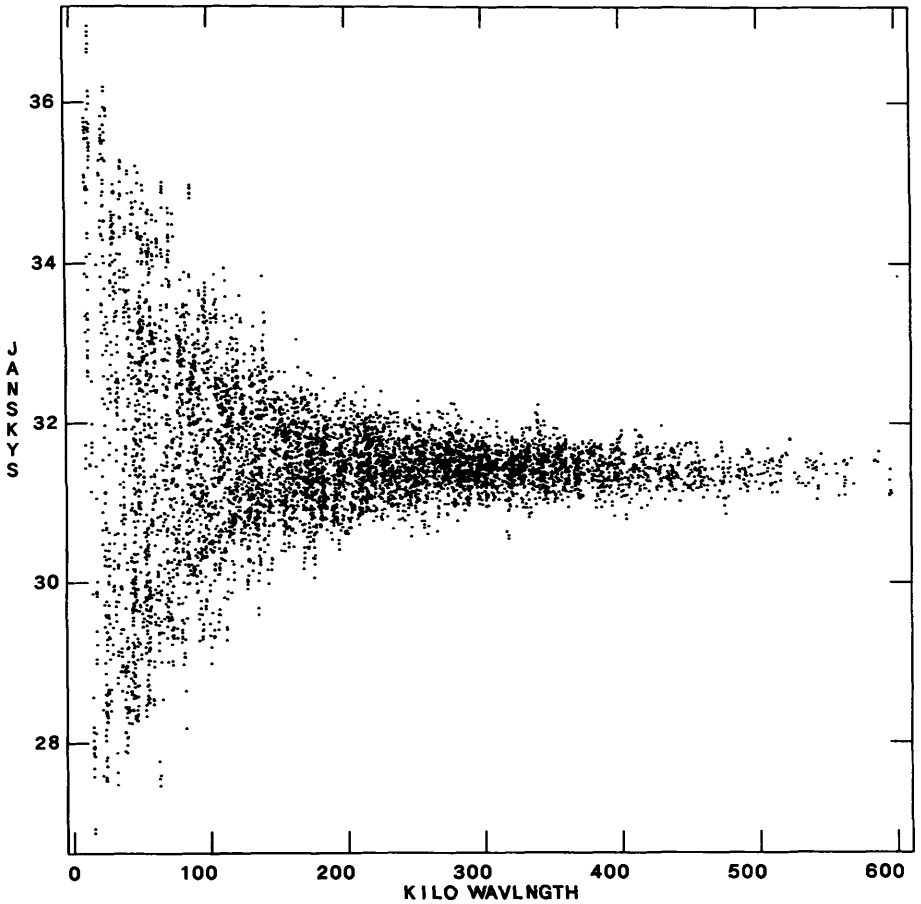


Figure 11–1. The visibility plot of 3C 273 at 6 cm in the A configuration. The 'trumpet horn' shape is indicative of a point-dominated source with larger-scale structure which is completely resolved out on longer spacings. A point source of flux density 31.5 Jy provides an excellent initial model for self-calibration.

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 array configuration is shown in Figure 11–2. A point-source model is obviously a poor choice here, as the secondary is not resolved out on any baseline. The usual procedure is to make a dirty image and deconvolve—however, in this case, the first negative 'CLEAN' component came before the first component from any point other than the core, so the usual prescription would return only a point-source model. This situation was the result of some very poorly edited data—I was was a little lax in my standards, since I was sure that self-calibration would work! Under these circumstances, the A configuration image was used to self-calibrate the B configuration data, with a u-v restriction to baselines longer than 50 Kλ.[1] Because the core flux density had changed between epochs of observation, a phase-only solution was made.[2] The resulting improvement was spectacular, and is shown in Figure 11–3. This gave a much improved model for the second round of self-calibration, allowing phase and amplitude gains to be simultaneously calculated.[3] The result of this is also shown in Figure 11–3. It will be immediately apparent that the noise is

[1]This restriction is not required, but is a useful precaution as the A configuration undersamples the u-v plane in this region.

[2]The visibility phases are less sensitive to a change of flux than are the visibility amplitudes.

[3]Adherence to the usual rules in amplitude self-calibration can introduce siguificant errors. Specifically, if a strong point source lies between two cells in the image plane, and the model employs only positive

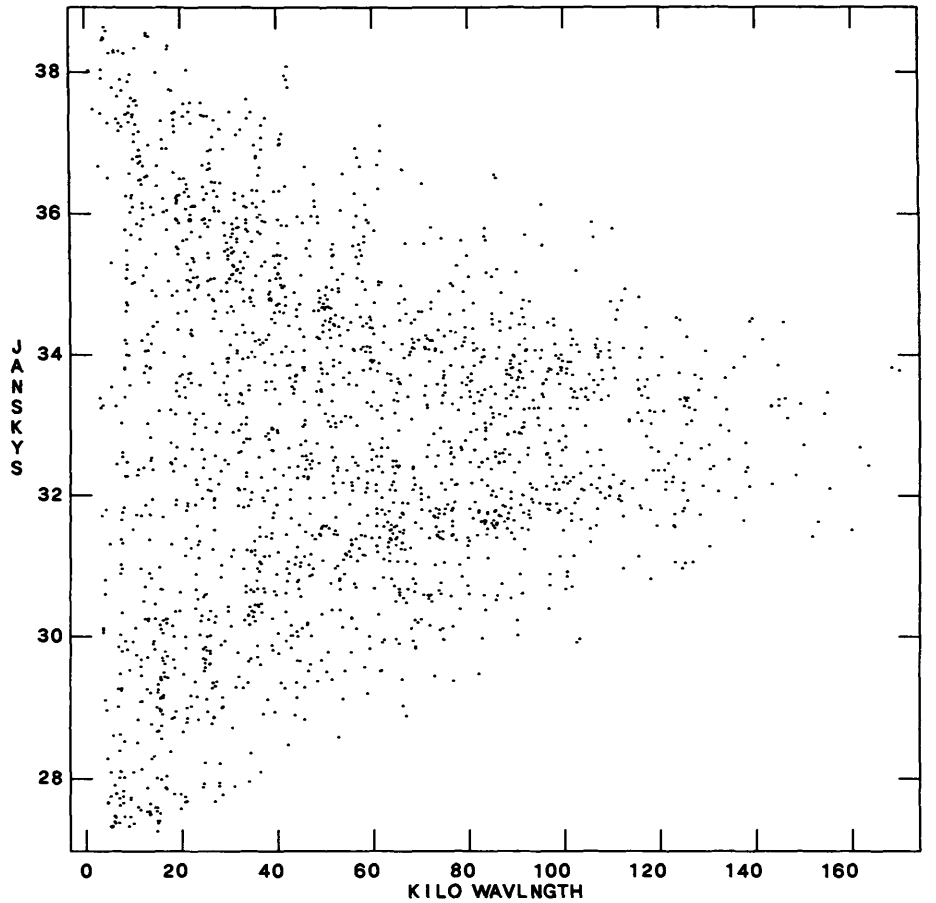


Figure 11–2. The visibility function of 3C 273 at 6 cm in the B configuration. Here the secondary noted in the previous figure is not resolved out, so self-calibration will benefit by a model more sophisticated than a point source.

not uniform. This is generally true for all objects, but is especially so for this object since its declination is 2°. The distribution of noise must follow the beam sidelobes, since the origin of the noise is limited to sampled u-v cells. At most declinations, the distribution of filled cells is sufficiently uniform to allow the noise distribution in the image to appear uniform. However, at low declinations, all tracks are nearly E-W, so the resultant noise is primarily distributed N-S. At high declinations, all tracks are circular, and the resulting noise patterns are similarly 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—for example, due 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

components, it is easy to see that the self-calibration will force a double source model. With good signal and good data, this situation will be avoided if negative components are also included. However, these simultaneous requirements are rarely met, and the only solution allowed by current software is to greatly oversample the beam—using 5 or even 10 points per beam. However, this obviously requires much larger images, and a better solution might be to allow fractional-cell cleaning. This problem is unimportant for partially resolved objects, or when the point objects are weak.

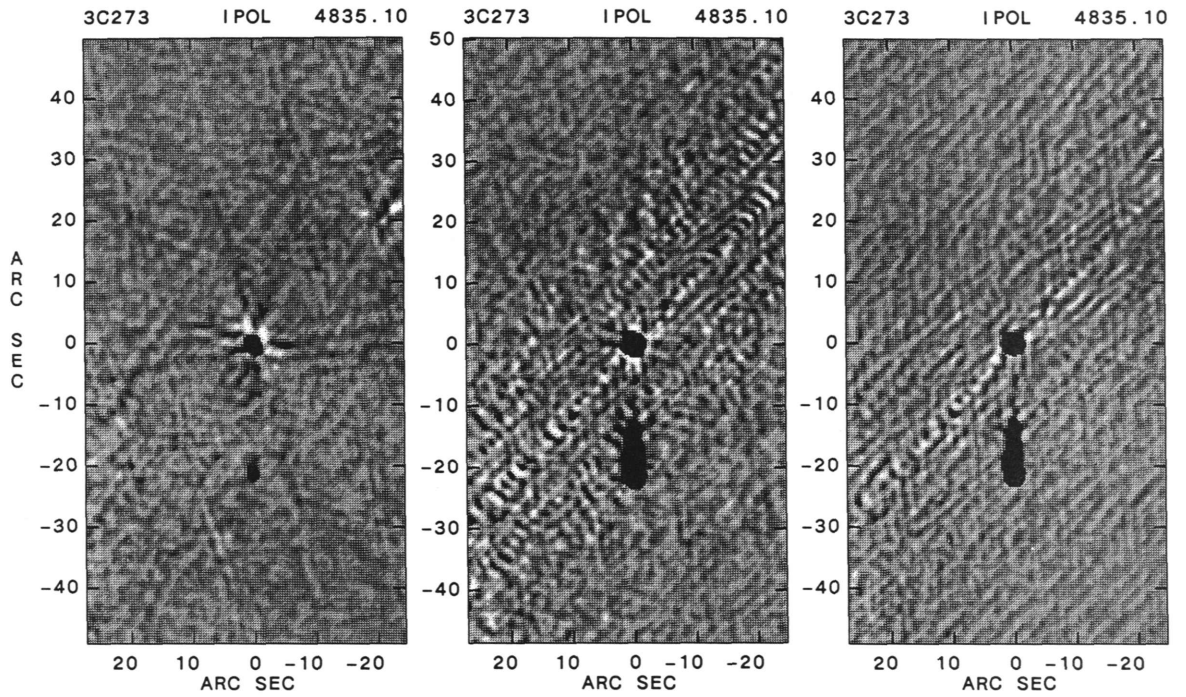


Figure 11–3. Images of 3C 273, made from B configuration data, demonstrating three stages of self-calibration. All three images have been rotated to make the structure vertical. *(Left)* The image without any self-calibration. The greyscale extends from −1 to 1 Jy/ beam. The peak is 28.2 Jy/beam, the 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.

are in strong disagreement with the best image. 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 11–4 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 now flagged. The subsequent image, shown in Figure 11–5, improved dramatically. Considerable time and effort could have been saved had I done the required flagging initially.

The techniques given 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 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.

3.3. Baseline-based error correction.

The principles of baseline-based calibration are identical to those of antenna-based calibration. By measurements of strong, isolated sources, estimates of the (complex) mul-

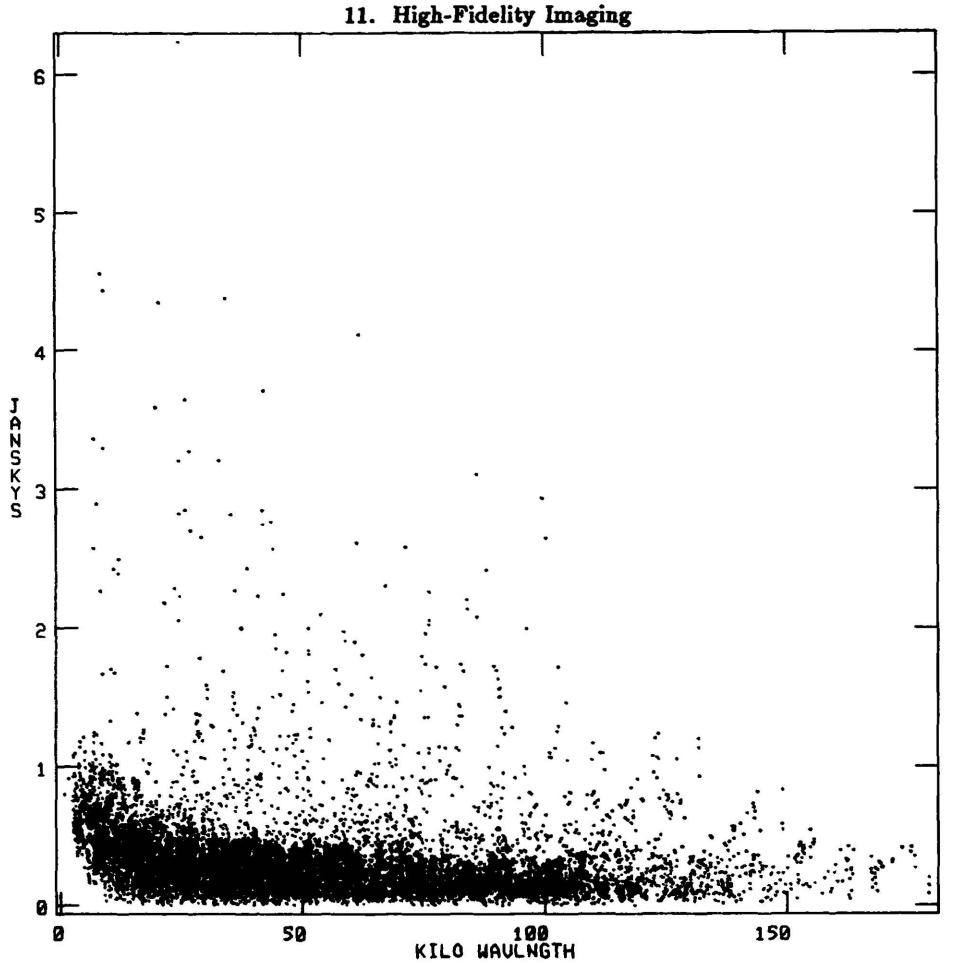


Figure 11–4. A plot of the visibility amplitudes after the inverse Fourier transform of the model has been subtracted. The points lying under 1 Jy represent normal residual closure errors, while the points scattered above this are all due to a malfunctioning correlator.

tiplicative 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. The software within AIPS necessary to make these calculations has only recently become available, so that the extensive testing required to properly and fully utilize the new technique have not yet been completed. However, the results so far are extremely encouraging, and I present below the current ideas for implementation of this technique.

Let me 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 to 10,000 (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 without baseline-based calibration. If traces of the beam sidelobes are still present, then this form of calibration is likely to be effective. For example, the residual sidelobes in Figure 11–3(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 0.5° in phase. The distribution of errors is non-Gaussian, with a few baselines showing

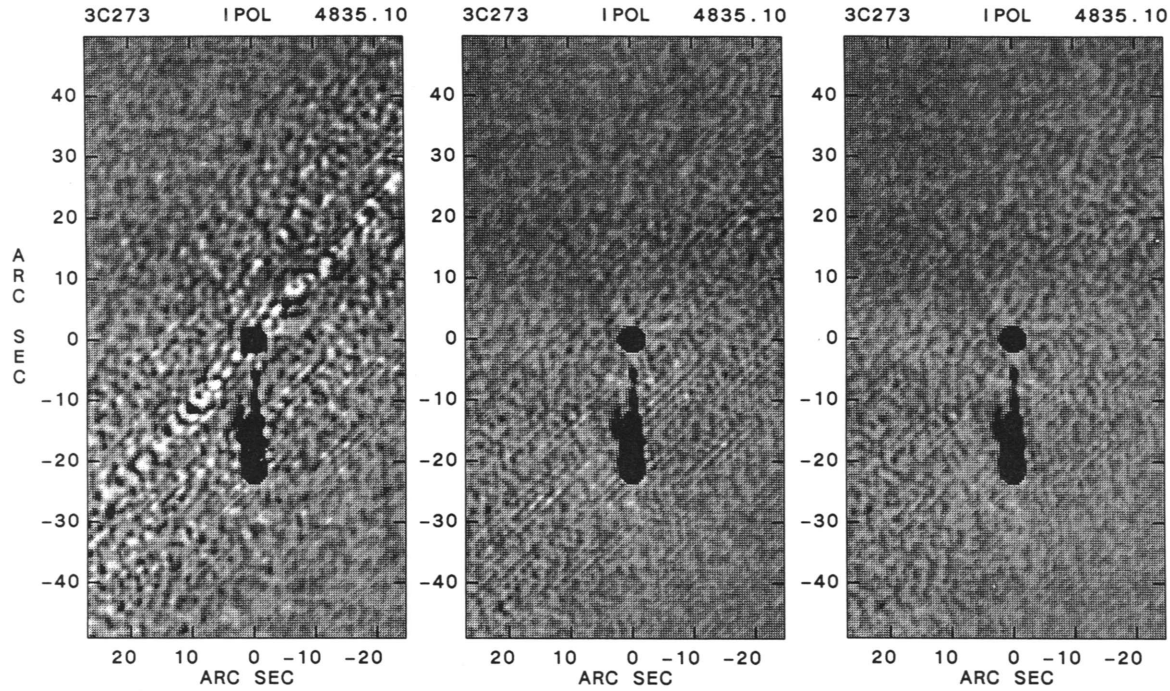


Figure 11–5. Images of 3C 273 after further stages in processing. All are shown with greyscale wedges from −10 to 10 mJy/beam. *(Left)* After self-calibration, followed by removal of a strongly variable correlator. The r.m.s. noises are 2.5 and 0.9 mJy/beam in the N-S and E-W directions respectively. *(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.

3 to 5 percent/degree errors. 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 ABSOLUTELY ESSENTIAL 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 'BCAL1'.

(4) Apply the corrections to the source database. Within AIPS, this is done by 'BCAL2'.

Some results of applying this procedure to observations of 3C 273 are shown in Figures 11–5 and 11–6. In almost all situations, one must use a strong, isolated, simple source for these corrections. However, for this case, 3C 273 itself satisfies these criteria, so I have used it to calculate its own closure errors (a procedure called by some 'incestuous self-calibration'). The reason for using a simple isolated source for closure corrections is to prevent 'pre-defining' the structure.[1] 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 only be effective 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 11–5(left), is shown in Figure 11–6. Any unmodeled flux shows up as a drift of the mean ratio away from 1.0—so in this case it appears that all the flux is represented. Note that the closure levels are as expected, generally less than 1%. Application of the mean correlator-based offsets to the data results in the image shown in the center panel of Figure 11–5. 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 11–5. The dynamic range here is 78,000.

The dynamic range of the image shown in Figure 11–5 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. It is hoped that the necessary changes to software will soon be made to allow this suspicion to be tested. I have already noted that the baseline-dependent errors are very much smaller with the spectral line mode operating, and tests have shown that the dynamic range achievable in this mode is at least a factor of four higher than in continuum. Thus, it is likely that the best results will come from the use of this mode. Unfortunately, at present this means giving up polarization capabilities.

3.4. Coverage errors.

It may 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 11–7(a) 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 secondary. These are a result of inadequate short-spacing u-v coverage. Recall that any given VLA array 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.

[1] To better grasp this problem, note that closure corrections are modifying every visibility. If a time-varying correction is calculated from a model of a source, and applied back to the data at each time the correction is calculated, then clearly the data will be modified to exactly reproduce the model. Thus, for example, a double source could be turned into an unresolved point source.

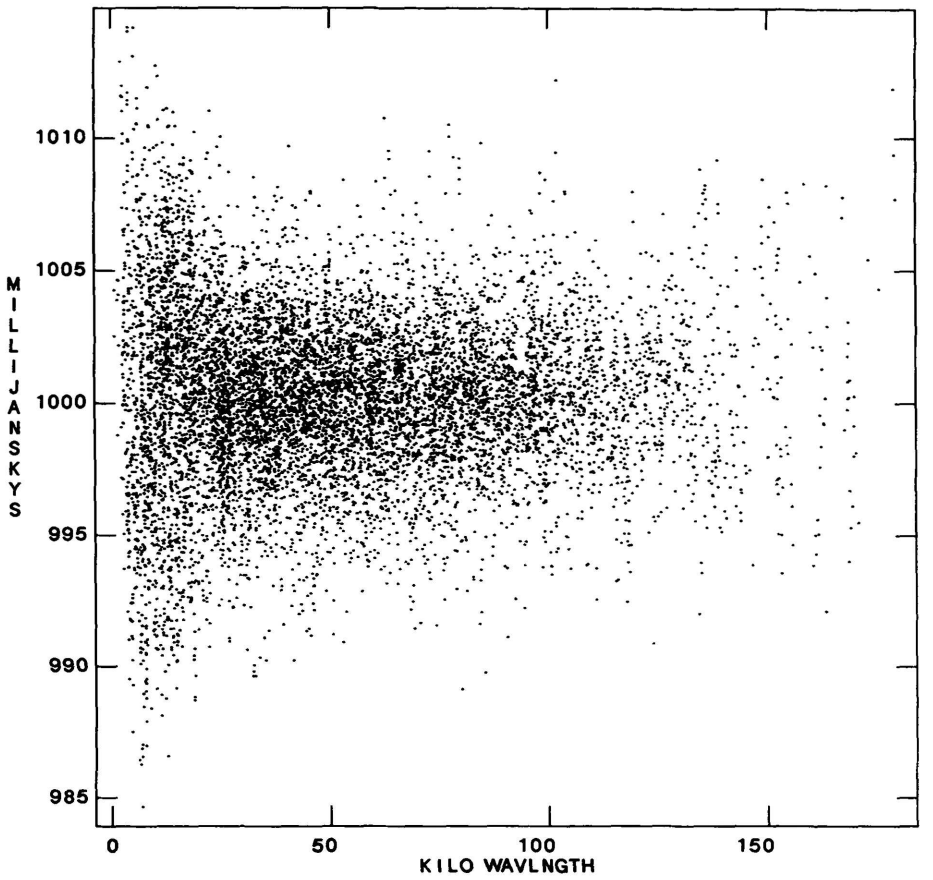


Figure 11–6. The visibility after division by a model provided by the image after self-calibration. The scatter about 1.0 is due to multiplicative closure errors

For this example, the resolution is approximately 0.35 arcsec, so any structures larger than approximately 7 arcseconds 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 indicates 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

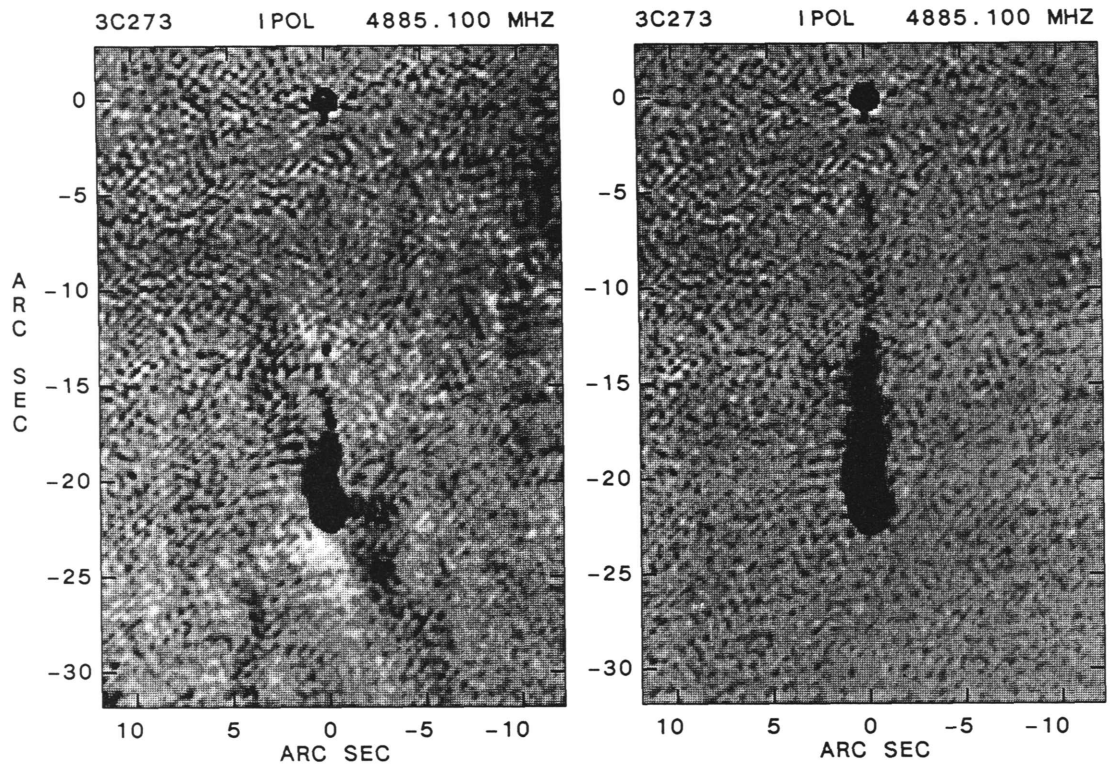


Figure 11–7. Illustrating the effect of missing u-v coverage on 3C 273 at 6 cm. In both panels the grey scale runs from −10 to 10 mJy/beam. The left panel shows the image using **A** configuration data only, so that structures of scale $\geq 7''$ are severely attenuated. Note the large scale undulations 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.

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 was of excellent quality, and it 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 slightly different frequencies (different 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.

12. Spectral Line Imaging

JACQUELINE H. VAN GORKOM AND RONALD D. EKERS

1. INTRODUCTION

In this Lecture, problems specific to multi-frequency aperture synthesis observations are discussed. Such observations may be required either in order to measure kinematic or physical conditions in line-emitting regions or, in continuum observing, when the monochromatic approximation is inadequate (see, for example, Lecture 8). Spectral synthesis differs from single-band synthesis in more ways than the number of frequencies. These 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 one of the largest and most uncertain errors in single dish observations—is not a problem, because the receiver noise is not correlated between elements.

Usually the bandpass calibration is performed by observing a calibrator for sufficient time to reach the required signal-to-noise ratio. Bandpass normalization by the autocorrelation 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., in the case of galactic HI observations), then one can do the bandpass calibration 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 aperture synthesis telescopes, the spectral line capability is provided by a digital cross-correlation spectrometer. Although this avoids many of the errors involved in an analog system, it does introduce an additional effect, the Gibbs phenomenon, which must be considered. The Gibbs phenomenon is the ringing around sharp changes in the frequency spectrum which occur because of the truncation of the temporal cross-correlation measurements. The consequences of the Gibbs phenomenon can be more serious in spectral line image synthesis 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 lacks these symmetries, and 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 visibility phase is nonzero (see Lecture 3). 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 in the field (usually at the phase center). The effect can be attenuated by a suitable tapering of the cross-correlation measurements prior to 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 situation where it might be advisable to use a different cross-correlation taper for the source than for the calibrator. A uniform taper can be used for the source spectra to obtain maximum frequency resolution, while Hanning smoothing should be used for the calibrator. In this case, the passband edges will 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 the calibrator coincide (e.g., when both are at the field center). The same weighting must be used for source and calibrator. It can 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 does become more confined to the neighborhood of the discontinuity. In particular, as the number of channels is increased, the Gibbs effect that is due to the bandpass skirts becomes more highly concentrated at the band edges. Since 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 was discussed in Lectures 2 and 8, synthesis radio telescopes have an inherent chromatic aberration which results from the formation of an image by adjusting the phase $\Delta\phi$ of the correlated signals for each point in the image plane (the Fourier transform relation), instead of the arrival time of the wavefront $\Delta t = \Delta\phi/2\pi\nu$ at each point, as would occur in the focal plane of an imaging optic. This aberration causes a radial smearing which increases linearly away from that point in the image for which the time delays have been equalized by the delay tracking system. It is commonly known as the delay beam smearing.

In order to overcome this defect it is necessary to subdivide the band into channels which are sufficiently narrow to have negligible aberration, and to use the actual frequencies of each channel rather than the band center frequency. In most spectral line observations the channel bandwidth criterion is automatically satisfied by the spectral resolution requirements, but for broadband continuum observations using spectral line mode the number of channels required will be determined by the smearing.

Even if there is negligible delay beam smearing in the individual spectral line channels, the sidelobe structure due to the array geometry will still change with frequency, and this will have to be taken into account in the analysis of the spectral line data cube. The effects of frequency-dependent sidelobes could be removed by use of 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 the image will all be in the correct place, but the

sidelobe pattern will be changing from channel to channel by a radial scaling factor proportional to frequency. This is preferable if the spectral line images contain a lot of weak, frequency-dependent structure spread out 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. In some spectral line observations, very high spectral dynamic range is required. 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—one of the main sources of error in a continuum observation—are independent of frequency and can be removed by subtracting a continuum image formed by averaging frequency channels outside the line emission. Errors that limit the spectral dynamic range are multiplicative errors (see Lecture 10) that differ from one frequency channel to the next because bandpass calibration is not perfect. The effect of baseline-dependent errors has not been investigated, but it is believed to cause problems at a level much lower than residual antenna-based errors (see Lecture 11). The Fourier transform of the error term is convolved with the strong continuum source distribution, and thus if the errors differ from one channel to the next, then that is exactly what remains when the continuum is subtracted. Examples are bandpass calibration errors in amplitude and phase. The patterns that arise from gain 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. In Figure 12–1, examples are shown of the effects of very small phase errors on images of the weak recombination line emission from the compact HII region DR21. In one example a small frequency-dependent phase error causes an apparent (and erroneous) rotation of the HII 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 very powerful way of locating errors is by looking at the inverse 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 any deconvolution is done. 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 intensity of the continuum emission is not very high. If the frequency dependence of the sidelobe structure is important, then it may be necessary to subtract the inverse

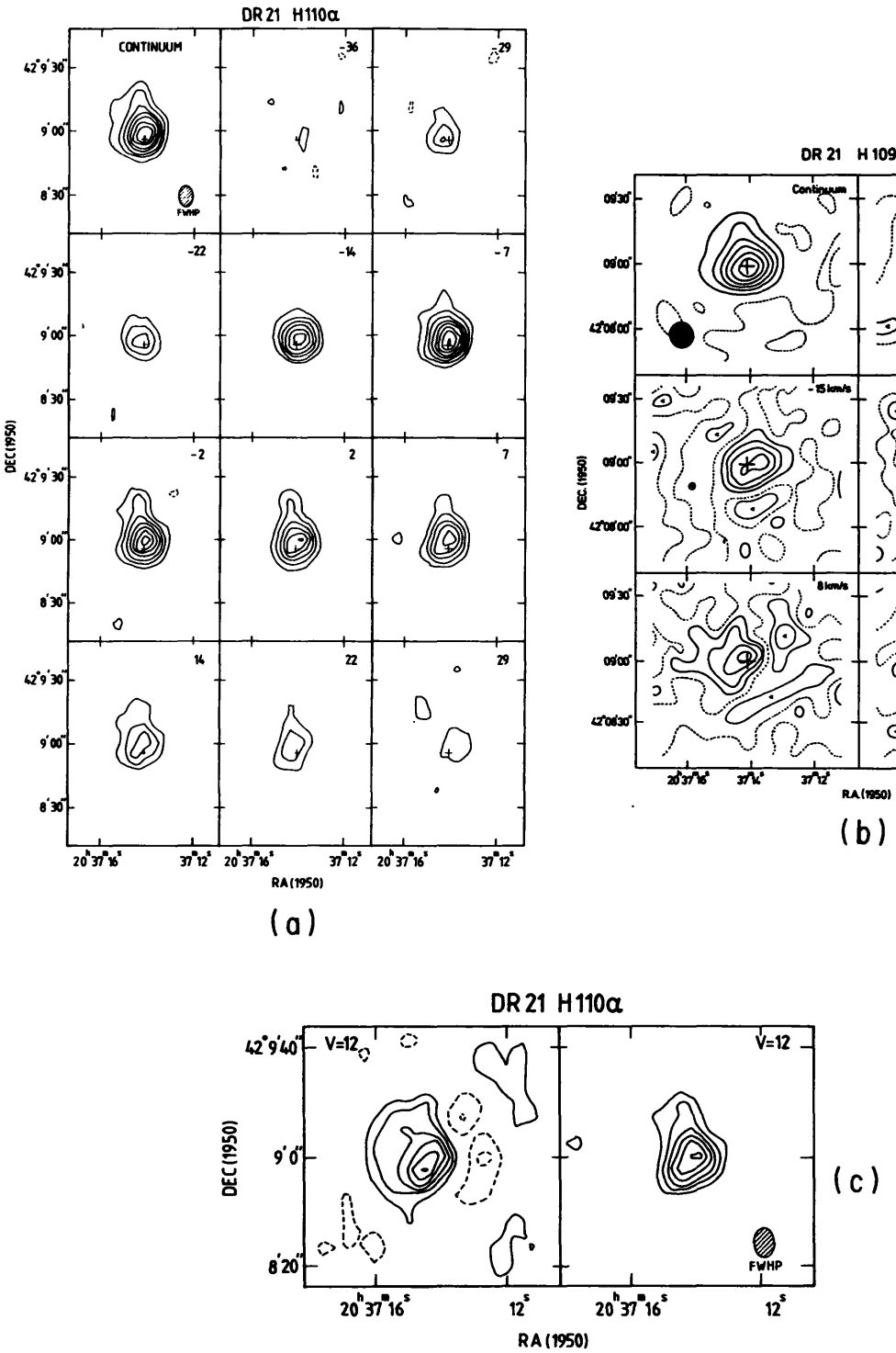


Figure 12–1. An example of the effects of very small phase errors on difference images. (a) shows images without phase error; (b) shows a similar observation, but here a phase error occurred which increased with distance from the band center, and changed sign at the center. As a consequence, the line 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 2 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.

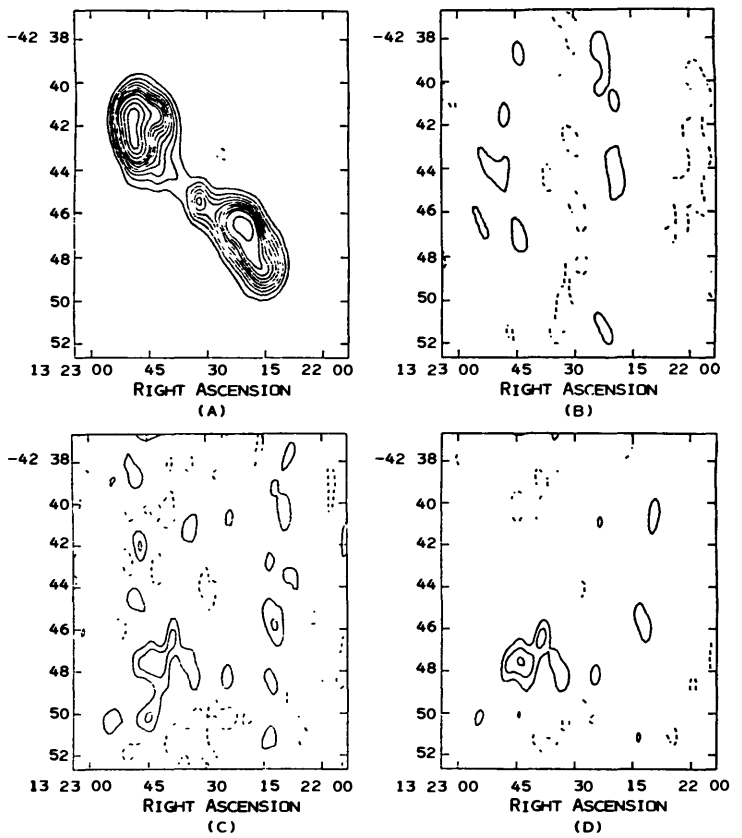


Figure 12–2. An example of the effects of chromatic aberrations on a line-minus-continuum image. (a) shows the continuum source, Cen A. The contours range from 1 to 18 Jy/beam. (b) shows a line-minus-continuum image at 260 km/s. The dirty mean continuum image has been subtracted from the dirty channel 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 HI 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 the reason that the residuals in (b) are so high that the line remains undetected.

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 in circumstances in which 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 visibility 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 flux density of the continuum source is very large. An extreme example is shown in Figure 12–2, showing weak hydrogen emission from the strong (200 Jy) radio galaxy Centaurus A. In this example, the mean continuum image was 'CLEAN'ed, the 'CLEAN' components were subtracted in the u-v plane from all frequency channels to correctly remove most of the strong continuum. Finally, the mean remaining continuum from the residual images was formed and subtracted from all frequency channels. The noise in the resulting line-minus-continuum images is a factor of ten lower than in an image obtained after simply subtracting the continuum image from a line channel. The example illustrates how a line can remain undetected if an improper subtraction of the continuum is done.

6. SELF-CALIBRATION

When applying the self-calibration relations to correct for the atmospheric phase errors in spectral line observations, the same set of phase corrections can be used at all frequencies. The optimum procedure 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). For cases in which all the channels have comparable 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 a number of 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 result in a substantial increase in the errors when these channels are subtracted to form the line-minus-continuum image. An example of this with the 'CLEAN' algorithm is shown in Figure 12–3. It shows the HI 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. The resulting image after subtraction can contain a mixture of positive and negative features, so algorithms which rely on positivity to suppress sidelobes (e.g., MEM) cannot be used.

Often we want to determine the ratio of the deconvolved line-minus-continuum to a deconvolved continuum image, in order to compute 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. In this situation a three-dimensional deconvolution algorithm would 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 (gridpoints) 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 HI observation) the hydrogen column density is given by

$$N_{\rm HI}(\alpha, \delta) \propto \Delta\nu \sum_{i=1}^{n_{\rm tot}} I_i(\alpha, \delta)\,,$$

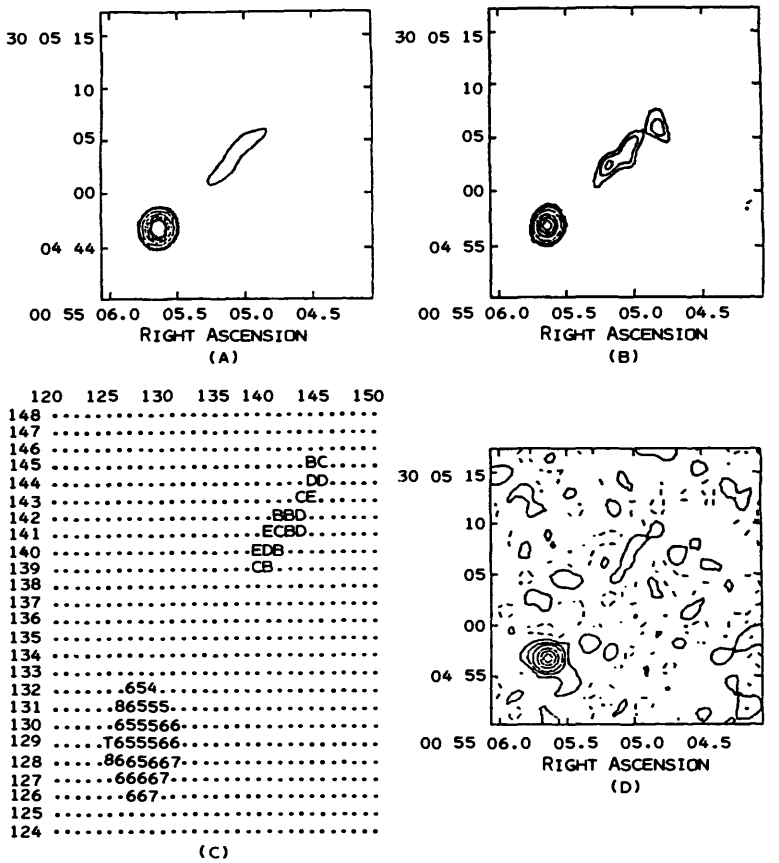


Figure 12–3. An example which illustrates how 'CLEAN' can introduce spurious emission or absorption features. **(a)** shows the continuum source NGC 315. **(b)** shows an image centered at the peak of the HI absorption line. The peak of the core is 400 mJy/beam; in (b) it is reduced to 250 mJy/beam. 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 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.

the intensity weighted mean velocity by

$$\overline{v} = \frac{\sum_{i=1}^{n_{tot}} I_i(\alpha,\delta) v_i}{\sum_{i=1}^{n_{tot}} I_i(\alpha,\delta)},$$

and the velocity dispersion by

$$\sqrt{\langle v^2 \rangle} = \sqrt{\frac{\sum_{i=1}^{n_{tot}} I_i(\alpha,\delta)(v_i - \overline{v})^2}{\sum_{i=1}^{n_{tot}} I_i(\alpha,\delta)}}.$$

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 will appear at a different position because of the Doppler shift. One obvious way to get the total emission is to take the sum over all the channels. However, at any given point in the image the line emission is only present in a few channels—the rest containing noise—so this straightforward procedure will increase the noise level significantly over that in individual channels. As an example, the neutral hydrogen emission in four different velocity channels

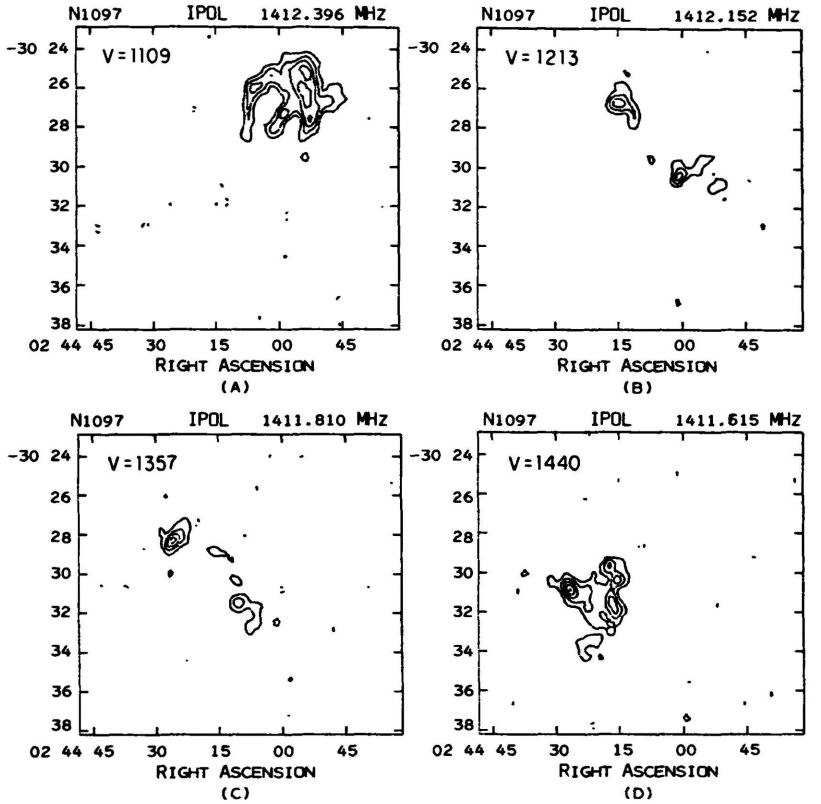


Figure 12–4. An example of four line images of the HI emission in NGC 1097. The velocity is indicated in the upper left corner. The contour interval is 6 mJy/beam.

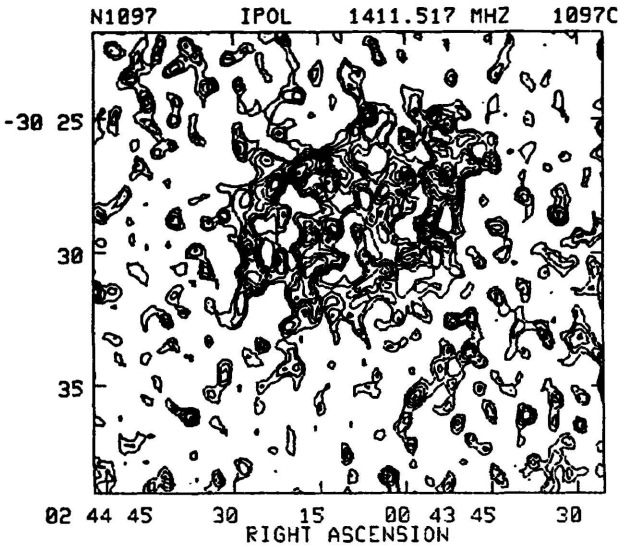


Figure 12–5. A total hydrogen image of NGC 1097. The image has been obtained by taking the sum of all channel images. The contour interval is 210 mJy km/s.

of the barred spiral NGC 1097 is shown in Figure 12–4. The total hydrogen determined by summing all the channels is shown in Figure 12–5. The increase in noise is clear. A number of methods have been developed to avoid this signal-to-noise degradation by including only the channels with line emission.

Methods that have been used to separate the line signal from the noise in the profiles are:

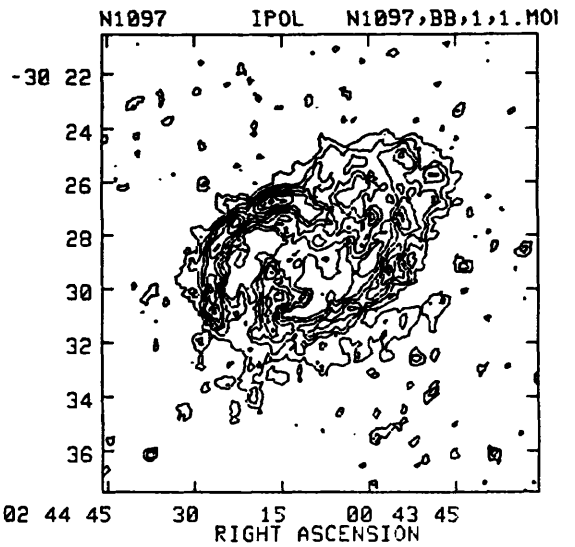


Figure 12–6. The same as Figure 12–5, except that points below a 2σ cutoff were not included in the sum.

(1) Apply an acceptance gate in intensity (CUTOFF method),
(2) Apply an acceptance gate in velocity (WINDOW method),
(3) Set the acceptance gate interactively, based on displays of 2–D sections of the 3–D image,
(4) Fit the line profile to a preconceived shape (GAUSS),
(5) First smooth the data, then use either (1) or (2).

Comparisons of methods (1) and (2) have been made by Bosma (1981) and of 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 zero and thus do not contribute in 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 12–6. Although the image is considerably improved, the method does have a 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 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. 12–7). 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, due to the strong impact of the factor $(v_i - \overline{v})^2$.

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 HI emission range is eliminated (Fig. 12–7). 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

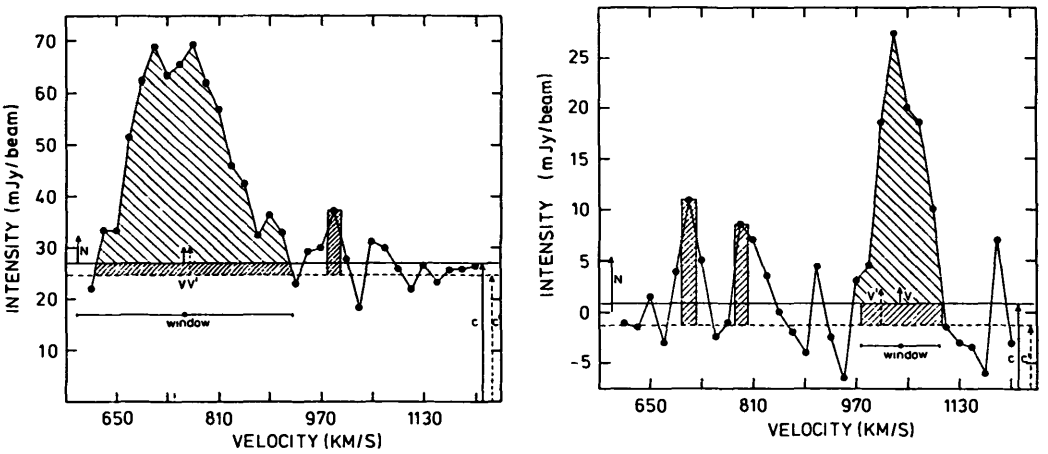


Figure 12–7. Some examples of WSRT profiles observed in the galaxy NGC 5033. N denotes the noise level (1σ), C and V are the results for continuum and velocity from the window method, while C' and V' are the results from the cutoff method. The profile integrals are shown with coarse shading. The fine shading indicates the additional contribution to the integral from the cutoff method. The cutoff level used is 1.5σ (Bosma, 1978).

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 subsequent 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).

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 is window set interactively using a suitable graphics input device. Although often superior, this does introduce the possibility of personal bias, and it is very time consuming for a larger database.

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; however, if the assumed functional form is incorrect, this can introduce a bias. Some limited testing has shown that this method and WINDOW give very 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 smoothed data—smoothed either spatially or in velocity—to determine the points to be included, but then going back to the full resolution data for the actual moment calculations. Figure 12–8 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 would like to thank Pat Palmer and Arnold Rots for comments on the manuscript.

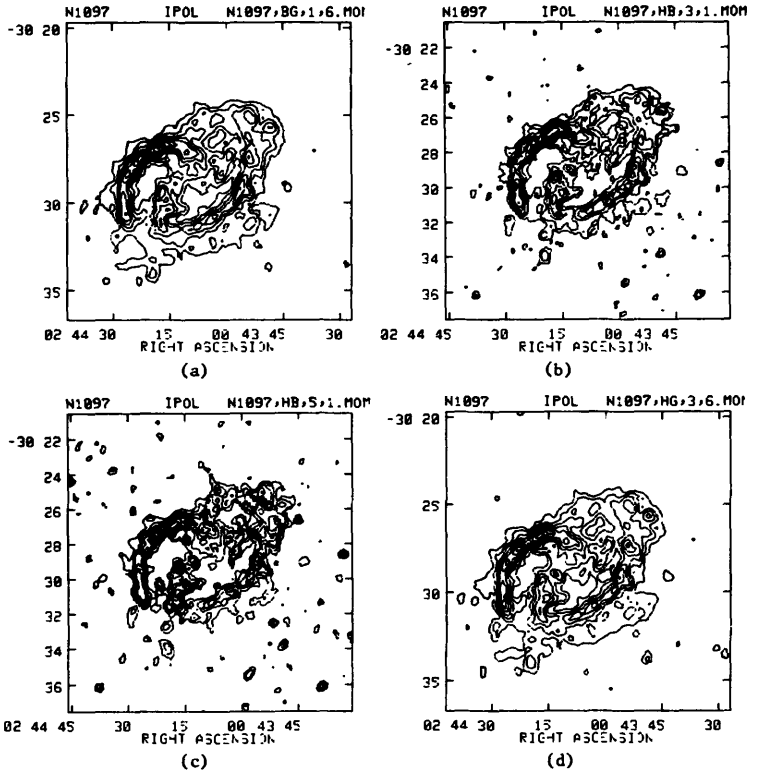


Figure 12–8. An illustration of the use of different smoothing functions to determine the acceptance window. **(a)** spatial smoothing with twice the resolution, **(b)** Hanning smoothing, **(c)** Hanning smoothing over 5 channels, **(d)** spatial + Hanning smoothing.

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 Leiden.

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, 2841 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 M 101", *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 HI gas in NGC 3938", *Astron. Astrophys.*, **105**, 351–358.

13. Very Long Baseline Interferometry

R. CRAIG WALKER

1. INTRODUCTION

Very Long Baseline Interferometry (VLBI) is the technique that allows the use of widely separated antennas as elements of an interferometer array. It is distinguished from other forms of radio interferometry by the fact that there need not be any communication between the elements while the observations are being made. Instead, independent, high-quality frequency standards are used in place of a distributed local-oscillator signal, and the baseband, digitized signals are recorded on magnetic tape for later correlation. Test experiments have been done using satellite links, but nearly all VLBI observations are done using separate frequency standards and recorded data.

Since no communication is required between the array elements, the baselines can be arbitrarily long. Currently, baselines are limited to about 11,000 km by the size of the Earth and by the practicalities of mutual scheduling of antennas. There are plans to extend the technique to even longer baselines using antennas mounted on spacecraft. The resolution typically 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 relatively common. 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 at a redshift of 0.067 ($H_0 = 100\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$). It is also about 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 that has been discussed in the previous lectures. However, some operational differences complicate the data analysis and lead to the use of calibration techniques for VLBI that differ in detail from those used for connected-element interferometers. The most important differences result 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. In addition, there are differences between traditional VLBI methods and those used for connected-element interferometry that exist either 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 any significant differences. In this lecture, I will concentrate on those areas where VLBI differs from connected-element interferometry.

2. VLBI SYSTEMS

The VLBI technique was developed in the late 1960's to study the very high brightness sources in the nuclei of active galaxies and quasars and to study masers in our own galaxy. The antennas used belonged to existing observatories and, in most cases, were not designed for use as part of an interferometer. Two recording systems were developed in 1967: the Mark 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 Mark II system was developed in the United States and is still in use worldwide. It records 4 Megabits per second (Mbs) of digital data on video tapes. There have been three generations of video tape recorders used with the Mark II system. The current version uses home-style video cassettes and records four hours per tape. There are Mark II correlators at NRAO in Charlottesville (3 stations), Caltech (5 stations), and in Bonn (3 stations). Caltech is modifying their Block II correlator (primarily designed to process Mark III tapes) to process 16 Mark II tapes simultaneously.

More recently, the Mark III system has been developed, 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 be used to record 13 minutes of data. A narrow head system is being tested that will extend that time to about 3 hours. The Very Long Baseline Array (VLBA)—currently under construction by the NRAO—will use an enhanced version of the Mark III system that can record at least 128 Mbs for 12 hours per tape. There are operational Mark III processors at Haystack (4 stations) and in Bonn (4 stations). Plans exist to upgrade both of these processors to more stations. A processor that uses the Mark III tapes has been built in Japan for geodetic work, a Haystack-design Mark III processor has been built for the geodetic community in Washington, D.C., and the Block II processor (for Mark III data) is being tested at Caltech/JPL.

The frequency standards used for VLBI are either rubidium vapor frequency standards or hydrogen masers. For the time-scales of interest (a few seconds to a few hours) the hydrogen masers are the best available standards. Unfortunately, the quality is reflected in a cost of several hundred thousand dollars per maser. On short time-scales, crystal oscillators are somewhat better and, in fact, the VLBI local-oscillator signals are generally derived from a crystal oscillator locked to a maser. It is somewhat 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. However, as long as 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 the VLA if all of the 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 part of the reason for the VLBA, and be able to anticipate some of the problems that arise in calibrating VLBI observations.

3. Data Flow

This section presents an overview of the data flow for typical VBLI observations—with special emphasis on those operations that differ from what is normally done in connected-element interferometry. Detailed discussions of some aspects of the data reduction will be presented in later sections.

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

[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.*

the same total local-oscillator frequency at all stations, but the frequencies could 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 VLA has three-level quantization). The sampling is typically at the Nyquist rate, which is the reciprocal of twice the bandwidth. The VLBA will have an oversampling capability that improves the signal-to-noise ratio for narrow band (usually spectral line) data. After digitization, the data are formatted and recorded on magnetic tape. Extra information, most importantly the time, is encoded with the data. The tapes are shipped to one of several correlators within a few days of the observations. The effects on the signal-to-noise ratio of the various aspects of the digitization of the data were discussed in Lecture 3.

3.2. Correlation.

The tapes are correlated on special digital equipment built for this purpose. Each correlator has a maximum number of stations that it can process simultaneously. Modern experiments almost always involve more stations than any existing correlator can process at once, so it is necessary to make many passes to process all baselines. The VLBA correlator is being designed to handle up to 20 stations for the most common kinds of experiments. It should be possible to process most experiments in one pass (the VLBA will have 10 antennas). The 16 station Mark II capability being installed on the Block II processor at Caltech should soon allow those large experiments for which the Mark II system has sufficient bandwidth to be processed in one pass.

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 Earth rotation Doppler shift difference between the stations; multiplies each pair of data streams for a range of delays to generate a correlation function; accumulates the results; and writes the results to some archive medium, usually a magnetic tape. For spectral line data (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. Actually, two shifted data streams are formed, using sine waves that are 90° out of phase. The results of correlation of both of these streams with the data from the other antenna of the baseline form the complex components of the correlation function. This dual fringe rotator is one of several possible ways of obtaining a complex correlation function that were discussed in Lecture 3.

3.3. Editing and fringe fitting.

The data need to be edited after correlation but before imaging. Any large blocks of bad data should be deleted before the fringe fitting (discussed below) to prevent poor results. However, prior to fringe fitting, the volume of data can be very large (typically 2 second records in current practice), so it would be tedious to do a thorough editing job. Any really discrepant points or very short integrations should be eliminated, but a modest proportion of low amplitude points will not seriously affect the fringe fitting. After fringe fitting, the data are averaged so that the data volume becomes more manageable. At that point, it is worth inspecting all of the data and eliminating any bad points. Interactive programs that use a TV or character graphics on a CRT are available for editing out bad data.

After correlation, continuum data are *fringe fitted*. This operation, which is critical to VLBI, is the process of solving for, and removing, 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[1]. Fringe fitting is discussed in detail in Section 4. It is hoped that the geometry and clocks for the VLBA will be understood well enough that fringe fitting—and even phase calibration—can be done on calibration sources, to allow imaging of very weak program sources. Such observations can be done 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 very few, if any, unresolved calibrators for VLBI baselines and essentially all very 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 available for typical VLBI observations, and partly for historical reasons. I suspect both communities could learn from each other in this area. Once images are made, there is no real difference between VLBI data and other interferometer data, except that, for VLBI data, absolute position information is totally lost in the fringe-fitting/self-calibration process while it will typically survive self-calibration for connected-element interferometer data to high enough accuracy to be useful.

3.5. Spectral line data.

The fringe fitting step cannot be performed on the program source data for VLBI spectral line observations. 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) use methods of finding the locations of features that use just the rate of change of phase (fringe rate) and not the phase itself. Both methods will be discussed in later sections. Typical VLBI spectral line sources (usually masers) have very 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 the autocorrelation spectra to be used 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 are much like those for any spectral line data. One difficult aspect of many VLBI line experiments is a result of the nature of the sources. Masers often have very compact

[1] 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 be calculated with sufficient accuracy 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.

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 of the ways around this problem are discussed in Section 8.

4. Fringe Fitting

For any interferometer, the correlated signal will appear 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. At the time the data are correlated, the expected delay is calculated—and the data streams from each element are aligned at that delay. For spectral line observations, a correlation function centered on the expected delay is formed and Fourier transformed, to obtain the cross-power spectrum. For continuum observations, 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. If the delay is not known well *a priori* (as is usually the case), the correlation function must be determined over a range of delays that is greater than the delay uncertainty, 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 since only one sideband is kept, and it is what is needed for global fringe fitting.

Significant delay errors can lead to very 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 any uncertainties in geometry) combine to make accurate, *a priori* determinations of the delay and phase very 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 difficult to predict. For these reasons, the correlator must deliver several delay (or frequency) channels and records integrated for only a few seconds. To reduce the volume of data to reasonable levels, either very careful observations must be made to determine clock and geometric parameters—consuming much of the available observing time—or the delay and fringe rate must be determined from the data. The latter is the traditional method, and the process is called *fringe fitting*. Note that—since the geometric information is contained in the delays—fringe fitting, like self-calibration, removes any absolute position information from the data. However, the delays determined in sophisticated versions of fringe fitting are just the information used for VLBI astrometry and geodesy.

4.1. Theory of fringe fitting.

Conceptually, fringe fitting is fairly 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 residual fringe frequency spectra for several delay lags are shown for a strong source in Figure 13–1a and for a weak source in Figure 13–1b.

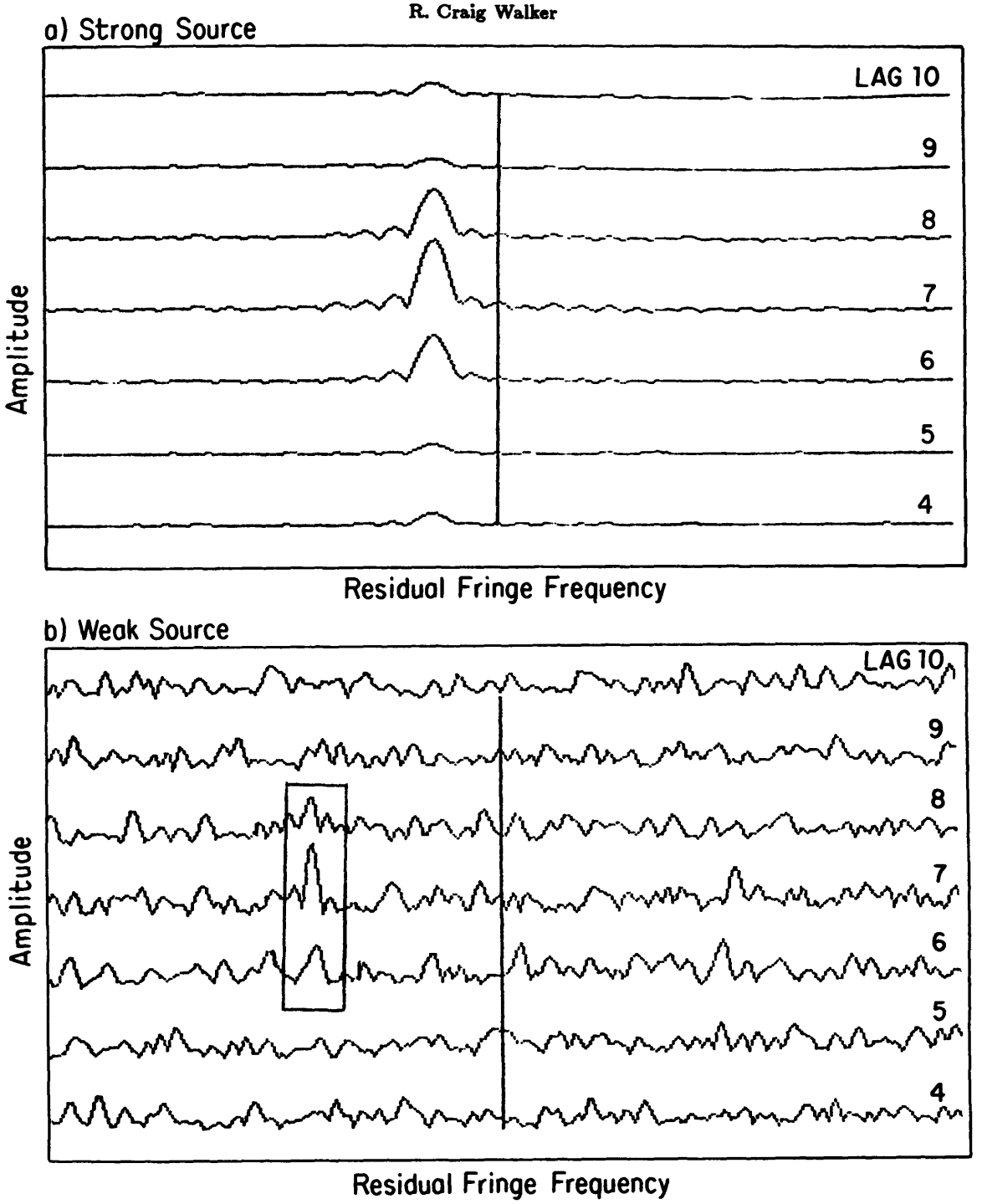


Figure 13–1. Plots of residual fringe frequency spectra for 7 delay lags for (a) a strong source and (b) a weak source. Note how the signal appears in adjacent lags at the same frequency. In the weak source case, the problem is to distinguish the signal from noise spikes.

Finding the signal in the strong source case is fairly easy. However, in the weak source case—often where the interesting science lies—it can be tricky. The methods used for fringe fitting are an area of active current development. The following describes the situation as of mid-1985.

Essentially all of the uncertainty in delay and phase, except that part due to source structure, is the result of uncertainties in parameters for each element of the interferometer.

The measured visibility data, in frequency-time coordinates, can be expressed as:

$$V_{ij}(t,\nu) = G_i(t,\nu)G_j^*(t,\nu)V'_{ij}(t,\nu) + \epsilon_{ij}\,, \tag{13-1}$$

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 gains,

$$G_i(t,\nu) = a_i(t,\nu)e^{i\phi(t,\nu)}\,, \tag{13-2}$$

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 13–1 explicitly assumes closure by asserting that the baseline gain can be expressed entirely as the product of antenna gains. In practice, a non-closing term should be included for each baseline, but such terms are less than a few percent and will be ignored for now. For high dynamic range imaging, such terms cannot be ignored, and work is in progress to find ways to calibrate or remove them. In most of the following equations, the noise term will be ignored—just remember that any measured quantity must include a contribution from noise.

Fringe fitting only concerns the phases of the visibility function. Amplitudes are calibrated by more traditional means, as will be discussed later. Separating the amplitude (A) and phase (ϕ) parts of Equation 13–1 and simplifying the notation by dropping the (t,ν) dependence gives

$$A_{ij}e^{i\theta_{ij}} = a_i a_j A'_{ij} e^{i(\phi_i - \phi_j + \theta'_{ij})}\,, \tag{13-3}$$

or for the phases alone:

$$\theta_{ij} = \phi_i - \phi_j + \theta'_{ij}\,, \tag{13-4}$$

where θ_{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 and noting that all of the ϕ_i, ϕ_j, ϕ_k's appear twice with opposite sign, and cancel,

$$\theta_{ij} + \theta_{jk} + \theta_{ki} = \theta'_{ij} + \theta'_{jk} + \theta'_{ki}\,. \tag{13-5}$$

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:

$$\begin{aligned} A_{ij}e^{i\theta_{ij}} &= a_{i0}a_{j0}A'_{ij} \\ &\times e^{i(\phi_i(t_0,\nu_0) - \phi_j(t_0,\nu_0) + \theta'_{ij}(t,\nu) + (r_i(t_0,\nu_0) - r_j(t_0,\nu_0))(t-t_0) + (\tau_i(t_0,\nu_0) - \tau_j(t_0,\nu_0))(\nu-\nu_0))}\,, \end{aligned} \tag{13-6}$$

where the r_i are the antenna fringe rates,

$$r_i = \frac{d\phi_i(t_0,\nu_0)}{dt}\,, \tag{13-7}$$

and the τ_i are the antenna delays (residual) expressed as phase slopes,

$$\tau_i = \frac{d\phi_i(t_0,\nu_0)}{d\nu}\,. \tag{13-8}$$

The object of fringe fitting is primarily to determine the τ_i and r_i, so that the delay and rate offsets can be removed from the data and the data averaged in both time and

frequency. For a single baseline, the net delay and fringe rate due both to the antennas and the source structure can be determined by fitting for phase slopes in frequency and time or, as mentioned earlier, by transforming to delay-rate space and looking for a peak.

Since all of the gain terms are antenna dependent, it is possible, as in self-calibration (Lecture 9) to use all of the data to determine the antenna gains. This allows greater sensitivity, allows the 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*[1] and is described in detail by Schwab and Cotton (1983). As in any self-calibration scheme dealing with phases, there is only 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.

There are two ways to do the global fringe fit—a least-squares solution and a Fourier transform solution. In practice, the Fourier transform solution is often used to give an initial guess for the least-squares solution. For the least-squares solution, the antenna phases, fringe rates, and delays are determined using all of the baseline data and assuming a source model. Equation 13–6 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 will depend on the signal-to-noise ratio of the data and on the quality of the initial guess model. If the model is very bad (radians of phase error), it may be difficult to obtain a decent fit and it may be desirable to restrict the number of baselines used for the fit. This consideration will be discussed further below.

For the Fourier transform solution, the frequency-time data discussed so far (e.g., Equation 13–3) are transformed to the delay-fringe rate domain. The fringes will appear at an isolated point in this domain, so one need only find the maximum and fit adjacent points for the location of the peak. 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 any other antennas. Only the one, two, and three baseline combinations are independent. Combinations of larger numbers of baselines can be expressed as sums of smaller combinations. Even the two and three baseline combinations are not totally independent because individual baselines contribute to several separate combinations. However each does contain some information not available in any other. The equations for the phase differences are, in the three cases:

$$\begin{aligned}
\phi_{ij} &= \phi_i - \phi_j = \theta_{ij} - \theta'_{ij}\,, \\
\phi_{ikj} &= \phi_i - \phi_j = (\theta_{ik} - \theta'_{ik} + \phi_k) - (\theta'_{kj} - \theta_{kj} + \phi_k) = (\theta_{ik} + \theta_{kj}) - (\theta'_{ik} + \theta'_{kj})\,, \\
\phi_{iklj} &= \phi_i - \phi_j = (\theta_{ik} - \theta'_{ik} + \phi_k) - (-\theta_{kl} + \theta'_{kl} + \phi_k - \phi_l) - (\theta'_{lj} - \theta_{lj} + \phi_l) \\
&\qquad = (\theta_{ik} + \theta_{kl} + \theta_{lj}) - (\theta'_{ik} + \theta'_{kl} + \theta'_{lj})\,,
\end{aligned} \tag{13–9}$$

where the jth antenna will usually be the reference. As can be seen, the phases of the intermediate antennas along any string of baselines enter into the sum twice, but with opposite signs—so they cancel. Each combination depends only on measured and true visibility phases.

[1] The program in NRAO's AIPS package that does global fringe fitting is called VBFIT.

A frequency-time data array can be formed with the complex, weighted sum,

$$F_{ij} = w_{ij}e^{i\phi_{ij}} + \sum_k w_{ikj}e^{i\phi_{ikj}} + \sum_{k,l} w_{iklj}e^{i\phi_{iklj}}, \qquad (13\text{–}10)$$

where the weights can reflect a variety of factors, including the signal-to-noise ratio of the baselines in the combination, preconceived notions of the value of certain stations or baselines, and the expected source visibility amplitude on the baselines. Use of a reduced set of baselines is equivalent to setting some of the weights to zero. Again, for details see Schwab and Cotton (1983). F_{ij} is now 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 13–9 and 13–10, 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 various terms in Equation 13–10 would cancel each other rather than enhance the signal-to-noise ratio. If the source structure is unknown, as it usually is in the first pass (or why is it being imaged?), it is often best to use a very restricted set of baselines to obtain the solution for each antenna. It may be best to use just the direct baseline to the reference antenna, or the two baseline combination through some other strong station. However, errors in the assumed true visibility phases that are not large enough to degrade the amplitude of the sum (less than about a radian) will not seriously degrade the solution for fringe rate and delay. Since the visibility phases due to structure are usually slowly varying in time and frequency, the main effect of errors in the model is to introduce errors 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 does not need to be especially good. All of these considerations with regard to the initial model also apply to the least-squares solution.

The result of the fringe fitting process will be a table of the delay, phase, and fringe rate offsets for each antenna relative to the reference antenna, as a function of time. That 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 averaging in delay are set by the field of view or the desired spectral resolution. In time, the limits are set by the field of view, the coherence time, 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 will be discussed below.

4.2. Practical considerations.

So much for the theory of fringe fitting—now for the art. For a number of reasons, the fringe-fitting step of the data reduction is not straightforward. There are questions about what baselines to use in a solution, how to weight the baselines (or antennas), what model to use, how long a time interval of data to use in the solution, what reference antenna to use, etc., that do not have clear answers in all circumstances. Many of these questions arise because of the wide range in sensitivity of different baselines in typical VLBI experiments, resulting from the use of a wide variety of antennas. Resolution effects also increase the range of signal-to-noise ratio's on the different baselines, often in a time dependent manner. Problems also arise in many VLBI experiments because no one antenna is on for the entire time of the observations. At least two reference antennas must 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 far from zero (e.g., more than roughly a radian), 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 the fact that some of the data will be lost. An image can be made from the results of this preliminary fit and used in a second fit, presumably with much better results. On the other hand, if the source structure is dominated by a point source, so that the true visibility phases on any baseline never get very large (several tens of degrees should not cause much problem), any combination of baselines can be used. Similarly, the model of a complicated source need not provide perfect phases in order to allow a good solution with many baseline combinations.

I suspect that there is a tendency to do second fits more often than necessary. This should be avoided, if possible, because the fringe fit places large demands on computer resources. The object of the fringe fit is to determine the delay and fringe rate with sufficient accuracy that the signal is not degraded by being averaged in delay and time. If the signal-to-noise ratio is sufficiently high, a single pass with single baseline solutions and a point source model may be all that is needed. 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 get fixed by self-calibration during the imaging process. A second fit is not needed.

How well does the delay need to be determined? The amplitude, as a function of delay, will follow a $\frac{\sin x}{x}$ function, somewhat smeared 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 $\frac{\sin x}{x}$ occurs two lags from the true delay. In this case, the amplitude will be degraded by 1 percent 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 Mark II) are a reasonable goal. Note that the tolerable delay error will depend on the distribution of baseband channels in the multiple band systems like Mark III and will have to be calculated based on the experimental setup.

The optimal way to use the delays derived in the fringe fitting for an experiment is a matter that 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. However, there is a certain amount of noise in the results, 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 due almost entirely to measurement noise, not to real fluctuations in the delay errors. Therefore it seems reasonable to try to smooth the fitting results before applying them to the data. To avoid errors, this process must be done carefully if delay steps were used in the processing, or if there are times when the delay is changing rapidly because of (for example) rapidly varying and unmodeled tropospheric variations in low elevation data.

For a recent 18 station experiment aimed at very high dynamic range, smoothing of the fringe fit results was done by hand to test the procedure. The results are encouraging, and the capability to do it automatically should probably be provided. Figure 13–2 shows the residual delays and fringe frequencies, as output by the fringe fitting program, for some of the stations of the 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 changed (note that the time ranges of the various plots are not aligned). Reference antenna changes were necessary because the source was not above the horizon at any one antenna for the entire time that a source was observed somewhere on the array. The delays were smoothed by first removing any steps introduced by clock offset changes used during correlation to keep the fringes centered in the

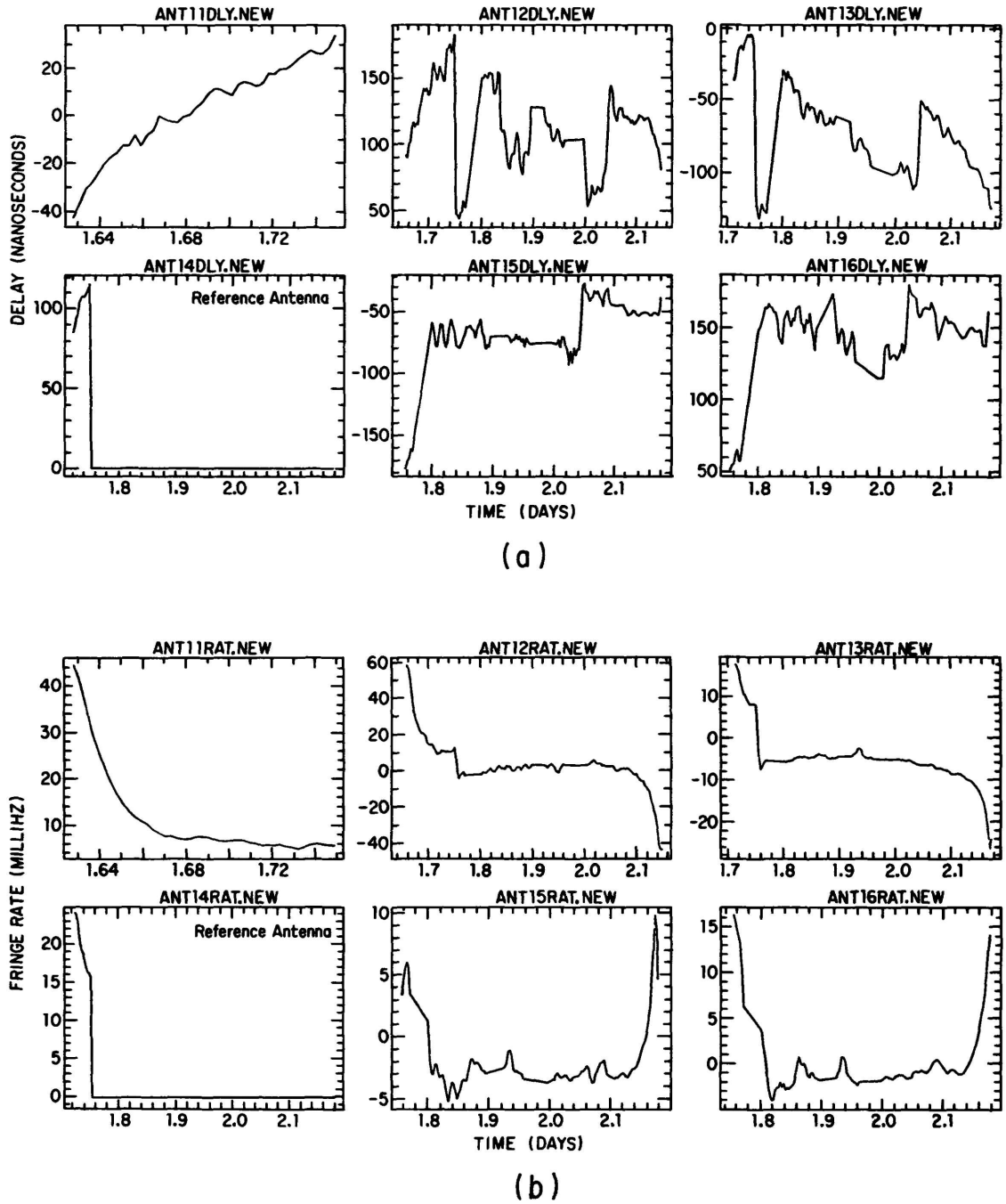


Figure 13–2. Plots of the fitted residual delays (a) and fringe frequencies (b) for 6 antennas of an 18 antenna experiment. Note that the time scales are not aligned. The reference antenna changed at about 1.75 days, causing a jump in delay and rate. Other large delay jumps were caused by steps in the 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.

delay window (some clocks drift rather rapidly), applying an offset at the reference antenna change to remove that step, generating a smooth delay for each station that consists of a few straight lines approximating the fit results, reintroducing all the steps, and using the results to correct the data. The fringe frequency fit results were not smoothed. Examination of the fit results using plots such as those in Figure 13–2 is a good method of gaining an idea of the quality of the fit and of determining any corrections that need to be made. The software for the display and manipulation of the fit results is still very much in a testing stage.

With the very wide range of sensitivities on the current networks, it is often best to just do the single baseline fits, using one of the strong antennas (e.g., VLA or Bonn) as the reference antenna. The signal-to-noise ratio is so much higher on those baselines than on most others that the others don't contribute much in a many baseline fit. However if the source is heavily resolved on the long baselines, there may be a problem. While a good reference antenna can be found for the US baselines and another good reference antenna can be found for the European baselines, there may not be a good reference antenna for the whole array. It would be good to be able somehow to use a two baseline combination rather than the direct baseline in such cases. For the 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. Unfortunately, the software is not conveniently set up for this method yet.

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. However, eventually the linearity assumption of Equation 13–6 breaks down, either because some of the offsets due to the troposphere change rapidly or because of fluctuations in the frequency standards. We shall refer to this situation, generally, as loss of coherence. The effect of coherence loss on an integrated signal is to reduce the amplitude and spread the power in fringe rate. For imaging, it is important to have good amplitudes, so the post-fringe fitting integration time should be kept well under the coherence time. However, for fringe fitting, amplitudes are not important, so integration can be extended as long as the signal-to-noise ratio for detecting a signal increases. Since the noise decreases with the square root of integration time, and the amplitude loss due to loss of coherence decreases more slowly at first, it is often worth integrating considerably longer for fringe fitting than for imaging.

The choice of an integration time will depend on the signal-to-noise ratio, on the scan lengths, and on the maximum integration set by coherence. If the signal-to-noise ratio is high, it is probably best to use relatively short integrations to avoid averaging over tropospheric fluctuations. The fit interval should be a reasonable match to the length of the observing scans. It is not wise to pick an interval that gives a very short effective fit interval at the end of a scan. The low signal-to-noise ratio in that interval will degrade the results. For current experiments, the maximum integration time allowed by coherence is likely to be established by one or two antennas that have especially poor frequency standards. In practice, the best way to determine the fit interval is to "fringe fit" a short subset of the full data set, using several different intervals, and to choose the one that works best. This is also a good way to determine what reference antenna and baseline combinations to use. Obtaining 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 13–1, along with knowledge of the antenna gains, to obtain an estimate of the true visibility amplitude from the measurements. Rewrit-

ing Equation 13–1, retaining only the amplitude terms of the gains as expressed in Equation 13–2 and ignoring the noise term, we have

$$S_{ij}(t,\nu) = \frac{A_{ij}(t,\nu)}{a_i(t,\nu)a_j(t,\nu)}, \tag{13–11}$$

where S_{ij} is the calibrated, correlated flux density and A_{ij} is the raw correlation coefficient from the correlator. With this formulation, the correction factors that account for the signal-to-noise ratio losses due to effects such as the digitization are absorbed into the a_i. It is more instructive to rewrite Equation 13–11 in a form that applies to one frequency and time and that shows more clearly how the various calibration parameters are used:

$$S_{ij} = A_{ij} b \sqrt{\frac{T_{\mathrm{sys}_i} T_{\mathrm{sys}_j}}{K_i K_j}}, \tag{13–12}$$

where b is the factor that accounts for digitization losses etc., the K_i are the antenna sensitivities in K Jy^{-1} (Lecture 6, Section 3), and the T_{sys_i} are the system temperatures in Kelvins.

Most current experiments use Equation 13–12 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 at the time of the observations, and gain curves are determined at some time. One of the complications of current VLBI is that each station provides the necessary information in a different form, 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 the system temperatures only sporadically. When there are few measurements, it is advisable to examine them carefully before using them, in case there are some bad points, as there often are.

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 known to have simple structure and enough flux density to give high signal-to-noise ratio are usually observed. Models and/or images of these sources can be used with self-calibration to determine constant offsets for each station; these then are used to improve the *a priori* calibration of the program source data.

6. Continuum Imaging

After fringe fitting and amplitude calibration, the data set is essentially the same as a data set from a connected-element interferometer that has not been phase calibrated. The procedure for making an image begins with deciding on a source model to use in the first pass of self-calibration (Lecture 9). In many cases, a point source will work, and some observers never use anything else. However, there may be circumstances in which a better model will help, especially when there is a very limited amount of data (this is an area of current debate). It also may be that some self-calibration software works better with a poor starting model than others[1]. If a more complicated model is needed, it can be obtained by fitting a number of Gaussians to some or all of the visibility data. Even model fitting requires an initial guess. Typically this is derived by examining plots of the visibility data

[1]NRAO's AIPS routines seem to work with a point source starting model.

and by guessing what structure in the sky would produce the observed minima and maxima. Deep minima in the visibility plots help this process greatly by indicating the position angle and separation of features. This whole procedure is rather difficult if the source structure is at all complicated.

Note that model fitting has uses other than to provide a first guess for self-calibration. When the source structure is simple and not confused, it is easier to get good error estimates by fitting source parameters in the visibility domain than by fitting them in the image plane. This could be especially important when trying to obtain a deconvolved size of a barely resolved feature. Such features look much like the beam in the image plane, but may have obviously reduced visibility amplitudes in the u-v data. However, any structure in addition to the features of interest affects the u-v data—so, in the presence of such structure, it is best to determine feature parameters by fitting to the final image.

Now for the hard part—getting the self-calibration started and on the right track. No recipe is known that works reliably for all cases. Because there is no preliminary phase calibration, and because the data sets are often small, providing relatively few constraints, this step is much harder for VLBI than for the VLA. The full imaging procedure involves several (often many!) passes through the self-calibration and imaging sequence. For each pass, the image from the last pass is used as a better starting model for the self-calibration. In effect, one is following an iterative procedure that solves for the antenna gains and the source structure simultaneously.

Because the imaging procedure is iterative, there is no need to try to get everything right on the first pass—in fact such an attempt would almost certainly be counterproductive. For the first several iterations, no effort should be made to correct the amplitudes in self-calibration. This is because the *a priori* amplitude calibration is probably better than what self-calibration would provide, while the model is bad. One should wait until the initially uncalibrated phases are reasonably good, before releasing the amplitudes. Also, as discussed in Lecture 11, it is usually worth restricting the range of u-v data that will be used in the self-calibration. An obvious case where this is worthwhile occurs when a large, complex source is dominated by a point component. The initial fit should only use the longest baselines, which are sensitive primarily to the point component. As the image improves, the shorter baselines can be used. Also as the image improves, the self-calibration can be extended to the amplitudes—first, just to constant scaling factors for each antenna, and eventually to point by point variations. The full range of u-v spacings might never be used, especially if the shortest baselines have excess correlated flux density from poorly modeled, large-scale structure.

A point source starting model biases the self-calibrated image towards symmetrical sources. Some common artifacts that can result from this are false counterjets in core-jet sources and triple structure in double sources. There is some indication that such bias can 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 very important to be alert to possible false features or other problems with the images. Experience with imaging from fake data sets for which the final result is known (not necessarily by the person doing the imaging) is very useful in obtaining a feel for how the procedures work. Certainly, such tests should be done when newly developed algorithms are being tested.

There are two primary clues to the quality 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.), there are likely to still be problems. 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 features that don't fit one's preconceptions of what should be there, a serious effort should be made to make the features go away. An example would be a counterjet in a core-jet source, or a feature off to the side of the jet. If the feature persists no matter what is done, that lends confidence to its reality. The noise level in an image is another indicator of the quality of the calibration, especially in high dynamic range images. As long as successive iterations reduce the noise level, the process should be continued. However, the noise level isn't a reliable indicator of quality—I've seen what seemed to be low noise images that contained spurious features and did not reproduce the u-v data well. Appearance is the traditional indicator used to judge the quality of VLA images but has only recently become a major indicator for VLBI.

The ability of the model to reproduce the u-v data has been the traditional quality indicator for VLBI. A quick search through the VLBI literature will reveal endless plots of models overlayed on data. With small quantities of data, this is important because it is possible to obtain images that seem to be of reasonable quality but whose associated models do not reproduce the data well. Such images can differ significantly from the final results that do reproduce the data. This is possible because, at low dynamic range, significant signal can be generated by spurious structures in the residual image. With large data sets and high dynamic range, the data should be reproduced better. In fact, when reaching dynamic ranges above 100, the minor changes in the predicted u-v data that are made as the image improves would be hard to detect in the normal u-v plots. However, the fit of the model to the data can still indicate problems. In a recent case where I had a spurious feature (counterjet or misplaced core) in an image with a dynamic range of 100 based on 11 station data, the best indication of trouble was not the appearance of the image—instead, it was small but significant differences between the model and data on a few baselines. The spurious feature and these differences disappeared together.

A factor that can be adjusted during image making with the 'CLEAN' algorithm is the size and placement of 'CLEAN' window(s) (see Lecture 7, Section 3.1). If there is some *a priori* reason, such as previous images of the same source, to believe that a source is confined to a certain region of sky, the imaging procedure can be guided by confining the 'CLEAN' windows to that region. With a large, high-quality data set, this is less important—but it can help with poor data sets. If the 'CLEAN' window is too small, it will be very 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, convergence was obtained quickly.

There are several antenna dependent factors that can be used to control self-calibration programs. One is the weight given to each antenna. This allows one to emphasize certain antennas in the fit based on, for example, the reliability of the *a priori* calibration (mostly useful for amplitudes). Another is a restriction that can be placed on the range over which the amplitude calibration of each antenna is allowed to vary. This allows one to force some antennas to remain nearly at their *a priori* values while others vary. With the wide range in the quality of the *a priori* calibration typical in VLBI, this can be very useful. Finally there is a smoothing time-scale for the amplitudes from each antenna. This allows one to let some antennas vary on a point to point basis while others only change very slowly. Again it allows one to utilize information on the quality of the calibration of each antenna.

As must be clear by now, there are lots of parameters that can be adjusted while attempting to obtain a good image. There is no set procedure that is guaranteed to work. The above discussion should be enough to get started, but each case is likely to require a slightly different method. Experience helps, but even with it, many false starts and very many iterations are often made before a good image is obtained. On the other hand, many images converge quickly with minimal fiddling of parameters. This whole area is clearly in need of some more advanced algorithms that make the imaging easier and more reliable.

An example of the imaging process is shown in Figure 13–3.

Figure 13–3a is the result of the first pass of self-calibration with a point source starting model. The image is obviously rather bad, but at least it is not a point, a definite position angle has been established and some asymmetry has been introduced. Figure 13–3b shows the results after 11 iterations. The source is now much more compact and asymmetric. The noise level is lower (note the lowest contour level). There are still hints of problems, however. There are some uncomfortably deep negative features to the east and there are suspicious, low level features off the line of most of the features. Also, comparison with images made at other times casts suspicion on the eastern-most feature that appears as a bulge on the side of the strongest feature. But overall, the image is very good on the scale of many VLBI images that are made so there is a temptation to stop and declare this to be the result of the experiment. As it turns out, that would be a big mistake. The eastern-most feature, that in this source would be treated as the core and would be the feature used to align this image with images made at other epochs, is not real.

The best indication that there are still problems with the image in Figure 13–3b is not seen in the image plane at all, but rather in comparison of the predicted and calibrated data.

Figure 13–4 shows the calibrated phases (crosses) and predictions of the 'CLEAN' components of the image (smooth lines). On one baseline (HSTK-OVRO—Haystack to Owens Valley), the prediction misses the data by a small but significant amount. A couple of other baselines, not displayed here, showed somewhat smaller problems. These offsets may not appear to be very significant but they are—the prediction should be much better and was for the final image. In this case, the problem could not be fixed by further efforts at self-calibration. Closer examination of the fringe fitting results revealed that the clock at one station (not HSTK or OVRO!) was very poor and that too long and integration time had been used. Also there were some problems with data at times when the raw fringe rate or delay rate were near zero. After redoing the fringe fit and editing some more data, the image of Figure 13–3c was obtained. This image is much better. It shows low level features that are far from the main emission region and that have been confirmed by observations at other frequencies. Also, it contributes to a consistent story about the evolution and motions of components in the source.

The data displayed in Figure 13–4 demonstrate the problem of the wide variation in signal-to-noise ratio for VLBI data. Clearly the HSTK-VLA data is very much better than the NRL-GRAS (Maryland Point to Fort Davis) data at this frequency. In fact, when the scatter in phase is as large as it is on the NRL-GRAS baseline, it may be best to edit out the baseline entirely.

The example of Figures 13–3 and 13–4 was something of an extreme case in the trouble encountered during imaging. However it was not because of limited data—the Mark II observations were of a 3 Jy source and involved 11 stations. It should have been easy. Clearly the procedures are not yet routine.

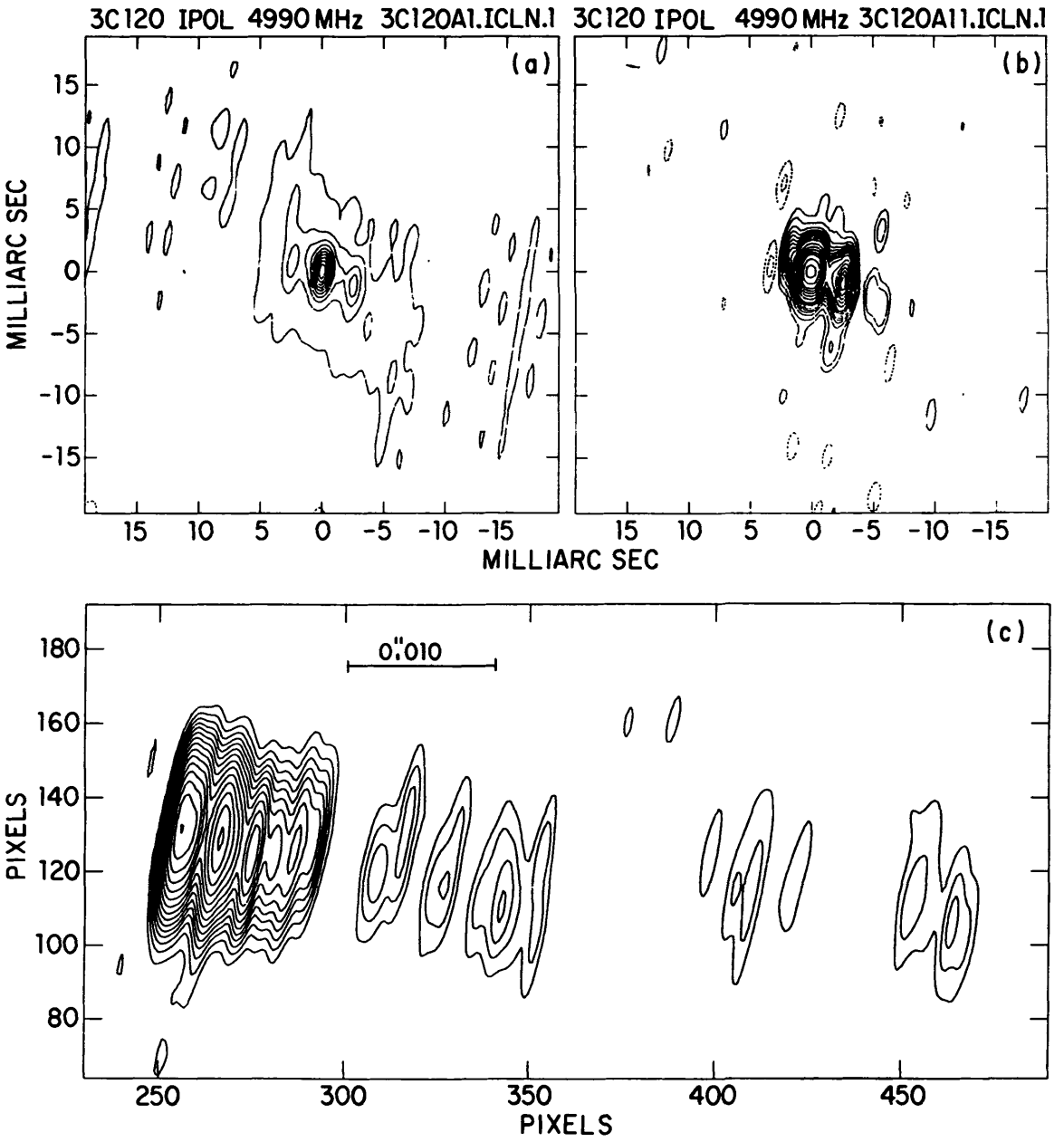


Figure 13–3. VLBI images of a compact continuum source (3C 120) at 3 stages of the imaging process. **(a)** The first iteration. One self-calibration iteration, phases only, has been done using a point source model. The contours are at −35, 35, 69, 139, 208, 278, 347, 417, 486, 556, and 625 mJy/beam. **(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.

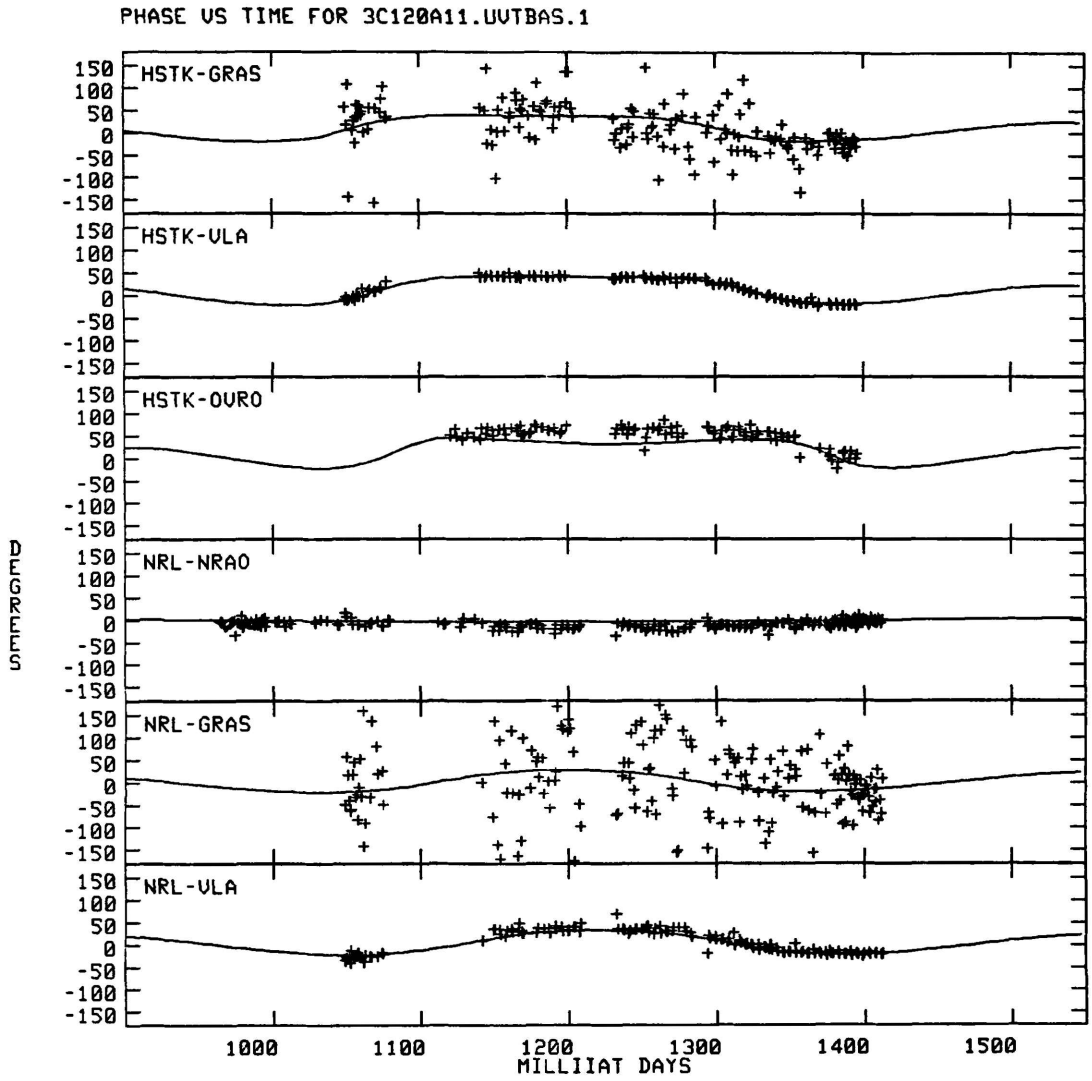


Figure 13–4. A display of data phases and phases predicted by the 'CLEAN' components of the image of Figure 13–1b for 6 baselines. The crosses are the data, the smooth curves are the model. Note the systematic offset on the HSTK-OVRO baseline. Also note the wide range of signal-to-noise ratio on the various baselines.

7. Spectral Line Calibration

Spectral line data are traditionally handled rather differently from continuum data. It is not possible to fit for residual delays using spectral line data, because phase slopes across the band are likely to reflect frequency dependent source structure—not just delay errors. It might be possible to use an iterative procedure involving imaging two or more well separated channels to solve for a delay term, but such a method has not yet been developed. However, since the spectral information is to be retained because it contains the science, there is no need to fringe fit to allow averaging in frequency (delay). The delay does have to be calibrated eventually in order to make synthesis images without dealing with phase offsets. This is usually done by observing continuum sources relatively often and fringe fitting them. There is a commonly used mapping method—fringe rate mapping—that doesn't care about constant phase offsets, so the delay errors don't necessarily need to be calibrated[1].

[1]The word "map" is used when discussing this method. The method, as commonly used, gives the spatial

Recall from Lecture 3 that one of the factors that degrades VLBI data is the so-called *fractional bit error*. It is the result of the fact that the delay cannot be set finer than one bit, so that there is a constantly changing phase slope of up to $\pm 90°$ across the band. For high delay rates that give high bit update rates, this just causes a small signal-to-noise ratio loss for continuum data. However it causes a frequency dependent loss for spectral line data that can be very large at the edge of the band. The best way to deal with this is to only integrate for a time short compared to the bit update rate, and then transform and apply a phase slope, to remove the known delay error due to the inability to set exactly the desired delay. The NRAO VLBI correlator in Charlottesville can do this, except on the very longest baselines, when the bit updates happen faster than the minimum integration time of 0.2 seconds. For those cases, one just has to live with the signal-to-noise ratio degradations, although the effects are known and amplitude corrections can be made.

For a typical spectral line experiment, the exact velocity range covered by each spectrum will depend on the Doppler shift produced by the Earth's rotation and orbital motion and on the local oscillator settings. These shifts can be significant in that they can correspond to a significant fraction of the width of a spectral line. Therefore, before imaging and before any calibration steps that depend on the source spectra, they must be removed. Each spectrum can be shifted to a common velocity by transforming to delay space, applying a phase slope, and transforming back to frequency space[1].

Fringe rate residuals need to be calibrated and removed, regardless of the imaging method. This could be done by solving for the fringe rate residual of one channel as if it were a continuum source. However, for many spectral line experiments, there is very strong signal in at least one channel. The traditional way of removing fringe rate residuals is to rotate the phases of a reference channel to zero, and rotate all other channels by the same amount. This is known as phase referencing. If the reference channel has structure, it is possible to image that structure using the continuum imaging techniques and then rotate the channel phases to match the image. Phase referencing does more than just reduce the fringe rates; it actually calibrates the phases of all channels, except for any phase slope due to a delay offset. If a reference channel image has been used for final phase referencing, and if continuum sources have been used to calibrate the delay offsets, then the phases of all other channels are fully calibrated and can be used directly to make images without self-calibration. These images will all have the correct position relative to the reference channel image. Self-calibration could be used to improve the dynamic range of the images, much as it is used to improve images made with well calibrated data from connected-element interferometers, but it is not essential.

The overall amplitude calibration (i.e., a constant factor for each cross-power spectrum) can be done in the same way as for continuum data. This is the most effective method for weak sources. However, for strong sources, such as most masers, there is a much better method that uses the autocorrelation spectra for each antenna as a function of time. The method takes advantage of the fact that the autocorrelation data are subject to the same effects that affect the cross correlation data. This includes pointing, atmosphere, receiver fluctuations, etc.

To use the autocorrelation spectra for calibration, they must be calibrated themselves, in the sense that the bandpasses must be removed. A raw autocorrelation spectrum will include the noise power, so it will look like the passband of the final filter, with the source signal added on top. The noise must be removed, usually by observing off-source near

---

locations of spectral features, but gives no information on their shapes. Positions can be plotted on a "map" without other information, but it is hard to call something with no shape information an "image".

[1] The VLBA correlator will be able to do this on-line.

the time of the on-source observation and then subtracting the off-source spectrum from the on-source spectra. Since many maser sources can contribute significantly to the total system noise, the on- and off-source spectra should be scaled by the system temperatures prior to subtraction. If the system temperature information is not available, the spectra can be scaled so that frequency channels that do not contain signal average to zero.

For Mark II data, and for good absolute calibration of any data, it is necessary to divide all cross- and autocorrelation spectra by the normalized bandpasses. For cross-correlation spectra, the geometric mean of the two bandpasses involved should be used. This removes any channel to channel variations in sensitivity. It is especially important for Mark II because the bandpasses are poorly matched (yes, this causes closure errors for continuum data). One doesn't want such fluctuations to show up as spurious structures in the images.

Once the autocorrelation spectra have had the bandpass effects removed, some of the best should be averaged and calibrated in an absolute sense (i.e., in Jy rather than correlation coefficients). This well calibrated spectrum now serves as a template. For all other autocorrelation spectra, the scale factor needed to match each spectrum to the template is derived using a least-squares fit. That scale factor is just the antenna gain. The geometric mean of the factors from the two antennas at each end of a baseline is then used to calibrate the cross-power spectra. Of course, there will be a constant term in addition to account for effects of the fringe rotator that don't affect the autocorrelations.

One final calibration would be useful, although it is not traditionally done for VLBI. That is to calibrate the phase passbands. This is important if the instrumental phases are not flat across the passband. In fact they are not, but the deviations from flat are only a few degrees and, for the low dynamic ranges achieved so far, are not important. They will become important eventually. The best way to do this calibration is probably to determine the phase passband using observations of a strong continuum source.

Note that there is more information available for spectral line calibration than for continuum calibration, so, although there are more steps, it can be easier and more powerful. For this reason, synthesis images were made from spectral line data at 1.35 cm long before the advent of amplitude self-calibration opened the possibility of continuum imaging at that frequency. At 1.35 cm, most of the telescopes have low efficiencies and poor pointing. It is not uncommon to see factor-of-two amplitude fluctuations in a matter of minutes, especially at scan boundaries, when the pointing is checked. The autocorrelation spectra allowed calibration of such problems. Until methods were developed to determine amplitudes from the data, continuum results at high frequencies were very poor. The spectral line calibration has the great advantage that, after one channel is imaged without the benefit of calibrated phases, all other channels have calibrated phases, and imaging is straightforward. Often there is at least one channel that contains a point source or simple double, so even imaging the first channel may be easy.

8. Spectral Line Imaging

In the last section, it was concluded that imaging of spectral line VLBI data is easy. What was meant is that one does not have the complex art of trying to obtain a good image with phases that obey the closure relations but are otherwise uncalibrated. What was not mentioned is that the sources often contrive to make the imaging very difficult, by being spread over regions of sky often four orders of magnitude larger than the resolution. This is especially true of the water masers at 1.35 cm wavelength. 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. Needless to say, an image with 3 points per beam covering twice the field containing the source—the

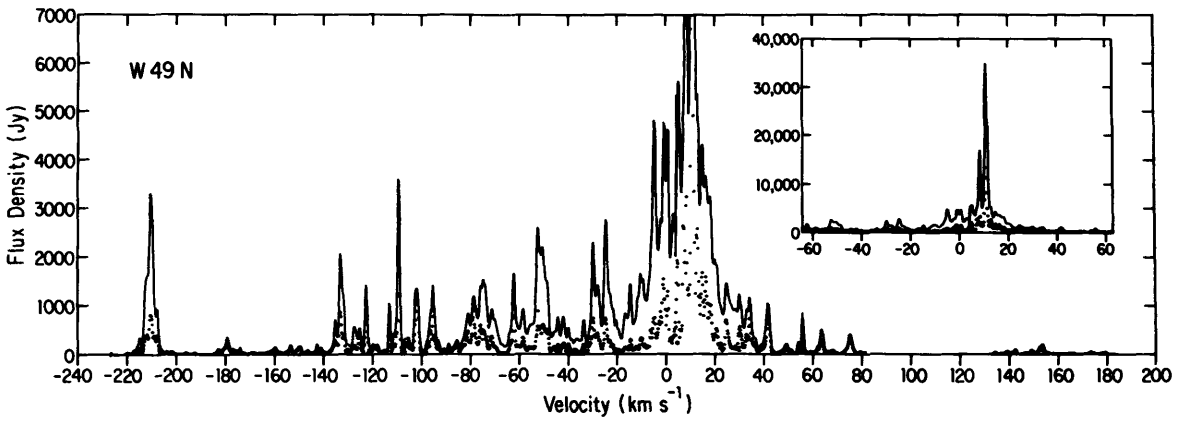


Figure 13–5. A spectrum of a complex water maser source (W49 — Walker, Matsakis, and Garcia-Barreto 1982). There are about 1500 spectral channels in the spectrum. The solid line shows the total power received by a single antenna while the dots show the portion of the total power that was successfully 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.

typical imaging parameters—is out of the question for any computer. It would take over 7 Gbytes just to store the image of one channel in the 2 arcsecond case! To compound the problem, the masers often have very complex spectra that require images to be made of a very large number of channels.

An example of such a complex maser source is shown in Figures 13–5 and 13–6.

Figure 13–5 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 13–6 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 is less than 1 milli-arcsecond. This display only attempts to show 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 very long. Typical water maser cases are limited to one or two seconds integration. Orion (30 arcseconds) is limited to 0.2 seconds, and less might be preferred. These short integration times, combined with the large number of spectral channels needed, lead to very large data sets. One reason that more spectral line experiments haven't been done, and that some of those that have are taking so long to reduce, is just that the data volume is overwhelming.

Clearly, clever methods must be found to make the imaging problem reasonable. To a large extent, the sources consist of a small number of point sources at each frequency. Calculating and storing 7 Gbytes, just to determine the parameters of a handful of separate features, is rather inefficient. The method that has been used most is called *fringe rate mapping*. It not only doesn't use a grid, but also it only uses 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

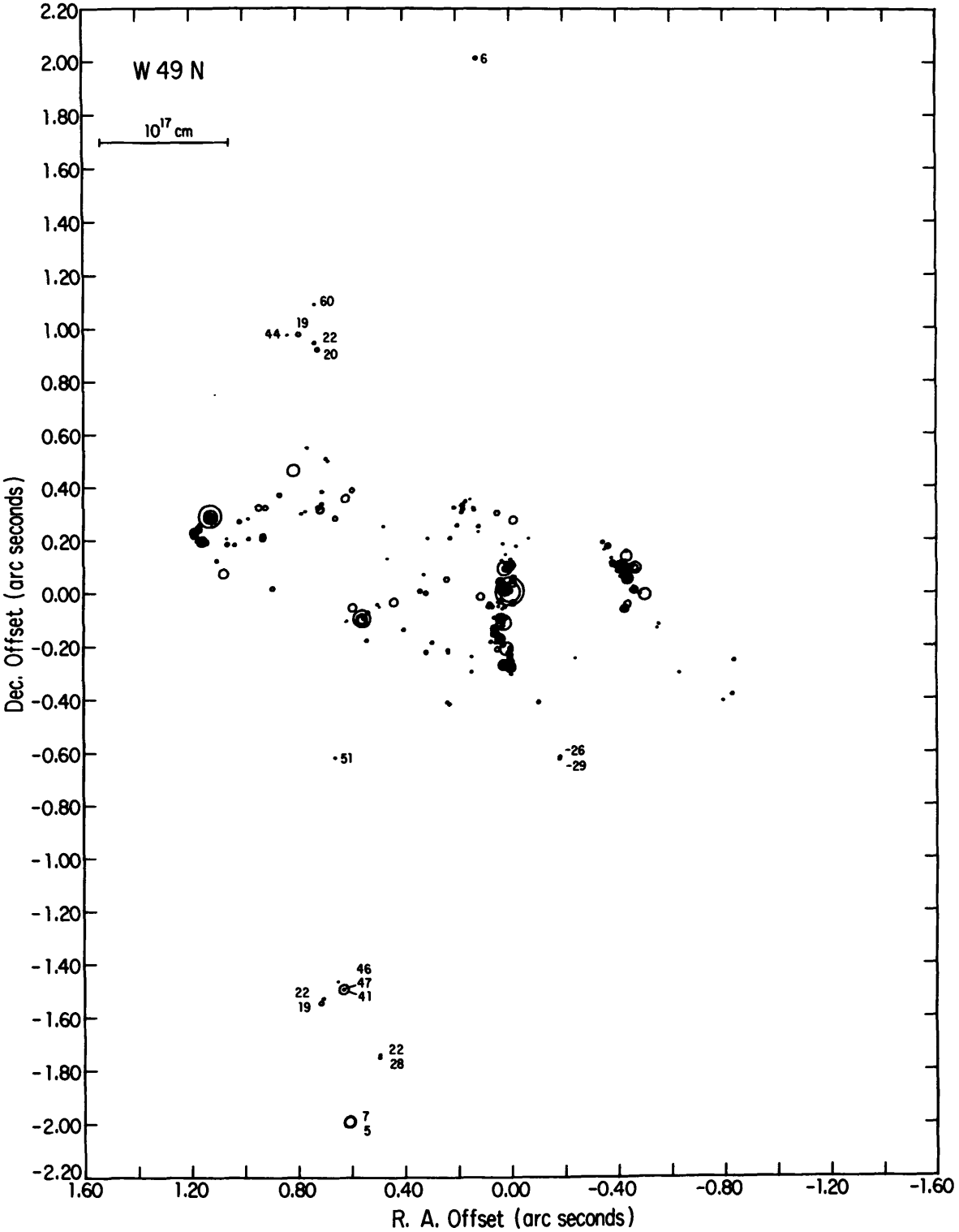


Figure 13–6. The fringe rate map of the water maser source whose spectrum is shown in Figure 13–5. The main concentration of features is spread over 2 arcseconds with a few features over a larger region. The size of the symbols represents the flux density of the features. The resolution is better than a milli-arcsecond so it is difficult to display the full spatial dynamic range in one image. The large ratio of resolution to source area 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.

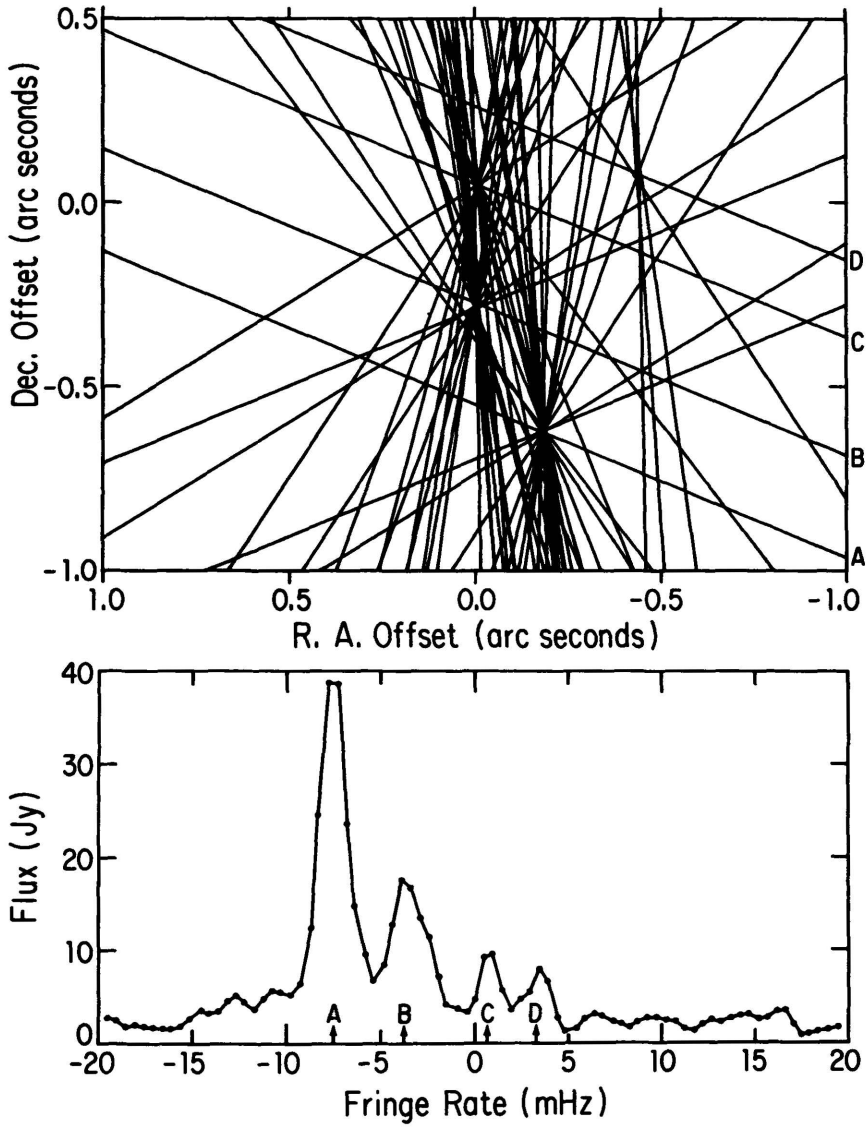


Figure 13–7. The lower plot is a fringe-rate spectrum of one velocity channel for the source in Figure 13–5. There are four peaks, each corresponding to a separate feature on the sky. Each peak confines its corresponding feature to lie along a line on the sky. The upper plot shows such lines from many 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 so 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).

critical. The method is based not only on the fact that there is a phase offset between any two features 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. The choice of the interval of data over which to calculate each fringe rate spectrum is based on a tradeoff. Long integrations give higher signal-to-noise ratio and finer

fringe rate resolution. However, relative fringe rates change with time, 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 such fringe rate peak constrains the feature to which it corresponds to lie along a line on the sky. The lines for all peaks from one fringe rate spectrum are parallel. 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. 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 3 station experiment are shown in Figure 13–7. 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 is complicated by the fact that sometimes features overlap in fringe rate space, sometimes some features aren't seen in all fringe rate spectra—for noise or dynamic range reasons, and 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 checking by eye. Probably more sophisticated mathematical techniques could be used if someone would take the time to code them.

The accuracy of positions found with fringe rate mapping is a few times worse than what can be achieved with synthesis imaging. The sensitivity of relative fringe rate to a position offset is much lower than the sensitivity of relative phase. The accuracy will depend both on the sensitivity to position offsets and on the accuracy with which the location of each fringe rate peak can be determined.

The 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[1]. 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 either require 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.

Once the maps are made, one is faced with all the usual problems of how to display spectral line data. But that is the subject of another lecture.

9. HAZARDS

In this Section I will note again some hazards that I have already mentioned, and discuss a few new ones.

The digital fringe rotation and delay setting causes losses of signal-to-noise ratio, as discussed in Lecture 3 and as mentioned above. However they can also cause much worse problems, under certain circumstances, if proper corrections are not made. Those circumstances involve short baselines and certain times when either the raw fringe rate or delay update rate are small. These are not the residual rates discussed under fringe fitting, but the total rates removed by the processor. If the fringes go through less than a few turns, or

[1] The AIPS program MX is well suited to this.

the delay change is less than a small number of bits during an integration time, the degradations due to the effects will be different from the statistical effect at large rates. The magnitude of the offsets tends to follow something like a $\frac{\sin x}{x}$ law, so it takes several turns or updates for them to become insignificant. This is easy to imagine by considering the zero fringe rate case or the constant delay case. For these cases, there is no degradation, so when such data are combined with data from longer baselines, there can be large closure errors. The effects are calculable, so data in one of the dangerous regimes should be corrected or deleted. For the VLBA, all data will be corrected. Baselines shorter than about 500 km may show problems for significant amounts of time in 18 cm Mark II observations. The effects scale such with frequency and bandwidth. Higher frequency observations have higher fringe rates, so slow fringe rate problems only occur 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 bandwidth that counts is the one that determines the sampling rate. For multi-band data such as Mark III, the effect depends on the individual channel bandwidth.

Finally, I emphasize again the need to for effort to understand the reliability of an image. Remember that self-calibration likes symmetric structure, so be suspicious of counterjets and of other symmetric features. Also remember that the appearance of an image alone may not be a reliable quality indicator, especially with small amounts of data. It is worth checking both the appearance and the fit of the significant parts of the image (e.g., 'CLEAN' components) to the data. If anything looks suspicious, try varying the imaging parameters or even starting over with a different initial model or different u-v range, or something. See how well features repeat.

REFERENCES

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

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

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.

Schwab, F. R. and Cotton, W. D. (1983), "Global fringe search techniques for VLBI", *Astron. J.*, **88**, 688–694.

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 in W49. I. Maps", *Ap. J.*, **255**, 128–142.

14. Image Analysis

EDWARD B. FOMALONT

By image analysis I shall mean the general procedures and techniques which are used to interpret and parametrize useful information from an image or a set of images. An essential part of this analysis is determining the reliability and sensible error estimates to associate with the intensity distribution and derived quantities. This analysis concept is, clearly, somewhat vague and dependent on the nature of the observations and the type of questions which motivated the observations. Nevertheless, some general analysis techniques are useful to discuss. I will emphasize the philosophy of most of the techniques and not go into implementation details except when necessary. VLA software will be mentioned in connection with specific algorithms.

I shall assume that the set of images has been appropriately processed. For aperture synthesis, this processing includes data editing and calibration (Lecture 4), as well as deconvolution of the point spread function (Lecture 7) 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, covered in Lecture 15, is an important aid in image analysis. For simple images, grey-scale TV-oriented displays and contour diagrams are suitable visual aids 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.

Throughout this Lecture, wherever an image analysis function is described I will also mention the name of the program implementing the function in the NRAO Astronomical Image Processing System (AIPS).

1. Image Modification

Several types of image modification are useful in analysis and display. Two that are described in this section are image convolution to change the apparent resolution, and image interpolation to change the grid network on which the intensities are defined. Other general types of image correction are also mentioned.

1.1. Smoothing an image.

The calculated intensity distribution represented by an image is generally a smoothed version of the true intensity distribution. One is at liberty to modify the resolution of the image in order to better discern very small or very large features. Figure 14–1 shows a contour display of a radio image which contains large-scale and fine-scale features. It is obvious that different image parameters are better suited for measurement at different resolutions. The overall appearance of the complex features is seen in image (a), the integrated intensity

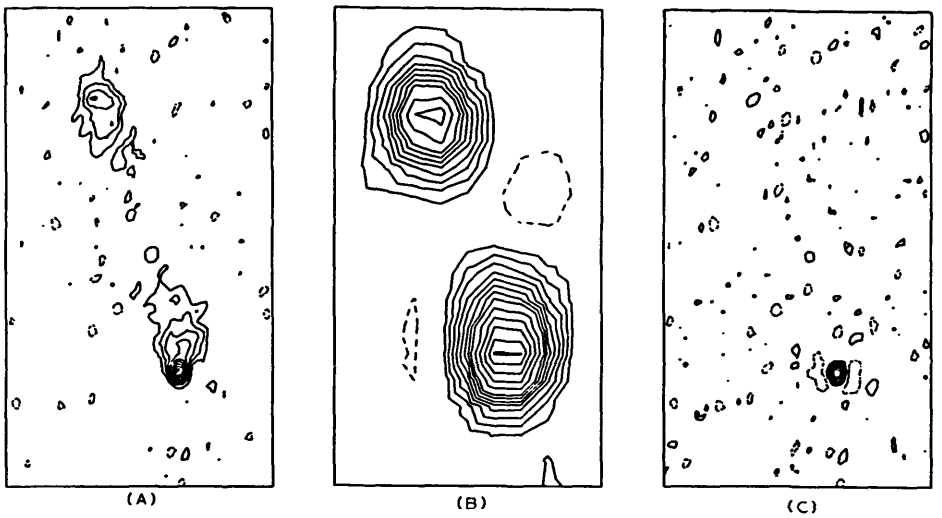


Figure 14–1. Relative emphases of image features achieved through use of differing resolutions. (a) $0''.5 \times 0''.4$ resolution; (b) $2''.0 \times 2''.0$ resolution; (c) $0''.5 \times 0''.4$ resolution, with high-pass filtering.

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 which 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 $I'(l)$. For simplicity I shall assume that the intensity is defined on a regularly spaced grid, l_i, and in one dimension the convolution is

$$I'(l_j) = \sum_i I(l_i) K(l_j - l_i) \,. \tag{14–1}$$

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

Kernel (1) produces a slight smoothing called Hanning; kernel (2), a heavier smoothing; and kernel (3) will sharpen some of the features. Each kernel is symmetric. Extensions to n dimensions are obvious (AIPS tasks CONVL, SMOTH).

A related method of convolution uses Equation 14–1 transformed into spatial frequency space, u. If $V'(u)$, $V(u)$, and $k(u)$ are the (inverse) Fourier transforms of $I'(l)$, $I(l)$, and $K(l)$, respectively, then the convolution formula becomes

$$V'(u_j) = V(u_j) k(u_j) \,. \tag{14–2}$$

The expression $k(u)$ can be interpreted as a weighting factor in spatial frequency space. This method is easily applicable to those instruments where the (inverse) Fourier transform of the image is directly measured (use of UVTAPER in AIPS tasks UVMAP, MX). And, even with the extra overhead of two Fourier transforms, many convolutions are more quickly

calculated in this manner, using Equation 14–2. For example, high-pass filtering can be accomplished by setting $k(u_j) = 0$ for $u_j < U$, $k(u_j) = 1$ for $u_j > U$. The associated kernel, $K(l_i)$, is an oscillating function similar to kernel (3), except that many more terms must be kept in the convolution in order to approximate an accurate high-pass filter.

The third convolution type is associated with the deconvolution/reconvolution methods that are used in aperture synthesis to remove that distortion in an image which is produced by non-uniform Fourier sampling (Lecture 7). These methods decompose an image into a set of point components and then reconvolve this set of components as desired. In principle, this set of point components can be smoothed to any desired resolution (AIPS tasks APCLN, MX, VM).

Unfortunately, the deconvolution needed for this third method may not work in a uniform manner across the image and can produce an image which is of variable resolution. This will happen if the deconvolution does not recover all the flux in the image field of view. For example, in the 'CLEAN' algorithm, if the subtractions are terminated while large-scale emission is still present in the dirty image, the restored components will have a resolution specified by the 'CLEAN' beam, while the unsubtracted emission, generally of large angular scale, will have a resolution specified by the dirty beam. The later smoothing will give different results with these two different beams. Thus, this simple and inexpensive convolution method ought to be used only if the image contains strong, isolated features with little extended emission, and has been fully deconvolved. Otherwise, the use of the first two convolution methods is recommended.

1.2. Interpolating an image.

The image intensity distribution is generally defined over a rectangular lattice which is often 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 determining the positions of isolated features, registering a set of images of the same field, and mosaicing a set of small images into one large image.

If $I(l_i)$ represents an intensity distribution defined over a grid, then the interpolated intensity $I'(l')$, where l' is at an arbitrary point, is also given by Equation 14–1 with l_j replaced by l'. If the intensity distribution is band-limited—i.e., contains no frequencies higher than U—then a perfect interpolation kernel is $K(z) = \frac{\sin 2\pi U z}{2\pi U z}$. The Fourier transform of this kernel, $k(u)$, has the properties $k(u) = 1$ for $u < U$, $k(u) = 0$ for $u > U$; so that $V(u) = V'(u)$ and $I(l) = I'(l)$. If $U\Delta l \simeq \frac{1}{2}$, then this interpolation is expensive to calculate. If the image is well-sampled—that is $U\Delta l \ll \frac{1}{2}$—then the adjacent points are not independent, and much simpler kernels will produce an accurate interpolation (AIPS task LGEOM). 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.

1.3. Primary beam correction.

The antennas which comprise an array are sensitive to radiation coming from a small region of sky. As discussed in Lecture 1, correction for the relative sky sensitivity across the image (the primary beam correction) is generally made after the best-quality image has been obtained. If the image contains only a few bright features, the correction need only be applied locally. Other imaging techniques have similar sensitivity variations over the image area.

1.4. Other image defects.

Many second-order image corrections were discussed in Lecture 8: corrections for bandwidth smearing, integration time smearing, non-coplanar aperture sampling, and grid curvature. It is generally very expensive to correctly modify an image for these defects. However, minor corrections to the parameters of discrete features can be made 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—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 can be easily compared at different frequencies and different epochs. The obvious properties which are useful for discrete components are the integrated and/or peak intensity, the mean position, and the size. In two-dimensional image analysis there are six relevant parameters (one intensity, two position coordinates, and three size descriptors). The knowledge of the image resolution is also necessary in order to interpret the intensity and the size of the component.

The simplest set of parameters defining the component properties consists of the moments of the distribution (AIPS tasks MOMNT, MAXFIT). In one dimension the (first three) moments are

$$\begin{aligned} F &= \sum I(l_i)\,, && \text{Integrated Intensity,} \\ l &= \tfrac{1}{F}\sum l_i I(l_i)\,, && \text{Mean Position,} \\ \sigma &= \tfrac{1}{F}\sum (l_i - l)^2 I(l_i)\,, && \text{Width,} \end{aligned} \qquad (14\text{–}3)$$

and they have been normalized in the usual manner. Extension to two dimensions is straightforward. Higher moments can be defined, but they are of little use for most astronomical applications.

2.1. Model fitting.

It is often more convenient to solve for these parameters in the framework of a specific model intensity distribution. For components which are not too resolved, the point spread function is generally chosen. Somewhat extended features can be decomposed into several model components, suitably displaced. Finally, images of extended objects with known shapes, like planets or stars, can be compared, for instance, to a uniformly-illuminated circular disk with several appropriate free parameters. After selection of the appropriate model intensity distribution, $M(l_i)$, the free parameters of the model 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, V,

$$V = \sum (M(l_i) + Z - I(l_i))^2\,. \qquad (14\text{–}4)$$

The zero offset, Z, which has been added in here is often useful to include as a free parameter.

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 need an initial guess of the model parameters to converge quickly and at the deepest minimum in V. An additional parameter, the zero-offset in the fitted region, should be included, since many image defects can produce a local bias near a discrete component.

INPUT MAP

	84		86		88		90	
823	0	0	0	0	0	0	0	0
822	0	0	0	0	0	0	0	0
821	0	1	0	0	0	0	0	0
820	0	0	0	0	0	0	1	0
819	0	-1	0	1	1	0	0	0
818	0	1	4	8	5	0	0	0
817	0	3	20	38	21	3	0	0
816	0	4	38	78	46	8	1	0
815	0	4	35	76	48	9	1	0
814	0	2	15	35	24	5	0	0
813	0	0	3	7	4	1	0	0
812	0	0	0	1	0	-1	1	0
811	0	0	0	1	1	0	0	0
810	0	0	0	0	1	0	0	-1
809	0	0	0	0	0	0	0	-1
808	0	0	0	0	-1	-1	0	0
807	0	0	1	0	0	0	0	0

RESIDUAL MAP

	84		86		88		90	
823	0	0	0	0	0	0	0	0
822	0	0	0	0	0	0	0	0
821	0	1	0	0	0	0	0	0
820	0	0	0	0	0	0	1	0
819	0	-1	-1	0	0	0	0	0
818	0	0	-1	-1	0	-1	0	0
817	0	0	0	0	1	0	0	0
816	0	-1	0	0	0	0	1	0
815	0	-1	0	0	0	0	0	0
814	0	0	0	0	1	0	0	0
813	0	0	0	-1	-1	-1	0	0
812	0	0	0	1	-1	-1	0	0
811	0	0	0	1	1	0	0	0
810	0	0	0	0	1	0	0	-1
809	0	0	0	0	0	0	0	-1
808	0	0	0	0	-1	-1	0	0
807	0	0	1	0	0	0	0	0

```
Residual of fit          =  0.39
Peak Comp. Intensity     = 85.0±0.2
Integrated Intensity     = 97.7±0.7
Position =   87.10±0.01, 815.53±0.01
Comp. Size = 2.73±0.01 x 2.10±0.01 in  7.3±0.6
Resolution = 2.50      x 2.00       in  0.0
Intr. Size = 1.16±0.02 x 0.55±0.03 in 23.8±1.7
```

Figure 14–2. Image fit to a bright component.

An example of a fit to a strong feature is shown in Figure 14–2 (AIPS tasks IMFIT, JMFIT). The feature is only slightly resolved. The residual image 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 5 times the r.m.s. fluctuations, most model fits will add bias into the parameter estimates. The detailed corrections depend on the character of the noise and the type of fitting algorithm.

Correction for image defects, many of which were discussed in Lecture 8, can be applied directly to parameters from the model fits. These effects include the primary beam distortion, bandwidth smearing, integration time smearing, and distortions of the image grid. For example, the component which was fit in Figure 14–2 is displaced 70″ in position angle 30° from the phase center of the observations. With a 25 MHz bandwidth the calculated bandwidth smearing is 1.2 units in the direction of the displacement. Thus, the effective resolution of the image at the location of the component is 2.5×2.0 units in position angle 0° from the point spread function, plus a radial smearing of 1.2 units in 30°. The size of a point source would then be about 2.7×2.1 in position angle 10°. The component thus is nearly unresolved, with a size < 0.3. The convolution of the radial smearing and the Gaussian point spread function produces a final shape which is not precisely Gaussian. More exact methods of analysis are possible.

The fitting of extended components which require several model components for an adequate description can lead to ambiguous results. The parameters (up to 12 for two Gaussians) can be strongly coupled, especially the peak intensity and the size parameters, so that particular values and error estimates may be in error even if the residual image is satisfactory. Less ambiguity will occur if some of the free parameters are held constant.

2.2. Errors of the parameters.

Error estimates which are obtained directly from most fits should be viewed with skepticism. There is generally a built in assumption that the image errors are stochastic and independent, which may not be valid for a variety of reasons. Let R be the post-fit r.m.s. error associated with the pixels in the image. Then the minimum errors of the parameters

for most models are approximately,

$$\begin{aligned}
&\Delta Z = R/3\,, && \text{Zero Bias,}\\
&\Delta P = R\,, && \text{Peak Intensity,}\\
&\Delta F = \sqrt{R^2 + (\Delta\Sigma/\sigma)^2}\,, && \text{Integrated Intensity,}\\
&\Delta l = R\sigma/2P\,, && \text{Position,}\\
&\Delta \sigma = R\sigma/P\,, && \text{Width.}
\end{aligned} \qquad (14\text{–}5)$$

The above expressions are only rough guides, and the true errors may be larger. When fitting in two dimensions the same expressions apply.

2.3. 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 for poor image quality are:

(1) the paucity of input data,
(2) the inaccuracy of the measured visibility phase, and
(3) the distortions of very large features in the image.

Examples of this type of fit are the determination of the positions of strong isolated components using only the visibility phases, fitting the visibility amplitude data for 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. With a reasonable amount of data and good phase stability, parameters can be obtained from the image as accurately as from fits to the visibility data.

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. The fitting of the brighter sub-components does make sense, and the discussion of the preceding section is relevant. Intelligent image display (Lecture 15) at this stage is needed to determine which aspects of the image or set of images to analyze. Of course, there are many morphological properties of some images which cannot be parametrized, and a suitable display is all that is needed in these cases.

Comprehension of the properties of a complicated feature is usually enhanced if the dimensionality of the analysis of the feature can be decreased. For example, one-dimensional analyses of filamentary features are useful, and various distributions along lines parallel and perpendicular to the axes of these features can lend considerable insight. A radio image of a source with a jet is shown in Figure 14–3 (left). A one-dimensional analysis has been made along lines perpendicular to the jet axis at increasing distances from the core (Killeen, Bicknell, 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

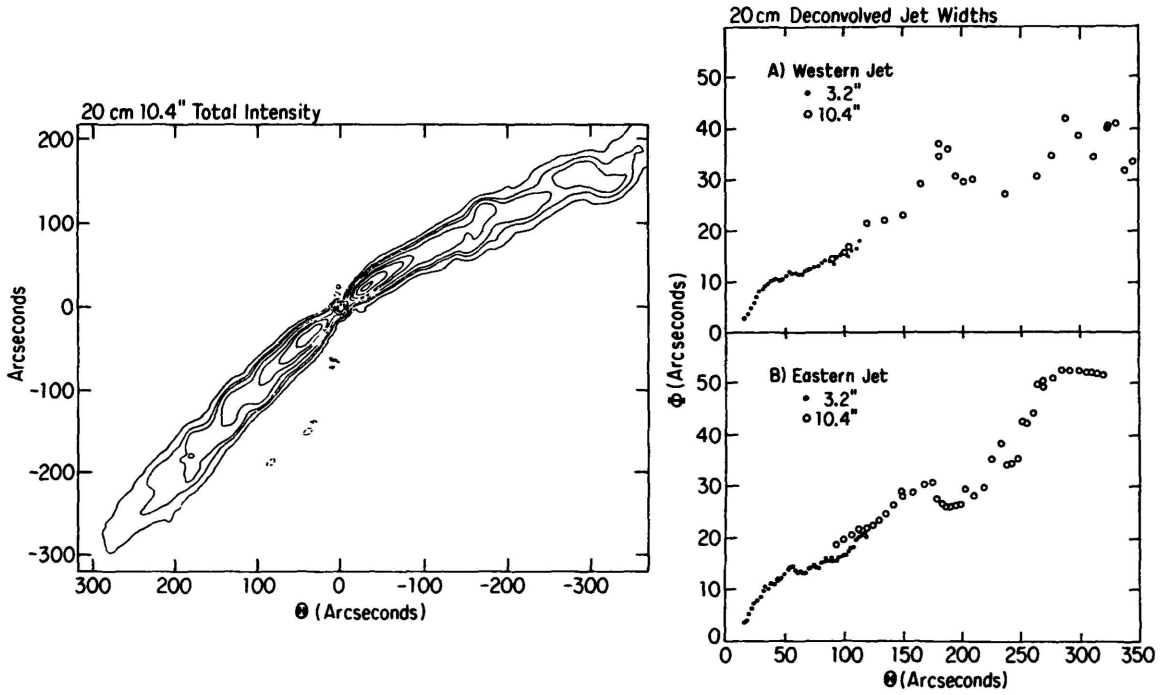


Figure 14–3. An example illustrating the analysis of an image of a radio jet. *(Left)* A contour display of the total intensity. *(Right)* Model fits of the jet width.

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 in 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 of an extended feature, along with its error, are useful parameters. Because an integrated intensity estimate usually is derived from data covering a large area, the estimation of integrated intensity is very 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 ensures that the correct boundaries will be chosen (see Lecture 12 for a fuller discussion). Also, some algorithms will not respond to low-level emission in the presence of much noise. The use of several alternative methods for determining the integrated intensity of an extended feature is illustrated below, using the feature in Figure 14–4.

1. Sum up the intensities within the feature as a measure of the integrated intensity (AIPS task IMEAN). A simple summation of the pixel values is usually accurate enough, although a Simpson's rule integration should be chosen for images with near-critical sampling. Choose several control regions which surround the feature, and use the same integration technique. The integrated intensities in the control regions can be fit to a constant offset, or to a higher-order polynomial 'baseline' around the feature, and an error estimate can be ascertained from this fit. The analysis shown in Figure 14–3, where the control regions were used to solve for a zero offset and error, gave the result 7.09 ± 0.14. A similar analysis is often used to remove the sky background from an optical image.

2. Try model fitting the feature with several components (AIPS tasks IMFIT, JMFIT). Do not pay too much attention to the individual parameter values. If the post-fit residuals

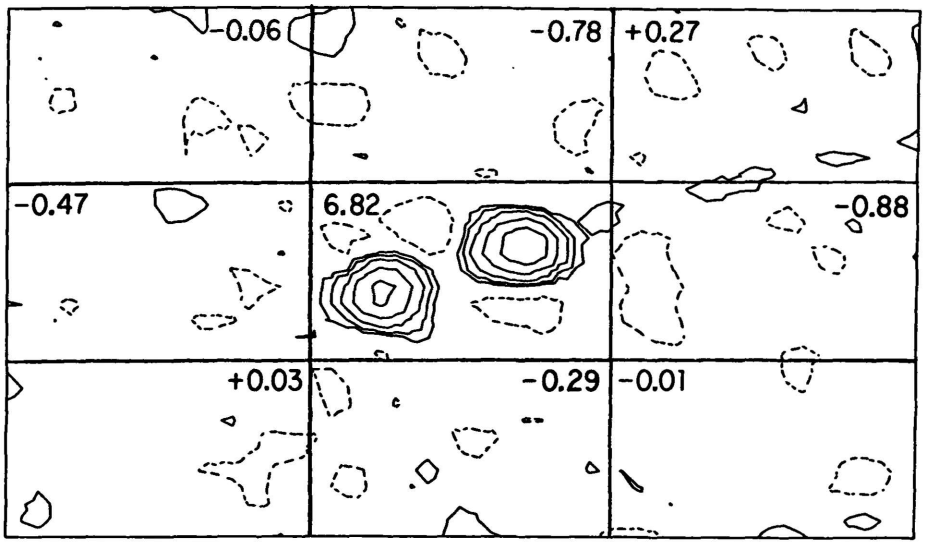


Figure 14–4. The integrated intensity of an extended feature. The peak intensity within the feature is shown at center, as is the peak intensity within each of the eight control regions. This is the same feature that is shown in Figure 14–1.

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. If the 'CLEAN'ing deconvolution technique has been used, sum up the 'CLEAN' components within the boundary of the feature (AIPS tasks APCLN, MX). Deep 'CLEAN'ing may be necessary. No error estimate is given in this method. The value 7.1 was obtained.

Method 1 is preferred, although methods 2 and 3 are satisfactory for features which are somewhat less extended than the one in Figure 14–4. The final estimate depends on the quality of the various methods and their agreement. For this feature a flux density of 7.1 ± 0.2 was used.

3.3. Very extended features.

An estimate of the integrated intensity of a very large feature is affected by small biases in the image. Simple sums over the feature can lead to poor estimates, and very low-resolution images often have extremely poor image quality. The integrated intensity 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 relevant part of the image (AIPS task FFT) or by direct measurement in aperture synthesis techniques. Extrapolate the lowest frequency Fourier components to zero frequency (AIPS tasks 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

There are many ways in which a set of images can be obtained. A region of sky is often observed at a number of frequencies and in several polarization states simultaneously. Repeated observations of the region are also made to improve the image reliability, to cover

a wide range in frequency, or to obtain differing resolutions. Finally, observations can be made with several different telescopes at totally different frequencies. There is the choice of analyzing each image separately and then comparing the results in some manner; or appropriately combining the images pixel-by-pixel to obtain image distributions of secondary quantities—which may be more easily and meaningfully analyzed and interpreted.

If each of the images is defined over the same regularly-spaced coordinate grid, then the set of images can be combined naturally into one three-dimensional image, which is commonly called an image cube (AIPS task MCUBE). Spectral line observations produce a homogeneous set of images over frequency. Observations at equal time intervals also naturally produce an image cube. These image cubes can be transposed to rearrange the ordering of the data in the computer. Many display and analysis techniques can be achieved much more efficiently with proper storage of the data.

4.1. Image compatibility.

Image combination makes little sense if the input resolutions are not identical. Otherwise, strange effects will occur 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. The set of images must also be interpolated onto the same spatial coordinate grid. Occasionally, the scale in a set of images will change linearly with frequency because of an assumption or simplification made during the calculation of the image.

Images with different resolution, but which are otherwise comparable, can be combined linearly. However, the corresponding sum of the point spread functions must also be calculated in order to interpret these images. For example, many short observations at the VLA can be summed at various stages of reductions. Depending on the instrument, there may be a loss of signal-to-noise ratio in such a combination.

The proper alignment and orientation, called registration, of a set of images usually is necessary before they can be combined and compared. For images which have been obtained simultaneously from a single observation, it is likely that the images can be aligned precisely, unless the set of images covers a large range of frequency or a large field of view. For example, dispersive refraction will change the relative position as a function of frequency for the images.

For observations made at different epochs, even using the same instrument in an identical configuration, offsets between the images may occur because of inaccuracy in the determination of the absolute positioning. For images from 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 4) and by careful monitoring of the instrument while the observations are recorded.

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 exceedingly difficult to determine (see Lecture 13).

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, and
(4) Mosaicing of adjacent images.

The forms of these distortions are generally known and can be corrected by using Equation 14–1 in order to interpolate all of the images onto the same grid.

4.2. Image errors.

The distribution and the properties of errors in an image are important for the determination of the reliability of the image and the parameters derived from it. In Section 2, where discrete features were parametrized, errors were assigned on the basis of the r.m.s. variations in the vicinity of the feature; little regard was made to their nature except to note whether the errors produced a bias in the image. However, more complicated analysis of images and sets of images requires more intricate techniques.

The error distribution in an image usually consists of two components, due to: (1) fundamental limits in the telescope and instrumentation which produce errors that are usually stochastic and have reasonably well-defined properties, and (2) systematic effects caused by a variety of 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 detectors used to intercept the radiation. For correlation-type detectors, as used by most synthesis arrays, the noise is distributed with a normal probability about zero intensity (a small offset is possible) with an r.m.s. dependent on many observing and receiver parameters. Systems with total-power detectors produce a noise distribution which follows a Rayleigh function. The correlation scale of the images is equal to the resolution. An example is shown in Figure 14–5. 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 instrument resolution. Finally, for observations where the number of detected events per pixel is small, the image will have Poisson statistics. The magnitude of the error distribution across the image may change. However, for arrays the error is relatively constant across the image (before correction for the instrumental sky sensitivity). The error distribution should be about the same as that expected from the observing parameters (see Lecture 6).

Any discussion of systematic errors in images 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 11 should be consulted. A different set of problems occurs for optical and X-Ray imagery.

4.3. Linear image combinations.

A linear combination, I_c, of a set of J images, I_j, has the form

$$I_c(l) = \sum a_j I_j(l) + b_j \,. \qquad (14\text{–}6)$$

Examples are: (1) the sum or difference of images at different frequencies; (2) the sum or difference of various polarization states; and (3) sums of images made at different epochs (AIPS tasks SUMIM, COMB). The noise distribution in the combined image is the appropriate combination of the noise on each image. Stochastic errors tend to add in quadrature; systematic errors may be highly correlated among the set. For example, the difference between two images at slightly different frequencies, observed simultaneously, may be of much better quality than either of the input images, because errors associated with the point spread function or tropospheric phase fluctuations are almost identical. On the other hand,

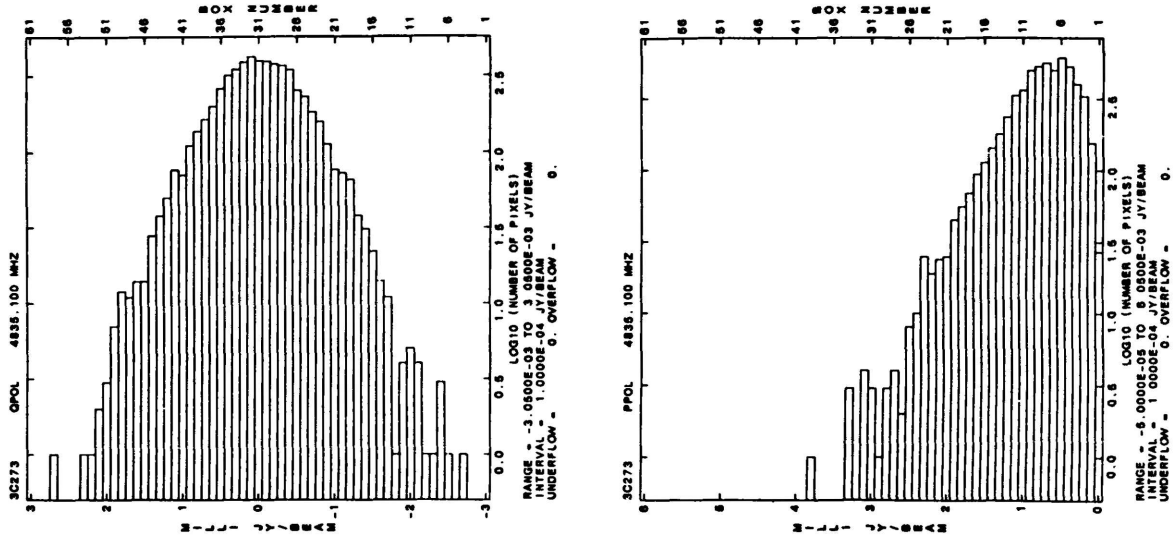


Figure 14–5. The noise distribution within an image. *(Left)* A nearly-Gaussian noise distribution. *(Right)* An approximately-Rayleigh noise distribution.

the image difference may be substantially worse (see Sec. 5.2). The resolution property of the combined image is identical to that of the composite images, and subsequent model fitting of discrete features is identical to that discussed above.

Since many reduction schemes are approximately linear, it may be possible to combine the images at an earlier stage in the reduction process, with a subsequent large saving of computer processing. In aperture synthesis, the calibrated data associated with different polarization states are generally combined before imaging and other application of reconstruction techniques. Many frequency channels can be combined, in the visibility data stage or in the image stage, before expensive deconvolution algorithms are used. However, slight registration problems, image scaling differences and nonlinearities, as discussed above, may preclude such combinations at an earlier stage of reduction. Some of these problems are mentioned in Lecture 12.

Although these derived images have well-defined resolutions, those images whose pixel values can be negative as well as positive can startlingly change with a modest amount of smoothing. For example, two neighboring features in such an image which are nearly equal in absolute intensity, but opposite in sign, will completely disappear when the image is smoothed. The same problem occurs with sums of images over frequency. With a sufficiently wide bandwidth, some strong features can disappear completely because of a cancellation of positive and negative intensities with frequency range.

4.4. Nonlinear image combinations.

Many useful properties of features can be derived from nonlinear combinations of im-

ages. A partial list of such combinations (AIPS task COMB) would include:

I,	Total Intensity,	$I = \frac{\text{RCP} + \text{LCP}}{2}$,
m,	Magnitude,	$m = -2.5 \log I$,
V,	Circularly Polarized Intensity,	$V = \frac{\text{RCP} - \text{LCP}}{2}$,
P,	Linearly Polarized Intensity,	$P = \sqrt{Q^2 + U^2}$,
ψ,	Linear Polarization Position Angle,	$\psi = \frac{1}{2} \tan^{-1} \frac{U}{Q}$,
F,	Fractional Linear Polarization,	$F = \frac{P}{I}$,
RM,	Rotation Measure,	$\text{RM} = \frac{\psi_1 - \psi_2}{\lambda_1^2 - \lambda_2^2}$,
D,	Depolarization,	$D = \frac{F_1}{F_2}$,
I_C,	Continuum, wide-band,	$I_C = \sum I_i$,
$I_{\text{line},i}$,	Line Emission, channel i,	$I_{\text{line},i} = I_i - I_C$,
$I_{\text{abs},i}$,	Line Absorption, channel i,	$I_{\text{abs},i} = I_C - I_i$,
τ_i,	Opacity, channel i,	$\tau_i = \exp \frac{I_{\text{abs},i}}{I_C}$,
α,	Spectral Index,	$\alpha = \frac{\log I_1/I_2}{\log \nu_1/\nu_2}$,
B,	Blanking,	Blanked image = I where $I > 0$, else 0 or a "magic value" ignored in subsequent processing.

Many more combinations are possible, of course.

The major complication of such secondary images is that the error distribution can vary enormously from pixel to pixel and, in fact, some output pixel intensities may be undefined. For example, the spectral index between two frequencies is proportional to the logarithm of the ratio of the intensity on the two images. If the intensities are of different sign, the spectral index is undefined. Generally a peculiar number is assigned to such a pixel which is then suitably ignored in subsequent analysis.

An error image whose pixel values are defined as the error of the derived intensity is the best solution to the image error problem. Of course, accurate knowledge of the error distribution of the composite images is necessary. In order to avoid the generation of an additional image, a simpler approach is often taken. In the derived image, those pixels whose errors are greater than a specified amount are blanked or given the special intensity assigned to an undefined pixel. The remaining pixels are of low error, although their relative weights are lost (AIPS tasks RM, COMB). Further analysis and display algorithms then can be used with these blanked images, without having the confusion of pixel intensities which are grossly in error. Often, the most useful blanking level is unknown, so that the image combination must be repeated several times to give the desired displays or to produce useful analysis.

The resolution of the images of these quantities is equal to that of the composite images. Further smoothing of the "nonlinear" images must be done on the input linear images, and then the appropriate combination recalculated. Useful displays of these quantities are

difficult to generate because of the large errors and the rapid changes of the quantities between resolution elements. For complicated features, one-dimensional displays like that shown in Figure 14–3 can be very useful.

Other special analysis techniques associated with a set of spectral line images (a set of images at equal frequency intervals) are covered in Lecture 12. These images are often analyzed in order to determine the velocity characteristics of galaxies. Many of these techniques are useful for other types of image sets.

Angular quantities are particularly nasty because of the 180° ambiguity of angles. Calculation of rotation measures and magnetic field orientations must be done carefully to remove such ambiguities.

5. Selected Image Analysis Topics

In this section I have chosen a potpourri of topics which are important for 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. Problems associated with noise-dominated images.

Often, even though an image is dominated by noise, parameter values or detection limits of possible weak features may be useful. First, plot a histogram of the number distribution of the intensity (AIPS task IMEAN) to determine whether the distribution is compatible with the expected noise. Such a distribution from the VLA is shown in Figure 14–5a. For the VLA, the noise caused by the receivers should be normally distributed with mean near zero. Because I have plotted the logarithm of the number, the shape of the distribution is parabolic. The highest positive and negative intensities should be about four times the r.m.s. of the distribution. An extended positive tail may be produced by faint sources. If the negative tail is also extended, this is an indication of an additional error component. Of course, the r.m.s. of the intensity distribution should be consistent with the observational parameters (see Lecture 10). The error distribution for derived nonlinear parameters is not Gaussian in general. For images of the total polarized distribution and for those from most total-power telescope systems, the noise distribution is Rayleigh (see Figure 14–5b). In this case, the presence of many weak sources, along with large systematic errors, produces an extended positive tail in the number distribution.

Parameter estimation of a discrete feature whose peak is greater than four times the image r.m.s. can be handled using the techniques of Section 2. For weaker features, many model fitting techniques may give a significantly biased solution because of the presence of the noise. Also, noise-dominated images are likely to have undergone little in the way of sophisticated processing, like 'CLEAN'ing and self-calibration, so that the point spread function may not have a shape at all resembling that of a Gaussian, but, instead, be of a rather more complicated shape (associated with the aperture sampling).

The following procedure is suggested for determining the intensity, or an intensity limit, at a particular position in an image. For weak features, it is not advisable to determine the size of the feature directly. Use the same analysis used in the illustration for Figure 14–4. The box area should cover the width of the point spread function between its zeros. Determine the integrated intensity, F, at the desired position. Then, determine the integrated intensity in several control regions which surround the position. Use these control regions to define a bias, C, and a scatter ΔC. An estimate of the integrated intensity, F', at the desired position is given by

$$F' = \begin{cases} F - C \pm \Delta C, & \text{for noise symmetric about 0,} \\ \sqrt{F^2 - C^2}, & \text{for noise asymmetric about 0.} \end{cases} \tag{14–7}$$

To establish the intensity scale, integrate the point spread function over the same area. This summation is then the integrated intensity which corresponds to one unit.

A conservative method of determining the size of a weak feature is to repeat the above analysis on a smoothed version of the image. If the integrated intensity of the feature increases significantly, then an approximate size can be estimated from the ratio of the intensity at the two resolutions.

5.2. Image bias problems.

Images often have a bias level which varies smoothly over the image field. In aperture synthesis images, these biases are produced by the lack of measured low spatial frequency Fourier components. For distinguishing and analyzing discrete features, most of the effects of the bias can be removed by solving for a zero bias, and perhaps for a linear slope under the feature. However, it is difficult to separate the bias variations from very extended features. This is a particularly nasty problem when the extended feature changes considerably over a set of images, because the shape of the bias depends on the amplitudes of the missing Fourier components—which in turn depend on the properties of the extended feature. Some improvement can be made in the deconvolution process by including a zero-spacing flux density estimate (see Lecture 8).

Further decrease in the bias problem can be obtained in two ways. The most straightforward method is to use the same array, but at lower resolution. Assuming a similar instrumental configuration in other regards, a simple combination of the data, together with re-imaging, will produce much better-quality images. Additional observations at lower resolution—with another array or with a filled-aperture telescope—will also work, but there are several complications. For example, the sensitivity of each instrument over the image field may differ between the telescope systems. Correction for this effect must be made before the visibility data are combined or before the images are combined. Further discussions concerning the practical problems of combining filled-aperture data with synthesis array data appears elsewhere (e.g., Bajaja and van Albada 1979).

5.3. Image intensity scale.

There are two intensity scaling calibrations: the peak image intensity associated with a point source in the sky, and the integrated image intensity associated with the resolution of the instrument. The first calibration is discussed in Lecture 4. The scaling factor is the ratio of the intrinsic strength of an observed point source to its measured intensity. For synthesis arrays, the visibility amplitude of the point source is used as the measured intensity. For optical images, the peak intensity on the image of a star is used. If many calibration stars are in the image field of view, the nonlinearities of the intensity scale can also be determined.

If a feature in the sky has an integrated flux density of F_s and an area of A_s, then its mean peak intensity is $P_s = F_s/A_s$. However, the measured peak intensity, P_i, in the image will depend on the resolution and on the detailed structure of the point spread function. The relationship between P_s and P_i is important, and it represents the second intensity scaling problem.

If the point spread function is very nearly Gaussian shaped, or at least well-behaved (monotonically decreasing with distance from the center, with no negative intensities), one can measure the weighted area of the point spread function by integrating its intensity in a region of nonzero response. Call this area A_p. The relationship of P_s to P_i is then

$$P_s = \frac{P_i A_s}{A_s - A_p}. \tag{14-8}$$

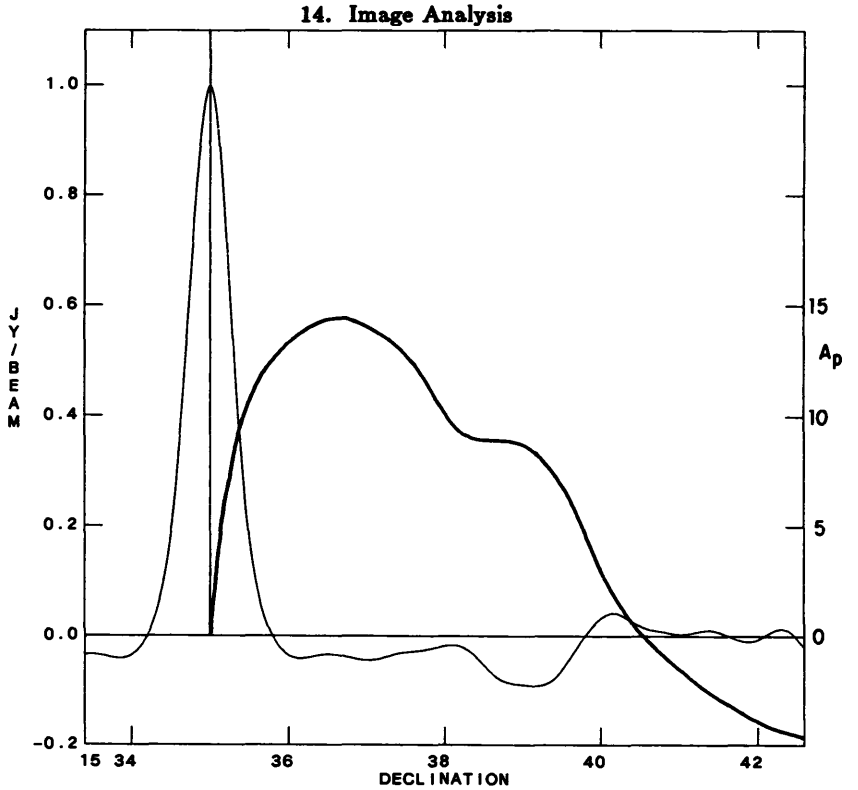


Figure 14–6. Some characteristics of the point spread function corresponding to a VLA image. *(Light)* The north–south intensity distribution of the point spread function. *(Bold curve)* The dependence of A_p with radial distance from the image center.

The point spread functions associated with the instrumental response are not usually well-behaved. An example is shown in Figure 14–6. The point spread function in the north–south direction has a long negative sidelobe with subsequent ripples, at the 5 to 10% level. The area of the point spread function, A_p, within the specified radius reaches a maximum just after the zero and decreases slowly, and even becomes negative. Any simple normalization of the type given in the above Equation is difficult if the feature is larger than the size at which the point spread function area begins to decrease.

One remedy is to try to reconvolve the image with a more desirable point spread function, by 'CLEAN'ing or by means of other techniques. This can be done on images which are noise-dominated, but it is very expensive in computing time. Another method is to smooth the image heavily enough that the feature is in fact not much larger than the point spread function. Finally, it is possible to calibrate P_e and P_i as functions of the feature size, by using known models in the sky and determining their peak intensities on the image.

5.4. Motion of features.

Determination of the relative motion of discrete features in different images (corresponding, say, to different epochs of observation) is limited in accuracy by the resolution of the images and the signal-to-noise ratio. If the resolution is R and the signal-to-noise ratio at the peak of each of the features is S, then the sensitivity to a displacement is approximately $R/2S$. This assumes that there are no registration errors between the two images. These can be minimized by suitable calibration of the data. If the images contain other features, then registration can be accomplished by attempting to superimpose all of the features and ignoring those which are obvious outliers. This is equivalent to the optical image comparison technique of blinking. Accurate parameters can be derived by model fitting the positions of the features on each image and then taking the appropriate differences.

REFERENCES

Bajaja, E. and van Albada, G. D. (1979), "Complementing aperture synthesis radio data by short spacing components from single dish observations", *Astron. Astrophys.*, **75**, 251–254.

Greisen, E. W. (1983), "Non-linear coordinate systems in AIPS", NRAO, AIPS Memo No. 27.

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", *Ap. 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", *Ap. J. Suppl.*, **54**, 291–334.

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

15. Data Display: Searching for New Avenues in Image Analysis

ARNOLD ROTS

1. INTRODUCTION

Display of data is an immensely important area in data processing, since it functions as the prime interface between the user and his data as the latter are being processed. Yet, most people are relatively unfamiliar with its potential, usually because the necessary manpower is not available to implement sophisticated software for display devices, so that most of the capabilities remain hidden for those working on data analysis software as well as for the users. For instance, the IIS image display systems used in GIPSY[1] and AIPS are quite powerful, not only in rendering very flexible image display, but also in performing certain simple image analysis functions. Full use of those capabilities would make data processing with these systems easier for the users, while it also could reduce part of the burden on the host computer. In neither system, however, are those capabilities fully realized; the IIS systems are basically used as frame buffers for the TV monitors, with rudimentary slope-and-intercept transfer function control. Also, from users' responses to more advanced features available in only one of these systems, it has become clear that users deem such features essential only after they have gained experience with them. This is true for sophisticated display techniques as well as for image analysis functions which provide interactive "quick-and-dirty" preview of equivalent functions available in the host computer; when done in the host they take longer and require more resources, thereby often precluding extensive experimentation.

Of particular concern is the processing of three-dimensional data, where users have so far been forced to interpret their observations from displays that are essentially mere modifications of designs made for the two-dimensional case. Some research has been done in astronomy on display of three-dimensional data. This has only scratched the surface so far, although it has revealed some enticing vistas.

These are the matters I would like to address in this Lecture, in the hope to raise people's awareness of the opportunities that are available right now and the possibilities that may become feasible through a concerted effort in research, design, and hard coding work. One should not imagine that the display tools come for free. It will need more manpower than traditionally has been invested in this area; but it will be warranted by the enhanced power of the data analysis system as a whole.

In terms of hardware and applications there are two types of data display: graphics (line drawing) and image (gray-scale) display. Both may be in monochrome or in color. In the next Section I will try to define the objectives for the design of display systems. Then, in the following two Sections I shall deal with the two display device types. Finally, there will be a discussion of off-line support functions and a concluding Section.

[1]GIPSY is an acronym for *Groningen Image Processing System*, a computer package used in the Netherlands for reduction of Westerbork Synthesis Radio Telescope (WSRT) data. — *Eds.*

2. Objectives for Data Display

The data reduction and analysis process consists of a string of reduction and analysis functions applied to the data. At each point in this sequence the astronomer has to decide on subsequent processing, based on his assessment of the data and his own objectives. He also wants to understand his data. The main objectives of data display are to facilitate this by giving the astronomer an instant overview of the contents of his data, as well as enabling him to lift out details. These functions should be performed fast, interactively, and under easy control of the user.

This may not be the most complete definition of data display objectives, but it is certainly a valid one and, in my opinion, quite adequate in the context of this Lecture.

The objectives may also sound rather trivial. One should be aware of the implications, however. "Instant overview of the contents of the data" means considerably more than just "a picture of the data"; it requires precisely such a display that the full contents can be grasped instantaneously, which is by no means a trivial requirement. Also implied is a good deal of capability to "play" with the data. Ideally, only "good" intermediate results should be calculated and stored by the host computer. Tools to determine the correct parameters and processes to achieve this should be provided, at least in part, by the display system. In some cases flexible display is sufficient to make a sound decision; in other cases one needs interactive "quick-and-dirty" versions of analysis functions present in the display device for preview.

Easy control of display functions similarly has far-reaching consequences: language—and thus interaction through a character keyboard—is a poor medium to control displays. Like theatre lighting systems, displays are much more easily controlled by analog devices such as switches, knobs, slides, and buttons.

3. Image Display

For aperture synthesis observations there are three types of data to be displayed, each with its own specific requirements: visibility data, two-dimensional images, and three-dimensional images (predominantly spectral line data). There are two elements in the application of display devices to the user's data: the pure display function which tries to translate as many bits as possible into a comprehensible image, and image analysis which allows the user to manipulate the bits interactively in order to gain an even better understanding. It will be clear that the distinction between the two is not always sharp. Finally there is some specialized hardware to be considered that is extremely helpful for the enhancement, the efficiency, and the ease of use of display systems.

3.1. Visibility data.

Aperture synthesis visibility data consist of complex numbers as a function of two coordinates: baseline and time, or u and v. Image display of these data forms a powerful tool in editing and calibration of the observations since discordant data values are easily recognizable when arranged in such a way that large-scale patterns are to be expected. Depending on the nature of the observations and the deficiencies to be detected, either baseline–time or (u, v) display may be preferable.

For detection of bad receiver behavior, for instance, a baseline–time display is the most revealing. Care should be taken that the baselines are arranged such that patterns caused by source structure can easily be recognized. This probably means for the VLA that intra-arm baselines for each arm should be grouped together, while inter-arm baselines need careful sequencing.

The most generally useful way to represent complex quantities is through an intensity-hue display (see also Sec. 3.3.5); here amplitude is represented by the intensity of the pixel, phase by the color (hue). A cyclic color scheme (blue-cyan-green-yellow-red-magenta-blue) is most appropriate for this purpose since then there will be no discontinuity at phases of $\pm 180°$ or $0°, 360°$. The display system should allow the user to independently vary the transfer functions for amplitude and phase in order to suit the particular range of values he is interested in at any one moment. This is an important requirement that not only determines the versatility and usefulness of this display, but also restricts the choice of acceptable models of image display machines since most models are incapable of supporting this feature.

Interactive support functions should include: dynamic amplitude, phase and antenna pair display for pixels selected under cursor control; and flagging (editing) of pixels under cursor control.

3.2. 2–D images.

Two-dimensional images are the traditional realm of image display systems. We assume that we are dealing with astronomical images containing some variable (e.g., brightness distribution) as a function of two coordinates—either two spatial coordinates, or one spatial coordinate and velocity. In this Section we shall restrict ourselves to the techniques that optimize the display of such a single image.

Image sizes will vary considerably, most commonly from 64×64 to 2048×2048. Image storage for at least 1024×1024 should be provided. TV monitors that can display a pixel grid of this size are available, although they are not (yet) common. One also ought to consider whether anything is to be gained by switching to this size monitor from the standard 512×512 displays, considering the amount of detail the eye can take in.

Some display systems (e.g., Vicom) will allow optimal use of memory space because they are capable of handling arbitrary image sizes which are not tied to fixed positions in memory.

3.2.1. Zoom and pan. Zoom and pan are very essential features, both for lifting out details and for accurate positioning of the cursor (down to the pixel level). Most systems provide zoom factors of 1, 2, 4, and 8; zooming is usually accomplished by pixel replication. Ideally, one would want any integer zoom factor, which some systems provide, a negative zoom (to display 1024×1024 images on a 512×512 screen), which is extremely rare, and some sort of interpolation instead of pixel replication, which is not available in hardware. In general, however, these wish list items either are non-essential niceties or can be emulated in other ways.

If the monitor screen is no bigger than 512×512 one wants at least a roam feature that allows interactive roaming through a 1024×1024 image.

Some systems have nasty definitions built into them. The IIS, for instance, has a very inconvenient definition of the zoom center.

3.2.2. Color schemes. Although a gray scale representation is often quite adequate and usually provides more dynamic range than most color schemes, pseudo-color is a useful option to have. Schemes presently in use include spectral colors and a variety of discrete (contrasting) color palettes. The latter options are often referred to as color contours and include the capability to compress or stretch (recycle) the colors which is especially useful for optical data.

There definitely is a need for multiple color schemes; choice of a particular scheme depends on the data, the user's taste, and the user's physiology (such as partial color-blindness).

It is highly desirable to be able to decouple the definition of the color scheme from the transfer function (see Sec. 3.2.3). This requires at least two stages of look-up tables, as IIS provides. Some other manufacturers provide this capability with a special color table, which is less flexible.

3.2.3. Transfer functions. Transfer functions form the heart of image display, in a sense. The question is how to transform the range of intensities present in an image—or, the part of that range that the user is currently interested in—in such a way that the maximum amount of information can be perceived by the human eye. There are two parts to solving this problem. First, a transformation has to be performed on the image intensities as they are kept on a storage device, to compress them into 8 bits; this is usually not a problem when the dynamic range of the image is less than about a factor of 100. Then, the 8 bit values in the display device have to be transformed into intensities on the monitor. On the whole, the approaches to both transformations are similar, although the requirements at any given moment may be different; e.g., the user may wish to load as much of the image intensity range into the display memory as possible in order to look at a number of sub-ranges in detail.

There are two classes of requirements for the transfer functions. One is to use the available viewing (or display memory) dynamic range as efficiently as possible for the entire range of intensities. How this is done depends on the data. For images that are buried in the noise, for instance, a straight linear transfer function (1:1) may be used, while high-dynamic range images (with relatively few points near the high end of the intensity range) usually profit from a logarithmic or a histogram-equalized transfer function. Histogram equalization is a particularly useful and powerful method that allocates the available output intensities as efficiently as possible by distributing them over the input intensities on the basis of the distribution of those input intensities. If the image device is equipped with a 16 bit mode and a histogram generation option, the bulk of the work for this technique can be done on the display device which significantly unburdens the host by halving the number of required I/O operations on the image. Whether used in loading the image or in displaying it, histogram equalization has the effect of maximizing the amount of detail one can see in a high dynamic range image, thus enabling the user to grasp as much of the contents of an image as possible from one picture. It is useful to have several options available for the equalizing algorithm since not all images can be handled the same way. Another important feature is the possibility to interactively vary the range in the histogram to which the equalization algorithm is being applied.

The second class is to lift out part of the input intensity range to look at it in great detail; it means concentrating the bulk of the output intensity range on a more or less small part of the input range. Traditionally, this often has been done by slope-and-intercept transfer functions. They have the advantage that they are very easy to control interactively since there are only two input parameters with functions that are easily understood by the user. However, they are also rather crude and do not allow designating a small portion of the output range to the remainder of the input range. Two-kink/three-segment transfer functions are far superior in that respect, but not perfect yet and control is more involved. It should be possible, though, to develop a family of transfer functions that is better than both of the previous ones in either respect.

Often, the quality of a picture can be improved even more by combining the use of sophisticated transfer functions with two-dimensional filtering functions. These will be discussed in Section 3.3.6.

3.3. Image analysis.

Modern image display devices usually incorporate a certain amount of processing capability that allows the user to perform a number of different basic operations on entire, or parts of, images in a very short time (typically one refresh cycle per operation). By stringing several basic operations together more sophisticated processes can be performed. There are two advantages to making use of these facilities: doing them in the image display device relieves the host computer from these tasks and the display machines can perform them much faster (essentially in an array processor fashion) with instant display, making it feasible to do it interactively. The disadvantage is that all operations are performed with limited accuracy (8 bits), although in certain cases one may be able to use 16 bit options, provided there is enough refresh memory. There are five areas of application:

- The case where a quick-and-dirty operation is sufficient for the user's purpose. An example is when the user wants to see the difference between two images and a limited accuracy subtraction in 8 bits is enough to show what he wants to see. Intensity-hue display essentially also falls in this category.
- Enhanced image display such as histogram equalization, intensity-hue display and the use of two-dimensional filtering for display purposes.
- The case where the user wants to handle the data interactively in order to make decisions on further processing. An example in this category is using a two-dimensional smoothing implemented in the display device (fast, but with limited accuracy) in order to determine the optimal parameters for a proper smoothing on the original data.
- Image statistics. Image display devices that have blotch and histogram functions offer the capability of performing image statistics (average, sum, r.m.s., median, histogram, etc.) of selected areas of an image almost instantaneously and usually with enough accuracy. The retrieval of single pixel intensity values is a special case of this.
- Definition of image segments, regions of interest, or blotches. These can be used for blanking or differential treatment either in further processing or for display functions.

In the following I shall briefly describe some analysis tools, not according to the applications outlined above, but rather according to the techniques used.

3.3.1. Image segments. Defining different segments in an image is often needed for a variety of applications: outlining the area in which one believes the source is contained or, instead, defining an area in which one believes there to be no radiation, for such purposes as "Clean boxes", spectral line windowing, image statistics, and different application of transfer functions. The most natural way to define these areas is by drawing polygons in the image itself as it appears on the TV monitor, with the aid of a pointing device (light pen, tablet, track ball) and outlining them in a graphics overlay plane. This is especially powerful if the image display device offers the capability of distinguishing between the "blanked" and "non-blanked" regions in performing statistics and arithmetic and logical operations.

3.3.2. Image statistics. Statistics of selected areas in an image provide essential information on the quality of the data and form an indispensable supplement to the image itself. The statistics include: average, median, sum, standard deviation, and histograms of intensity distributions. If the image display device is equipped with a histogram generator which can operate on the entire image, in the region of interest (see Sec. 3.3.1) or outside that region, the statistics can be obtained very fast and usually with enough accuracy. Increased

accuracy can be obtained by reading the image again from disk or by working in higher precision arithmetic, if it is available.

Histograms are also important for fast processing of histogram equalization algorithms (see Sec. 3.2.3).

3.3.3. Comparison of images. In judging the results of image processing techniques the user often wants to compare two images. The technique that has a long tradition in astronomy for this purpose is the blinking of the two images. This can easily be accomplished in a display device by loading the images into different image planes and then switching the monitor at some rate, interactively controlled by the user, between the two images. This requires the two images to have very similar appearance which can be done by fiddling with the transfer functions while both are shown in a split screen mode.

Another technique is to put both images on in split screen mode and give the user interactive control over the position of the split, so that he can move it back and forth over the area of interest. Although it depends on the personal preference of the user, this technique is on the whole more effective than blinking.

Finally, subtracting the two images may also provide a very effective means of comparison. This can be done by subtracting the original images in the host computer and displaying the difference in the display device, or by performing the subtraction in the display device itself, which is infinitely faster but has limited accuracy (see Sec. 3.3.4).

3.3.4. Arithmetic operations. Arithmetic operations between images are very common. Adding (or averaging) two images and subtracting two images from each other (either for comparison or subtracting the continuum from spectral line data) are obvious, but multiplication and division also occur. Image display devices can perform such operations very fast, by using an Arithmetic and Logical Unit in conjunction with a Feedback Unit, and/or the Look Up Tables. When two image planes are switched on, their contents (as modified by the Look Up Table) are added. Addition can be achieved by applying two linear transfer functions. Subtraction, by a positive and a negative linear transfer function. Multiplication and division can be accomplished by applying logarithmic transfer functions to the individual images and an exponential transfer function to the sum; this, by the way, requires separate Look Up Tables for each image plane, as well as for the sum (see also Sec. 3.3.5).

Having these functions available in the display device is useful for a quick-and-dirty preview and for cases where the limited accuracy is acceptable and the result does not have to be kept.

3.3.5. Intensity-hue display. Intensity-hue display forms an excellent tool for situations where the user wants to view two parameters simultaneously in two-dimensional space, especially when one of them carries intensity information and implies a credibility criterion for the other. Examples are: column density and velocity (spectral line), percentage polarization and polarization angle, amplitude and phase (visibility data), total intensity and percentage polarization, continuum flux density and optical depth, and flux density and spectral index. The former of the parameters in these pairs controls the intensity of the image, the latter the color (hue).

It will be clear that human beings in general think in terms of intensity, hue, and saturation (often the last two are combined into the household term "color"), rather than prime color mixes. Although there is nothing particularly difficult in making image display devices follow this natural trend, oddly enough, the industry is still in the Middle Ages where the architecture of the devices (with one exception) is dictated by the (TV monitor) hardware; henceforth, the programmer has to think in Red, Green, and Blue. Unfortunately,

since image display machine designers seem to be so far removed from reality, in most of these machines it turns out to be impossible to emulate intensity and hue display. IIS is one of the very few that can do it, for the same reason why it can multiply two images: it has three separate Look Up Tables for each image plane as well as for the red, green, and blue sums.

One should be aware that the range of fully saturated colors generally used in intensity-hue display does not offer a large dynamic range to the eye. It is therefore mandatory that the transfer functions mapping the pixel intensities into screen intensity and color can be controlled interactively and independently by the user. The choice of colors deserves some consideration and depends on the data being displayed. Angular units (phase, polarization angle) usually benefit from a cyclic color scheme, since they are cyclic themselves (see Sec. 3.1). For velocity, the jargon (redshift/blueshift) suggests a spectral color sequence, while for the other parameters either a spectral sequence, or a sequence going from "dark" to "light" colors is the most appropriate.

It should be emphasized that intensity-hue display is primarily an analysis tool which enables the user to recognize or discount certain features in his data quickly. When looking at the two parameter images separately, he might have missed this interpretation, or only arrived at it with considerably more effort.

3.3.6. 2-D convolutions. Two-dimensional convolutions (or filtering) have two basic applications in image analysis: for changing the image resolution (smoothing) and for image enhancement. The user may want to smooth (degrade the resolution of) his image in order to increase the signal-to-noise ratio for weak extended features, to make a strongly elliptical "beam" rounder, or to bring his image at the same resolution as other data for comparison; this last case may involve observations at different frequencies or observations made with another instrument. Two-dimensional filtering for image enhancement purposes usually involves combining original and/or smoothed and/or edge-enhanced images. It is interesting to see that these two applications can functionally be lumped together. They are both essentially image enhancement operations: one enhances the image through a low-pass filter bringing out extended features buried in the noise, the other through a high-pass filter accentuating the small scale details. Some examples of filtering are shown in Figures 15–1 and 15–2.

Although convolutions are conceptually simple operations, they do require a fair amount of resources when done in the host computer and can, in that mode, not be done at interactive speeds. Image display devices with an ALU/Feedback option can perform the basic operations needed for convolutions (multiplication, translation, and addition of entire images) at very high speed (burst speeds of 15 million operations per second) and have the added advantage that the result is immediately available in the display device for viewing. Admittedly, the precision is limited since one usually deals with 8-bit images, but in general 16 bits can be used for accumulation and possibly even for the entire operation if enough memory is available.

Even so, the result of such a limited precision convolution may be able to tell the user what he wants to know, or at least give him the information needed on optimal parameter choice to initiate a "proper" convolution operation in the host. Experience has shown that users, due to the lack of interactive capability of this sort, tend to experiment very little with convolution parameters, even though obvious improvements could be made. Rough timing estimates provide the reason for this: starting a convolution operation in the host and displaying the image on the monitor will typically take at least 3 to 5 minutes of real time, while the image display device can perform the same task in less than 15 seconds.

The field of two-dimensional filtering for image enhancement has hardly been explored

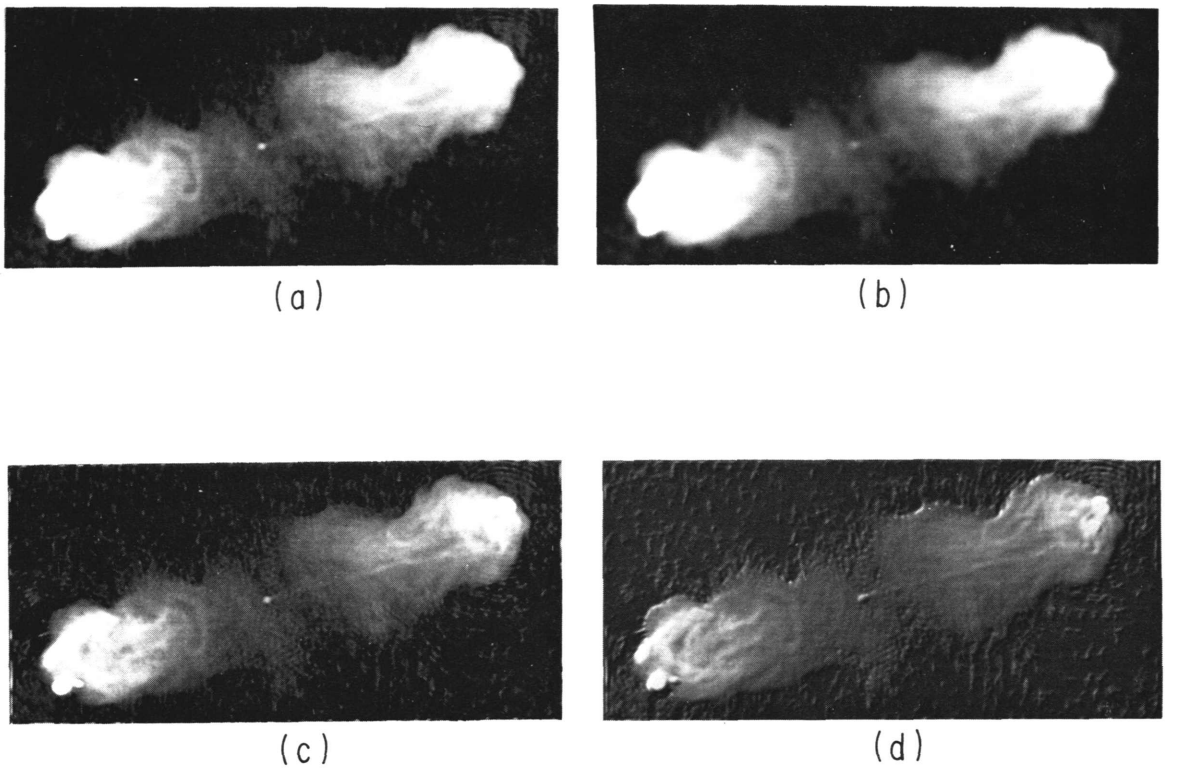


Figure 15–1. Original and filtered images of Cygnus A: (a) original; (b) low-pass filtering (smoothing); (c) high-pass filtering (sharpening); (d) edge-enhancement combined with the original. The filtering was done inside the IIS.

in radio astronomy, although users are increasingly becoming aware of the usefulness of high quality image display. It is not quite clear at the moment whether final quality display involving filtering can be handled with the limited precision of image display devices—although it may—, but it certainly would be useful to provide the option as part of the standard interactive image enhancement tools available during the data analysis process. An experimental program, MFILTR, doing this kind of two-dimensional filtering has been implemented on the PDP–11/44 computer DISPLY (recently replaced by a VAX–11/750 computer) and its IIS image display device at the VLA site. Figures 15–1 and 15–2 were produced by MFILTR.

3.4. 3–D images.

An especially massive display and data interpretation problem in radio astronomy is posed by aperture synthesis spectral line observations. In this case the user ends up with a three-dimensional image where the coordinate axes are formed by two spatial coordinates on the sky together with a frequency (or Doppler velocity) coordinate. For the interpretation of such data it is imperative that the user have a means of forming at least a mental picture of the brightness distribution in all three dimensions simultaneously; a clear understanding of the three-dimensional structure and continuity of the object(s) is necessary for intelligent analysis of the data. So far, this mental picture has usually been built up from a large number of displays of two-dimensional cross-cuts through the data cube, but obviously that assembly process could be made much faster and much more efficient if the data could be displayed in three dimensions directly. The following Sections will deal with some attempts at implementing such display tools. From this introduction it will be clear that the main objective is to provide tools for use during the analysis process, not the generation of fancy

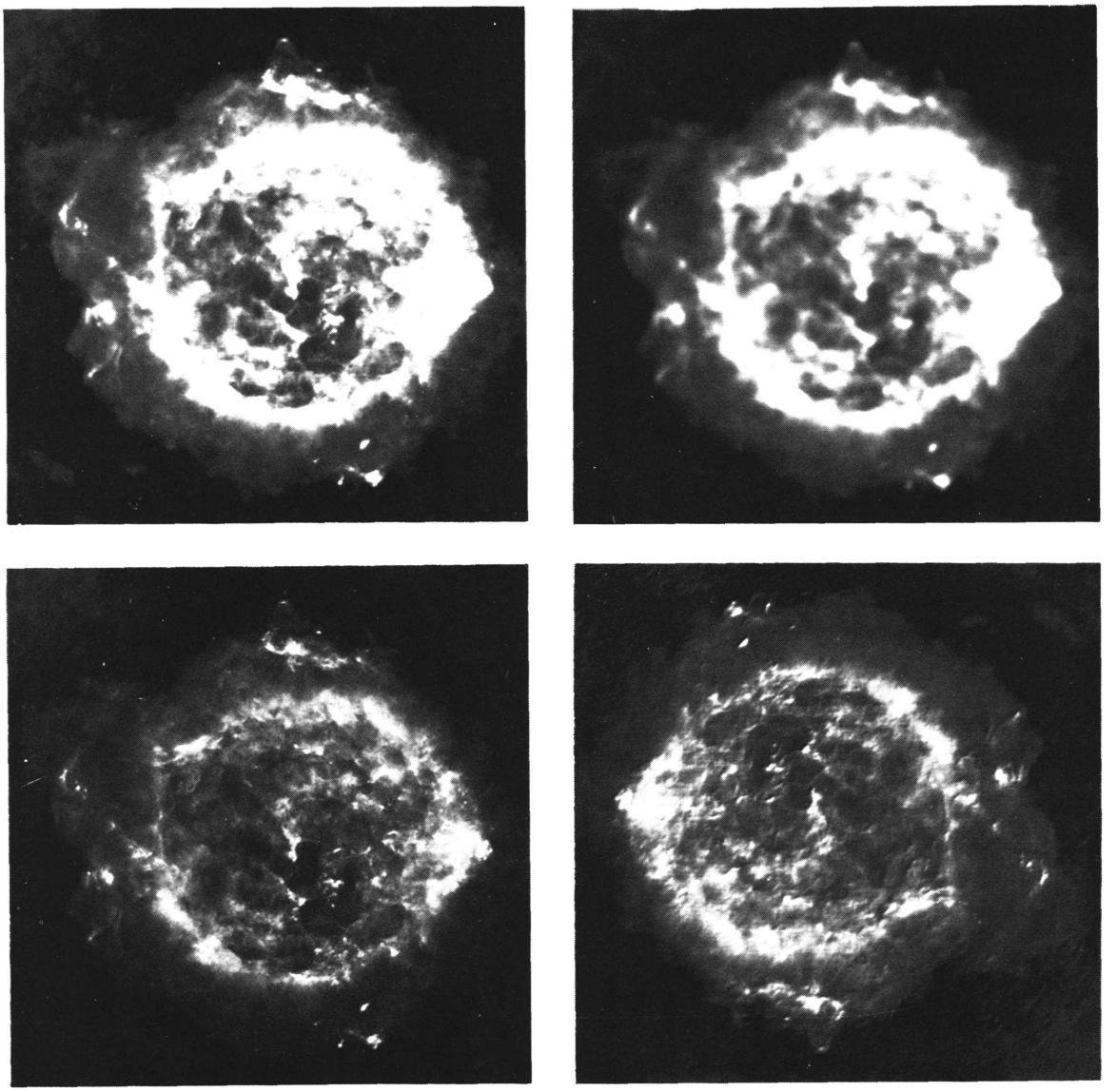


Figure 15–2. Original and filtered images of Cassiopeia A. The 2–D filters are the same as in Fig. 15–1.

final presentation displays.

3.4.1. 2–D representations. Because of the nature and geometry of common display materials (paper, TV monitor screens) two-dimensional representations will remain important. But in addition there is the fact that the three-dimensional displays that we can presently conceive of are not easily quantifiable. Hence, there is a need for two-dimensional displays, both as a substitute and for work copies to look at quantitative detail after the user has gained a three-dimensional understanding of the total contents of the data.

A mosaic of two-dimensional cross-cuts is very useful for the latter purpose. One may think here of contour plots in the spatial coordinates, one at each observed velocity, or in velocity and one spatial coordinate along parallel lines. It is in general useful if the display program can automatically compose such displays in a mosaic directly (like, for instance, the AIPS task KONTR).

Another application of a stack of two-dimensional (gray-scale) images, covering the entire cube, is to put them into a time sequence (animation) and to "travel" throughout

the cube in time. An example of such a program is the AIPS verb TVMOVIE. This technique is capable of aiding in the detection of three-dimensional continuity in a rather powerful way. Its drawbacks are that it requires the user to mentally accumulate what he sees over time in order to build up the three-dimensional picture in his mind, and that the display cannot be captured in a hardcopy to be studied at leisure. However, the latter disadvantage may be overcome by recording the scene on video tape and playing it over and over again on a cheap viewing station. The main problem in creating these displays is fast loading of the individual frames so that a realistic animation effect can be achieved; normal loading of the data from disk into the image display device is far too slow. There are three possible solutions. First, one can load each frame and transfer it to video tape individually. This requires a video recorder with editing capability (in order not to loose the sync) and is rather tedious since the whole process has to be repeated for every loop one wants to record. Secondly, each frame can be loaded and transferred to a specialized image storage device, from where the sequence can be retrieved at high speed (and, if so desired, recorded on video tape in real time); examples of such a device are the video disk used with GIPSY in Groningen and the image storage device being developed by NRAO. Finally, one can use the image display device itself, preloading all frames in sections of the refresh memories and displaying them in rapid sequence. An image display device like VICOM is completely flexible in its allocation of refresh memory and allows this type of operation easily. In a device like an IIS one can display a zoomed image and rapidly change the zoom center; naturally, the individual frames are then restricted to 512×512, 256×256, 128×128, or 64×64. In using this technique one obviously has to contend with a trade-off between the size of the individual frames and the number of frames in the time sequence, the product being limited by the amount of available refresh memory. But on the other hand, it provides much faster and easier access to time sequences than the other two methods and is therefore a useful (quick-and-dirty) option, even when any of the others are available.

There is one more application of two-dimensional displays to three-dimensional data: the use of false color. If one has a series of two-dimensional images at different velocities, each velocity could be assigned a slightly different color in, say, a spectral sequence. Although powerful for some data, there are severe drawbacks; it is especially tedious to change the color transfer function (requiring a complete rebuild of the image, reading through all the data again). In practice, the display is usually limited to an image built from a two-parameter representation of the profile (the intensity as a function of velocity at each spatial pixel): zeroth and first moment. One can then let the former image control the intensity, the latter the hue, while it is possible to vary the transfer functions of both images (like in the AIPS verb TVHUEINT, developed by Jim Torson). See Section 3.3.5 for details. A disadvantage is obviously that one has lost all information on the exact shape of the profiles and that the human eye does not have a tremendous resolution along the saturated colors of the spectral sequence.

3.4.2. 3-D solids. What one ideally would like to do is put the contents of a data cube in space and view it from various angles. To do this one has to make a three-dimensional object out of the data, or a 3-D solid surface. Such an object can be constructed by wrapping a surface through the cube at a particular threshold intensity: everything inside the surface has a higher intensity than the threshold, everything outside a lower intensity. Basically this is the true three-dimensional equivalent of a two-dimensional contour plot with a single contour level. This three-dimensional solid surface object can then be painted on the TV monitor screen as seen from any viewing position, including ambient light, direct lighting, perspective, and depth cueing (diminishing intensity at larger distances). For an example, see Figure 15–3. In principle one could also add other features, like surface texture,

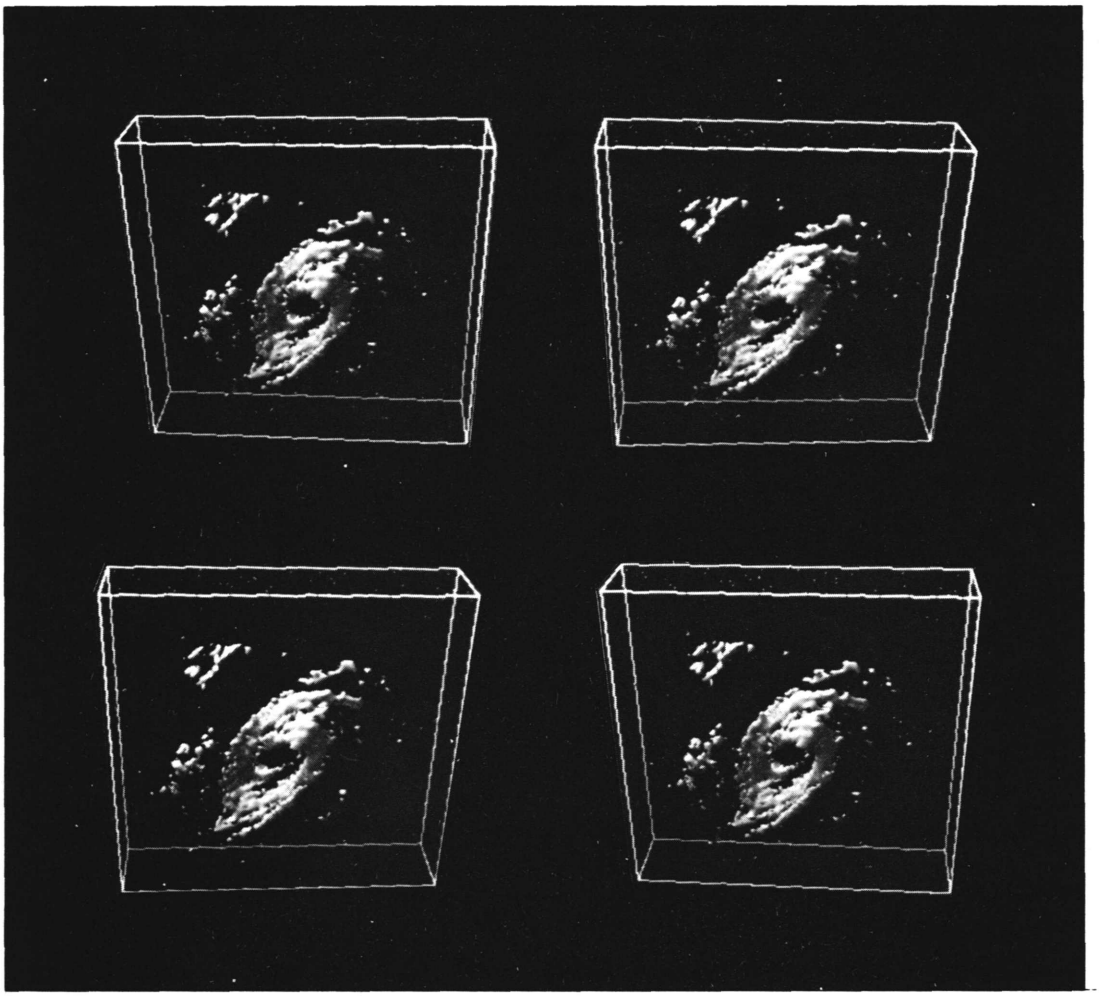


Figure 15–3. 3–D solid surface representation of HI observations of M81. The "short axis" is velocity; the main warp represents the rotation of the galaxy. Two stereo pairs are shown. The top one is "crossed": if one crosses one's eyes so that the right eye sees the left image and vice-versa, one may well be able to perceive the stereo effect without special equipment. The bottom pair is "parallel" and may be viewed in a similar fashion (but left image to left eye, right to right) or through a stereoscope.

reflectivity, true shading, and even transparency, but such sophistications do in general add so much overhead that they are not really warranted.

Increasing the threshold value would have the effect of peeling the onion skins off, getting closer and closer to the heart of the object. An objects can be cut open, in which case it may be advantageous to color-code the intensities seen in its bowels. It is useful to add a colored back drop in order to be able to discriminate between dark (shaded) parts on the object and holes in it.

Even though this is one of the better techniques to show three-dimensional structure, there are drawbacks inherent in these displays. The problem is that when we see a (two-dimensional) picture of a familiar three-dimensional object—such as a house, a chair, an animal—we can in our minds reconstruct the three-dimensional structure because of the familiarity: we have walked around these objects and we know how they are put together. This is not true for most astronomical objects, especially not in the strange (right ascension, declination, line-of-sight velocity) phase space. The images discussed in this Section do

therefore not convey enough information to enable one to understand the exact three-dimensional shape of the objects or to even decide (in the case of several unconnected structures) what is in front and what is in the back. Solutions to this problem are discussed in the next Sections.

3.4.3. Stereoscopic images. An obvious technique to turn one's attention to for the display of three-dimensional data is that of stereoscopic images. There are two ways of displaying the data and two techniques for viewing them.

The first mode of display is just to stack the images along the third axis in two pictures, with an offset depending on the position along that axis. This essentially produces a transparent object. More experimentation is needed with this form of display, but one of the disadvantages is confusion between foreground and background features.

The second mode is to produce a stereoscopic display of the three-dimensional solids discussed in the previous Section. This is done by generating two views with viewing angles slightly different in azimuth; this difference should typically be in the range 5 to 15 degrees. Figure 15–3 shows the result.

The best way to view a stereo-pair is through a stereoscope. For this purpose the two images can be put on the TV monitor and hard copies can be made which are then placed in the stereoscope; if necessary, the images may be reduced on a copying machine.

A fast way of viewing stereo-pairs is the use of anaglyphs: one image is put on the TV monitor in red, the other in green and the result is viewed through red-and-green "stereo-glasses". In positive, the right hand image is red, in negative it is green. Care has to be taken that the images do not "cross" too much in the back; they may have to be slightly offset in horizontal direction. In general, the TV screen phosphors are well matched to the lenses in these glasses. This mode of viewing is also useful for audience presentations where the use of stereoscopes is technically impossible. A more satisfactory technique is to use polarizers instead of red-and-green, but this involves more sophisticated machinery. For audience presentations, two projectors (each with the image for one eye) have to be used with orthogonal polarizers in front of the lens. One has to project onto a non-depolarizing screen and the audience has to be equipped with corresponding polarizing glasses; usually, linear polarizers (at 45 and 135 degrees) are being used, but circular polarization has great advantages because it makes the mounting of the polarizers less critical and allows more tilting of the viewers' heads. On a TV monitor something similar can be done. Stereographics Corporation will sell a complete system (including the monitor) for about $20,000, but it may be possible to do something simpler: one fills the odd lines of the image with one view and the even ones with the other; a circular polarizer and a variable retarder are placed in front of the screen; the variable retarder either lets light through unchanged or changes the sense of the polarization and is triggered by the vertical retrace; the user, finally, is equipped with circularly polarized glasses.

One thing one has to be aware of, however, is that the effectiveness of these techniques is highly subjective: some people simply do not have stereopsis—possibly because of some physiological defect—, some people are very good at it, and most people can "sort of see it", but could be good at it with some practice. Similarly, positive stereo-pairs work better for some people, while others prefer negatives; this may also depend on the objects, and may in particular be different for gray-scale and graphics images.

Finally, there are some three-dimensional imaging devices on the market now; one of these works with a vibrating mirror. However, it appears that the amount of data that such devices can handle is, at least at the moment, insufficient for our purposes.

3.4.4. Animation. By far the most effective way to show the three-dimensional structure of objects is to move them; for instance, by rotating the 3–D solids described in Section

3.4.2. Unfortunately, this is not always easy. For one thing, one has to construct a fairly large number of frames; one needs intervals no larger than 5 degrees and preferably half that. Then, the animation has to be effected; this can be done in four ways: using the zoom-and-pan of the image display device, using an image storage device with a high data rate to the display device, or making an actual animation movie, on film or video tape.

The zoom-and-pan option has been described extensively in Section 3.4.1. One should keep in mind, though, that the limited number of frames that can be displayed is probably a rather severe restriction for this application, since one is likely to deal with frames no smaller than 256×256. Nevertheless, it may be sufficient for a back and forth "rocking" of the object which may be satisfactory.

The image storage device will be dealt with in Section 3.5.3. It may be useful to point out here that to obtain a smooth animation (which this application requires) one needs to run at frame speeds of at least 10 Hz, which may not be achievable with digital image storage devices.

If one rejects the options above, one is left with making an animation movie, for which, incidentally, an image storage device still comes in handy. The problems involved in recording on video tape will be discussed in Section 3.5.4. The recording of animation movies on film is something we do have some experience with. Ideally, one would like to generate these on a high quality film recorder, such as a Dicomed film recorder. At the same time, though, one would like to use 16-mm film because processing, editing, and projection facilities are usually readily available. This can be done adequately on a Dicomed Model D–48 recorder; on a Model D–47, such as we have, it is not feasible. Alternatively, one can shoot a movie directly from a TV monitor screen. This requires some control software and equipment and one has to be aware that it must be done in single frame mode with exposure times of at least 1 second. This means that movie production, although not difficult, is a lengthy process—of the order of half a minute to one minute of recording time per second of screen time. Although the medium itself (film) is obviously more expensive than video—at about $10 per minute, original or copy—the equipment is cheap now that everybody is switching to video: used 16-mm cameras can be had for a few hundred dollars. Nevertheless, because of the hassle of set-up and processing, it does not lend itself easily for quick interactive work. At the VLA we have just been lucky to have 1–3 day turn-around available.

A new possibility that has recently emerged is the development at the University of Pennsylvania Hospital of a 3–D Solids machine that will provide the user with interactive control over the viewing of a three-dimensional solid surface (see Sec. 3.5.5), thus enabling real time animation.

3.4.5. Holograms. While thinking about display of three-dimensional data, one inevitably turns one's thoughts to holograms. It would especially be useful if production of white light holograms were easy and could be automated. However, the state of the art is currently not such that holograms are a practical medium for display in astronomy.

3.5. Specialized hardware.

There are some pieces of specialized hardware that are eminently suited to make life easier in interactive image display. They should be considered integral parts of the image display stations since they allow the user to realize the full capabilities of image display devices.

3.5.1. Control panel. Controlling the functions of an image display is, in the mind of the user, essentially an analog operation. The use of a keyboard for interaction with the image

is experienced as an extremely inconvenient intrusion of this cumbersome device. Manufacturers of display devices have realized this—to a certain extent—and provided trackballs, mice, light pens, tablets, push buttons, etc. They may be used to move cursors around and control various functions, like zoom, split screens, transfer functions, etc. Nevertheless, one has to resort to menus and multiple button definitions (not to mention multiple cursor definitions) in order to cover all control functions. And even then the system will at the very least be cumbersome to use, and probably confusing.

It therefore makes more sense to put the control of the image display device almost entirely on a hardware panel that looks analog to the user. The following Sections could be envisaged to be part of this panel. A module that controls which image planes are switched on. A number of transfer function modules that control transfer functions through rotating knobs or slides, with push button selection of the image planes they control. A zoom module. An image plane *cum* split screen control (including positioning of the split). A blink module with control of the planes involved, and of the blink rate. A graphics plane control module. A color scheme control module. And any others one may fancy. Such a control panel not only gives one easy—and easily understandable—control of the display, but it also provides an instantaneous status display. Basically, all this falls in the realm of ergonomics, and care should be taken that things are "right": the organization of the panel, the "feel" of the knobs, the slant of the panel, as well as placement of the monitor and lighting.

In principle the control panel could communicate with the image display device through the host, but it would make life easier if the panel could communicate with the device directly; this would also take a burden off the host. It would be necessary to build a small processor into the control panel.

If the display device cursor is used for any functions, it is helpful for the user to adopt different cursor colors to distinguish between these functions and to store text and/or numeric information in the cursor array (assuming that the display device has such a feature). The advantage of the latter technique is that this information is not zoomed with the rest of the image (as it would be when displayed through the graphics planes of IIS Model 70 machines) and that the information is displayed right at the center of the user's focus of attention. These techniques have been employed in GIPSY and in the program MFILTR (see Sec. 3.3.6).

3.5.2. Hardcopy. The capability of obtaining hard copies of the screen is a tool the user cannot do without. There are two types of hard copies, each with its own purpose and requirements.

The first will mainly be used for reference and as work copies. The main requirements here are: fast generation (less than a minute), reasonable size (at least 8×8 inches), and a material that can be written on. Black-and-white is sufficient for these purposes, and one does not need a tremendous dynamic range in gray scale intensity. The Honeywell hardcopy devices used at the VLA site are an excellent illustration of the kind of thing needed in this respect.

The second need for hardcopy is for pictures that can be shown to others to present the data. High quality in dynamic range, color, accuracy, and geometry is required here. Two forms of hardcopy are needed: slides (transparencies) and prints. Movie capability (film and/or video) could be added. Accurate film recorders and easy-to-use software are needed, as well as photographic facilities. The latter can in principle be fairly simple, but it probably would make sense to combine them with other, more sophisticated photographic needs such as producing overlays. Turn-around time can be longer (up to a few days).

3.5.3. Digital image storage. Experience has shown that there is a very important need for a capability for storage of images. There are two categories: just storing an image that has been fine-tuned for later retrieval (for comparison, display, or recording) and storing a sequence of images that can be run off as a time sequence. This latter function comes about because conventional loading of images is not fast enough to produce a realistic time sequence. The requirements for an image storage device are: fast loading (in both directions), flexible control, capacity of several hundred images, archival facility, and, preferably, a digital format.

Fast loading enables realistic time sequences and avoids irritation at having to wait for the device to do its job. Flexible control (of especially the time sequences) can be effected through the control panel described in Section 3.5.1. Digital format will allow the user to fiddle with the image after retrieval and makes the images compatible with any other images loaded from the host; this includes the capability (if adequate header information is stored in the device) of retrieving individual pixel intensity information. It also should produce a higher quality image. Archival storage, for instance by being able to dump or load the image storage device to or from cassette tape, enables users to pick up again at a later session without being bothered by intervening users; it also allows the user to build up a library of time sequences.

Such a device does assume, of course, that the image display device can be "dual ported"—i.e., that it can, in some way, be loaded from the host as well as from an external device. This requirement is not strictly necessary for the control panel, but does make things easier and faster for it; if the capability is there, it does make sense to integrate the storage device and the control panel.

3.5.4. Video recording. The image display device puts images on a TV monitor—so what would be more logical than to record these images on video tape for later display and/or the generation of time sequences? Unfortunately, things are not that simple.

To generate time sequences, frame by frame, requires a video tape format with frame encoding. The most likely candidate is U-matic 0.75 inch tape, not the most popular VHS 0.5 inch tape format. This is true for any sequence which is not recorded in one shot and where one does want to avoid irritating flicker. To be able to view such recordings on the garden variety VCRs then requires a copying capability to VHS tape.

The signal coming out of display devices is R-G-B-sync (four cables). In order to enable the user to feed this into a VCR one has to turn it into composite video, which is not a big deal, but still a $3500 box.

Our TV monitors display 512 lines. Standard (American NTSC) TV only displays 480 lines. So one will loose the bottom 32 lines, or one-sixteenth of the image. Our European colleagues are better off: they only get a black band underneath the image.

Our TV monitors run in "underscan" mode: to achieve a 1 : 1 aspect ratio, the TV beam only makes a partial horizontal scan. When displayed on a regular TV set, the image will be distorted—stretched in the horizontal direction by a factor 1.33. Our European colleagues are even worse off. Of course, it should be possible to squeeze the image horizontally to achieve a satisfactory aspect ratio, but to my knowledge boxes performing just that service are not available commercially; they are available as part of sophisticated (and expensive) studio equipment that will do much more than we will ever need. Alternatively, the squeezing could be done in software, before the images are loaded, but that is not really satisfactory either, since it violates the rule "what you see is what you get".

As one can see, there are some problems with video recording. They can be solved, but it costs money (about $10,000 to $20,000). Movie making (on 16 mm film) is actually cheaper as far as equipment is concerned. Nevertheless, I think the capability should be

developed. In the first place, it is in principle capable of producing canned displays with a very short turn-around time. Secondly, if we can record display sequences on video tape, the user can be provided with a cheap ($1000) work station to view his data *ad nauseam* without having to hog the image display device itself, its TV monitor, a host terminal, and, possibly, the image storage device and the control panel.

3.5.5. 3-D solids machine. The Medical Imaging Section of the Department of Radiology at the University of Pennsylvania Hospital has recently developed a Voxel Processor. Basically, it is a memory that can hold a cube of intensities with specially configured processors that calculate and load an image of a three-dimensional solid surface through that cube. The user has dynamic control over the viewing parameters (angle, perspective, clipping), as well as the threshold intensity. The current prototype can handle $64 \times 64 \times 64$ cubes and generate frames at 15 Hz. A $256 \times 256 \times 256$ Voxel Processor is under development. Such a device would be extremely helpful for the display and interpretation of three-dimensional data.

4. GRAPHICS DISPLAY

The realm of graphics display is vastly different from that of image display. This is not only expressed in the definition, but also, more importantly, in the commercial availability of products. In graphics we are dealing with geometrical objects (points, line segments, polygons, etc., most often translated into vectors), which may or may not have attributes such as color and intensity. Notwithstanding the vectorial nature of graphics display, much of it is displayed on raster devices today. In image display the data consist of actual intensities, measured on some regular sampling grid in space. Image display is more demanding in sophistication, I/O rates, and memory, while its application is confined to a relatively small number of specialized fields. Graphics, on the other hand, can be dealt with at a much lower level of sophistication and resources (although in the upper ranges it can certainly compete with image display in these respects), and has permeated virtually every facet of society, most importantly the business community. Since business has perceived graphics display as an important tool for increased productivity and effectiveness in nearly every type of work (public relations, promotion, management, production, process control, etc.) there has been a great incentive (as well as the financial resources) to develop an enormous spectrum of commercially available graphics hardware and software, at a reasonable price. This does, of course, mean that much of what is available—actually much of what gets most of the attention—is of little use in astronomy: pie charts, bar charts, over-sophisticated lay-outs, which all seem to have become the trademark of successful business.

The use of commercial software is very rare in astronomy, for various reasons. The astronomical community has traditionally had needs that either exceeded the capabilities of such software or were not covered by it; capital funds were often limited, whereas manpower was available; and the tradition of doing things ourselves has proven very strong. As a result, a situation has developed where it is accepted to spend money on hardware and even on software where it concerns things like operating systems and compilers, but where it is considered a waste to buy other commercial software because "we can do it much cheaper ourselves". However, it may be wise to consider the practice of using bought software by the business community a little more serious than just to assume that they lack the expertise to do it themselves. The reasons are three-fold: commercial software usually provides very flexible software tools at a price that is lower than the manpower cost of in-house production; the (often considerable) burden of maintenance is put on the software supplier; and commercial packages provide hardware independence, with regard to the host as well as the peripherals used.

Graphics software is probably the field where astronomy can benefit most from the use of commercial software packages, for the reasons outlined above. The only experience we have is with DI–3000, a product of Precision Visuals, Inc. This experience shows that commercial packages can provide very powerful, flexible software tools that greatly increase programmer production, are easy to use, have excellent documentation, and provide device independence. Usually with no extra effort a higher quality display is produced than one would otherwise bother to go to, while also flexibility for the user is increased. As far as cost is concerned, it should be kept in mind that software providers often offer significant educational discounts.

4.1. 2–D displays.

There are three types of two-dimensional graphics displays in use in astronomy: contour plots, graphs of various kinds, and interactive graphics in conjunction with (mostly) image display. This last type of display (drawing lines, boxes, polygons, or other types of regions of interest) is never used on its own, but only as part of another graphics or image display.

Similarly, there are three forms of output: non-permanent (on CRT or TV monitor), disposable permanent (i.e., work copies), and publication quality. The purposes and applications are self-evident, although it may be useful to mention the possibility of overlaying contour plots on images. Also, the three output media show a trial-and-error path toward final data display in the order given above. Ideally, the user should be able to get the same piece of graphics on different media by the flip of a switch.

CRT graphics display is rather poorly developed in astronomy. It has never really got beyond the Tektronics 4012. Attention should be given to the capabilities of more modern graphics terminals and the use of color.

In the field of work-copy devices, for a long time the choice has been Versatec-like devices—which are extremely messy. Laser printers are beginning to replace them; however, they have one disadvantage over the Versatecs: the size of the plot has a firm and rather small upper limit.

Publication quality, although attainable on a Versatec, is still most easily produced on a good old-fashioned pen plotter. With the advent of plotters with built-in microprocessor these devices have become faster and more flexible, taking much of a burden off the host. The improvements include publication quality character fonts. Great assets are the availability of different color pens (indispensable for the contouring of velocity fields; facilitating contouring of complicated regions; allowing overlaying of different plots) and pens with different thicknesses (for publication). One has to be aware, though, that pen plotters, by their nature, require a slightly different style of programming (see Sec. 4.3).

There are a few special needs for the contour plots that we usually deal with. One is that the contouring algorithm has to be able to deal with undefined points in an otherwise regular grid (for velocity fields, optical depths, dispersions, spectral index, etc.). Many algorithms either do not allow for this or are based on randomly sampled data; neither is acceptable. Another is that arbitrary sizes must be allowed; if necessary, it should be possible to assemble large plots from long strips. Yet another need is the production of contour plot mosaics for spectral line observations. Finally, we need the capability to overlay other information on the plots: star positions, polarization vectors.

Ruled surface plots deserve special mention. There are two applications: to display two-dimensional data (e.g., an image) as a stack of one-dimensional cross-sectional profiles, which is especially sensitive to gradients, instead of or in addition to a contour plot; and to display a collection of truly one-dimensional profiles (e.g., amplitude as a function of time for a number of baselines), which is very sensitive to "odd" points.

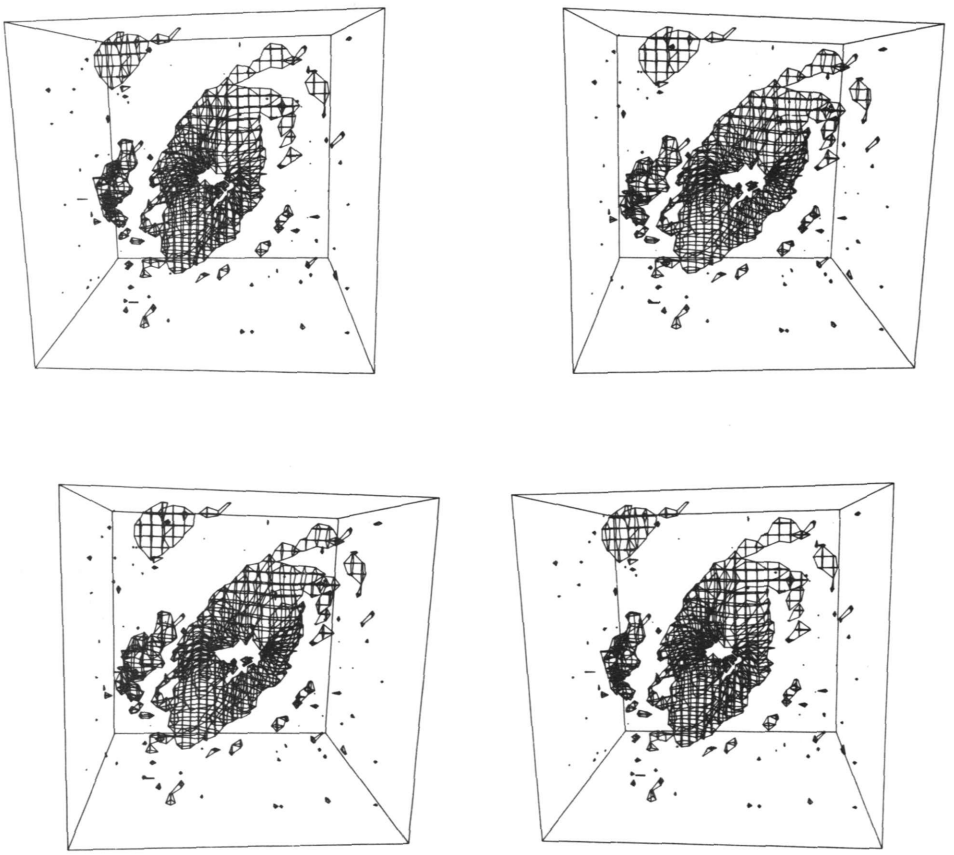


Figure 15–4. A 3–D wire frame contour representation of HI observations of M81. This is the exact graphics equivalent of Fig. 15–3; the arrangement is identical.

There is one more type of graphics that needs attention: plotting a number of related parameters on a two-dimensional grid. An example of this is the display of the parameters of multiple Gaussian components to spectral profiles. Ulrich Schwarz has developed a satisfactory scheme for this case.

4.2. 3–D displays.

There are two applications for three-dimensional graphics displays in radio astronomy: three-dimensional contour plots of spectral line data, and graphic display of visibility data. The former constitutes a substitute for as well as an extension of the three-dimensional solids discussed in Section 3.4.2 and needs no further clarification; an example is shown in Figure 15–4. Graphic display of visibility data forms an extension of the image displays discussed in Section 3.1, as well as of the graphs mentioned in the previous Section, and will be expanded upon below.

Three-dimensional graphics devices have been commercially available for a long time; Evans and Sutherland are probably the best known manufacturers. These devices allow real-time rotation, translation, hither- and yon-clipping, depth cueing, perspective, and zooming. Especially the motion cues introduced by real-time rotation are very powerful in conveying three-dimensional structural information. In addition, they will allow some form of stereoscopic display in conjunction with the other operations. Polygon fill has become available since manufacturers started switching to raster monitors and may, to

some extent, emulate 3–D solid surfaces. Because of all these features and capabilities, the three-dimensional graphics devices are very powerful tools in exploring multi-dimensional data. Traditionally, these machines have been very expensive (well above $50,000), but in recent years much cheaper devices doing essentially the same thing have come on the market, like the IBM 5080.

For the graphics display of visibility data, the 3–D graphics devices facilitate displaying amplitude or phase along the z-axis, while u and v lie along the x- and y-axes, as well as representing complex visibility as a true vector as a function of either baseline and time or u and v. It is true, of course, that any view of such a representation can be displayed on a 2–D device, but the dynamic changing of the viewing angle greatly facilitates quick understanding of trends and patterns (e.g., the twisting of phase). Jim Torson made a demonstration program that simulates the effect on the VLA PDP–11/40 with the aid of an array processor. A similar display mode, lacking the dynamic capabilities, is in use for MERLIN data and supports the usefulness of these representations.

4.3. Device independence.

The great variety of output devices has been a blessing and a curse in the graphics world. It has led to a large degree of flexibility in display choices, but at the same time to an even greater inflexibility in application of software, due to large amounts of device dependent code. With the proliferation of graphics and graphics software into every aspect of modern life (including those areas where there are large amounts of money), however, the impetus to standardize and introduce a large degree of device independence has become strong enough to achieve some results. The standards battle between CORE and GKS has been decided in favor of the latter. Three-dimensionality is still lacking in GKS, but will probably be brought in through the PHIGS extension. GKS also has, in principle, some image display elements, but it would, in my opinion, be unwise to rely on those; they are currently inadequate and it is doubtful whether they would ever become satisfactory.

The aim of the graphics standard is to achieve device independence. By defining the standard, the interface to the output graphics device has been defined; therefore, once a driver has been written for a particular device that adheres to this standard, the device is available to all software that complies with the same standard. The use of such a graphics standard *per se* is not that easy. However, by using a graphics package that interfaces with the standard, one can achieve considerable gains:

- The same code can be used for a number of output devices, i.e., the user can, in real time, choose the device that is most appropriate for his current needs (e.g., CRT or hard-copy).
- The quality of the display can be matched to the device and purpose of the output (e.g., quick-and-dirty or publication quality).
- Increased programmer productivity (this depends on the quality of the package used).
- Device independence from a management point of view: decisions on acquisition or replacement of graphics display devices do not have to be made anymore taking into consideration the match between a particular device and existing code. Also, mixing smart and dumb devices is not a consideration anymore since good device drivers will take advantage of all the capabilities of particular devices and simulate the rest in software.

The use of a commercial graphics package (discussed in the introduction to Sec. 4) facilitates the ease of programming and ensures the availability of device drivers adhering

to the standard; for the time being, device manufacturers cannot be relied on to provide the latter.

All in all, device independence is a wonderful thing. However, having said that, we also have to sound a warning. Device independence does introduce inherent inefficiencies: in writing code for a particular device one can take the peculiarities of the device into account and tailor the software to run as efficiently as possible. In using device independent code one has to pay in decreased performance as far as efficiency is concerned. In general, the advantages will outweigh the disadvantages, but in certain cases the result will be unsatisfactory. For instance, in writing a contouring program one has the choice between scanning the image line-by-line and following the contours. The former requires less memory, is simpler, and will run faster. In addition, it is the natural way to do it for dot-printing devices. But it is totally unacceptable for pen plotters: not only will it result in very slow plotting with a lot of unnecessary pen movements, but the appearance of the plot will be very bad as well, because of all the little discontinuities where individual contour segments join. Use of the latter technique is mandatory for pen plotters. Therefore, some considerations of the characteristics of the device actually used will still have to enter into the code.

5. Working Environment and Support Functions

It goes without saying that in order to take full advantage of the potential that data display techniques hold, the work environment and support functions offered to the users must be thought out rather carefully. One can have the most sophisticated display software, but it is not going to profit the user if the TV monitors are mediocre, or if they have to be viewed under cramped or otherwise uncomfortable conditions, or if there is no hard copy available, or if the most beautiful images can be recorded but there is no way to process the film; one can add to this list *ad libitum*.

5.1. Work stations.

There are three requirements for organizing work stations: they should be pleasant to work at, there should be a spectrum of different types of work stations tailored to users' needs, and there should be enough of them.

The first requirement involves such things as: enough space to move around, pleasant temperature, low noise level, enough desk space, good lighting, and general ergonometric considerations. All controls should be in easy reach and placed at the right height and angle. Chairs should be comfortable and not interfere with the work. TV monitors should be placed at the right distance, height, and angle. Lighting should be adequate for desk and monitor viewing, and be adjustable.

Not all work stations have to be the same. Efficient use of resources dictates that the user can choose a work station that fits his needs for a particular session, but no more than those needs. A proper mix should be established of very sophisticated image display work stations with adequate graphics facilities, sophisticated graphics display work stations with simple image display facilities, and simple image display work stations (on-line as well as off-line, possibly just viewing stations). In addition, regular alphanumeric terminals (e.g., for preparation of batch oriented jobs) should be available. In general, rather simple hard copy output will be adequate for the work stations. High quality hard copy devices (plotter, image recorder) should be present, but at a central location; not only does one need only a few (or just one) of each, but operations usually run smoother when they are under the care of a central operator.

Finally, there should be enough of these work stations. This implies that a proper mix, as outlined in the previous paragraph, is maintained. It also means minimizing the

frustration level of the users. Hence, there should be enough stations to keep all users at a given time busy at an acceptable level, but not so many that the systems become clogged and responses sluggish.

5.2. Image recording facilities.

There are probably three types of needs for high quality recording of images: slides for presentations, pictures for publications, and public relations work (posters, etc.). All three warrant the use of a high precision image recorder. The requirements are: recording areas of at least 4096×4096 pixels, high geometric accuracy, high linearity in intensity for at least 8 bits, good color rendition, and flexibility in film size uses. Experience has shown that the different needs have slightly different requirements for the type of film used and that the usage of the recording facility depends on how easy it is for a given user to get his slide or print made and how quickly he can get it. For these reasons the facility should be able to routinely handle 35 mm roll film and 4×5 inch sheet film, in color slide film, B/W negative, B/W reversal, and color negative film with no more than 24 hours turn-around time; in addition, a 16 mm film capability would be very desirable.

Obviously, there are other applications for image recorders. One might conceive of creating an alternative output path for graphics display through them. As a matter of fact, some graphics packages can provide a driver for a device like the Dicomed film recorder.

5.3. Photo/graphics facilities.

As indicated in the previous Section already, fast turn-around for images recorded on film is indispensable. This not only applies to the processing of the film but also to the capability of producing high quality prints. The same is true for the handling of high quality graphics work. The use of a film recorder for graphics display has been mentioned. Also needed are a high precision pen plotter and drafting services, as well as the photographic facilities to process their results.

In addition, astronomy requires some specialized photographic facilities: reproduction of optical material and the production of overlays.

Altogether, this calls for a well-equipped and well-manned photolab that can take care of all photo processing needs on time scales of one day to one week and a drafting/graphics department. Part of the photo work would be the routine processing of computer-generated outputs.

6. Conclusion and Recommendations

We conclude that judicious use of sophisticated data display hardware and techniques can make life a lot easier for the user, and at the same time unburden the computing resources. This is achieved by providing interactive display and analysis functions that are aimed at providing the user with powerful tools for a better understanding of the contents of his data, by moving the execution of those functions to the display devices. By doing this one not only shifts the compute power needed for the experimentation out of the host computers, but also enables the user to make well-founded decisions on how to proceed with the data processing, resulting in less unnecessary processing and a faster reduction process for the user. It does require, however, an investment in display hardware, software, and research.

As for the actual image display devices, we recommend that machines be used with capabilities comparable to those of the IIS Models 70 and 75, but preferably with 1024×1024 pixel resolution, and with adequate amounts of refresh memory. The capabilities should at least include multiple image refresh memories with 8 bits each, the double lookup table architecture, 16 bit mode, and ALU/Feedback unit.

A variety of graphics devices should be available. Each work station should have a graphics CRT and access to a high quality hard copy device. One work station with a sophisticated three-dimensional graphics device would be extremely useful. To make usage of the various devices transparent to the user and the programmer, as well as to provide flexibility in management, it is strongly recommended that a good graphics package be adopted that adheres to an international graphics standard and that provides drivers for as large a number of devices as possible.

Display devices can very profitably be supplemented by special purpose hardware, but only in those cases where the desired features cannot be achieved in any other way.

Hard copy capability is very important. Each work station should have access to instant hard copy devices, for images and graphics. A central facility should provide high quality hard copy for both, with good procedures and support to make these available to the user in a reasonable time.

Appropriate attention should be paid to the working environment and the ergonomics of the work stations. This includes designing controls that make using the displays natural and easy for the user. There should be a proper variety of types of work stations, reflecting the varying needs of different users as well as of the various stages of data reduction and analysis.

Display of three-dimensional images can be made far more effective than the current ones are. A strong research effort in this direction could yield considerable gains.

Finally, no system is better than its support. One is not going to improve matters without an adequate level of and balance between the three support branches: software, hardware, and services. The first two have been discussed above. The third determines whether the user can actually achieve lasting use of the results of his toils; it is hard to overspecify photo/graphics services for a facility where image display plays a central role.

16. VLA Observing Strategies

ALAN H. BRIDLE

1. INTRODUCTION

This Lecture discusses the choice of parameters for VLA continuum observing based on a mixture of astronomical and instrumental criteria. It suggests an orderly way in which to use the material of Lectures 2, 4, 5, 6, 7, 8, and 9 to choose critical parameters when planning and executing VLA observations. It also suggests strategies for avoiding some of the pathological image defects that were emphasized in previous lectures. Unlike most of the other lectures in this series, this one is explicitly oriented toward specifics of VLA continuum observing, though the general principles apply to observations made with other synthesis arrays.

Figure 16–1 shows a decision tree for preparing VLA continuum observations; Sections 2 to 6 of this Lecture detail the various levels of this tree. 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.

Sections 7 to 9 of this Lecture discuss calibration strategy, on-line observing strategy, and the observing proposal itself.

2. CHOICE OF ARRAY CONFIGURATION AND OBSERVING FREQUENCY

2.1. Resolution $\theta_{\rm HPBW}$—How much is enough?

An image made from untapered uniformly-weighted ≥ 4 hour tracks in a standard VLA configuration at positive declinations where foreshortening of the array is unimportant has a synthesized beam B with a half-power beamwidth given approximately by

$$\theta_{\rm HPBW} = 1\farcs25 \times \frac{1480}{\nu_0} \times 3.285^{n-1}, \qquad (16\text{–}1)$$

where ν_0 is the observing frequency in MHz and $n = 1, 2, 3$, or 4 for the **A**, **B**, **C**, or **D** configurations respectively.

The *minimum* resolution (i.e., maximum value of $\theta_{\rm HPBW}$) appropriate for the observations will be determined by the need to separate or resolve important features of the structure in the region to be imaged. For observations of extended emission, the *maximum* resolution (minimum $\theta_{\rm HPBW}$) that is appropriate should also be considered, by estimating the total integration time $t_{\rm int}$ needed to achieve the required brightness sensitivity. There is no point observing extended emission using such a small beamwidth $\theta_{\rm HPBW}$ that the interesting features of the source are close to or below the r.m.s. noise ΔI_m on the final images. To make sure that this does not happen, you must consider the *apparent brightness* (flux density per synthesized beam area) that you expect such features to have at the resolution you will use for your final images.

Recall from Lecture 6 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

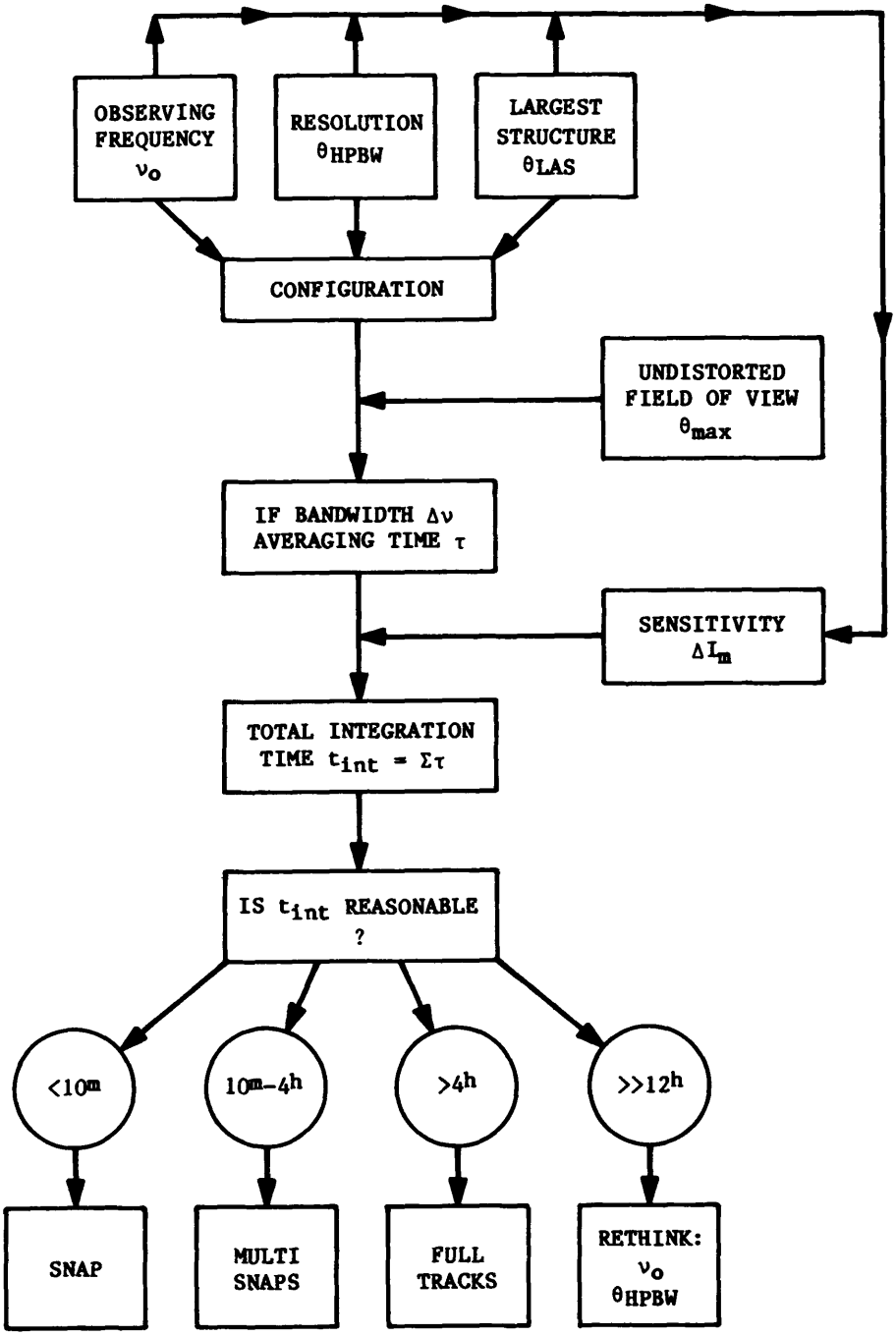


Figure 16–1. Factors Entering Into VLA Observing Strategy—A Suggested Decision Tree.

synthesized beam. It follows that, at a given frequency, all VLA configurations are equally sensitive to a given point source (apart from the effects of confusion and phase stability). In contrast, as described in Lecture 6, 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 at the imaging stage (Lecture 5, Section 2.2; and Lecture

6). When deciding on an observing strategy, it is usually sufficient however 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 in square arcsec can be calculated approximately as $\Omega_s \approx 1.13\theta_1\theta_2$ 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. Ensure that you do not observe with 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).

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 VLA 20 cm image made with the **B** configuration (resolution $\approx 4\farcs2$). It 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 $\approx 1\farcs3$). 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 16–1), but a 10σ detection in the **A** configuration using the same bandwidth would require about 31 hours of on-source integration! When studying extended emission, it is therefore *extremely* important not to use a configuration giving a smaller beam area Ω_s than is strictly necessary.

Note also that the effects of spectral index and resolution combine to make extended *steep-spectrum* emission much harder to detect in 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 **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 target without self-calibration.

You should also avoid unnecessarily high resolution in detection experiments at high frequencies. While the theoretical sensitivity to a point source is independent of the array configuration (apart from the effects of confusion), the phase stability, and hence the ability to integrate coherently between calibrations, will be poorer on longer baselines (see Lecture 4, Section 4.4). The phase stability will be highly dependent on the state of atmosphere over the array (the "weather"), so one cannot predict the severity of this effect in advance—but it is clear, for example, that the **A** configuration is rarely a wise choice for 1.3 cm point source detection experiments.

There are circumstances however when enhanced resolution improves the ability to detect interesting features in a source—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.

These competing factors affecting the choice of resolution cannot be estimated reliably in advance if the source structure is unknown or poorly known. If you are not sure what to expect your source to look like, the safest strategy is to guess on the side of low resolution in an initial observation. A preliminary low resolution image may tell you the source's total angular extent and could also warn you of any surrounding emission. This information would allow you to optimize the observing parameters for a more time-consuming high resolution study. It is also easier to justify reobserving a detected emission region at higher resolution than it is to justify reobserving at lower resolution what appeared to be empty sky!

2.2. Choice of frequency ν_0 at given resolution $\theta_{\rm HPBW}$.

The choice of observing frequency *at a given resolution* will be determined by astronomical criteria. A high frequency might be chosen 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. The spectral index of the emission being studied also influences the choice—optically thick thermal emission may be easier to detect at 2 cm than 6 cm despite the noisier system at 2 cm, whereas transparent synchrotron sources will be easiest to detect at a given resolution at 20 cm.

Returning to Equation 16–1, 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 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. "Scaled-configuration" VLA observations can be used to produce maps of spectral index, Faraday rotation or depolarization properties of extended sources that are relatively free from uncertainties stemming from differing resolutions at the different frequencies.

Note that use of the "scaled configurations" *optimizes your chances* of measuring frequency-dependent properties of a source accurately, but does not by itself *guarantee* success. Further careful planning, and *post hoc* examination of the visibility data, are also important. For example, the hour-angle ranges of "scaled-configuration" observations should be matched at the different frequencies. Also, even scaled configurations may sample parts of the visibility function of a source with differing sensitivities at different frequencies if the source structure changes radically over the frequency range of interest. This may happen if there are large spectral index gradients across the source in either its total or its polarized emission. Care must also be exercised when interpreting the final images if the databases at the two frequencies are differently affected by missing antennas or by bad data. In such cases, the reliability of inter-frequency comparisons may still depend on how well the deconvolution algorithm (Lecture 7) can interpolate in the u-v plane.

Finally, do not forget that the VLA continuum system allows you to 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 8, Section 1.1) or to study spectral or Faraday depth changes in your source across a "band" (the latter being especially worthwhile in practice at the VLA's L Band—1340 to 1730 MHz).

2.3. More than one configuration?

The above was concerned primarily with observations in the standard (A, B, C, D) configurations of the VLA, but other options are available. You may need to combine observations made in more than one VLA configuration if your observations require a *range* of baselines that exceeds the range provided by a standard configuration. The next step in planning your observations therefore involves thinking about $\theta_{\rm LAS}$, the largest angular scale of structure that you must sample well to produce an astrophysically useful final image. $\theta_{\rm LAS}$ will be the angular diameter of the most extended structure that you need to reconstruct accurately in the final image—usually the diameter of the most extended component of astrophysical interest in your source. (Do not confuse it with $\theta_{\rm max}$, the required field of view, which is discussed below—when observing a source $10''$ in extent in the presence of a point confusing source $1'$ away, you would set $\theta_{\rm LAS} = 10''$, not $\theta_{\rm LAS} = 1'$.)

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 $\theta_{\rm LAS} \approx 40\theta_{\rm HPBW}$ where $\theta_{\rm HPBW}$ is given by Equation 16–1 at the specified frequency. If the values of $\theta_{\rm LAS}$ and $\theta_{\rm HPBW}$ needed for your experiment do not *both* fall between $\theta_{\rm HPBW}$ and $40\theta_{\rm HPBW}$ calculated from Equation 16–1 for a given standard configuration and frequency, you should consider taking data in more than one VLA configuration. Obviously, any observation requiring $\theta_{\rm LAS}/\theta_{\rm HPBW} > 40:1$ falls in this category, but so do some with $\theta_{\rm LAS}/\theta_{\rm HPBW} < 40:1$; for example, your optimum $\theta_{\rm HPBW}$ might fall mid-way between two resolutions given by allowed values of n and ν_0 in Equation 16–1.

For example, Figures 16–2 and 16–3 show 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 16–2 is well filled by mixing data from the A and C configurations. You should consider mixing standard-configuration observations for any sources for which $\theta_{\rm LAS}/\theta_{\rm HPBW}$ will be significantly $> 40:1$. The total integration times to be spent observing in the different configurations should however be computed separately, as in Section 4 below; for most projects you will not need as long a total integration time in the more compact configurations as you will in the more scattered ones.

2.4 Hybrid configurations.

"Hybrid" configurations are those that become available during reconfiguration periods, when the arms of the VLA may be of different length, or may 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 (thus giving sensitivity to a wider range of angular scales). Some can assist self-calibration of data from a compact configuration by providing it with some unusually long spacings.

Hybrid configurations with long North arms are now regularly scheduled at the VLA. They are useful if you want to image 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 16–4 shows the u-v coverage for the B configuration at −40° declination, compared with that of a hybrid configuration in which the East and West arms are in the B configuration while the North arm is in the A configuration. The spacings obtained from 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 would be available for a brief period about every sixteen months, during a reconfiguration from A to B. The other such hybrids (B/C and C/D) are also 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.

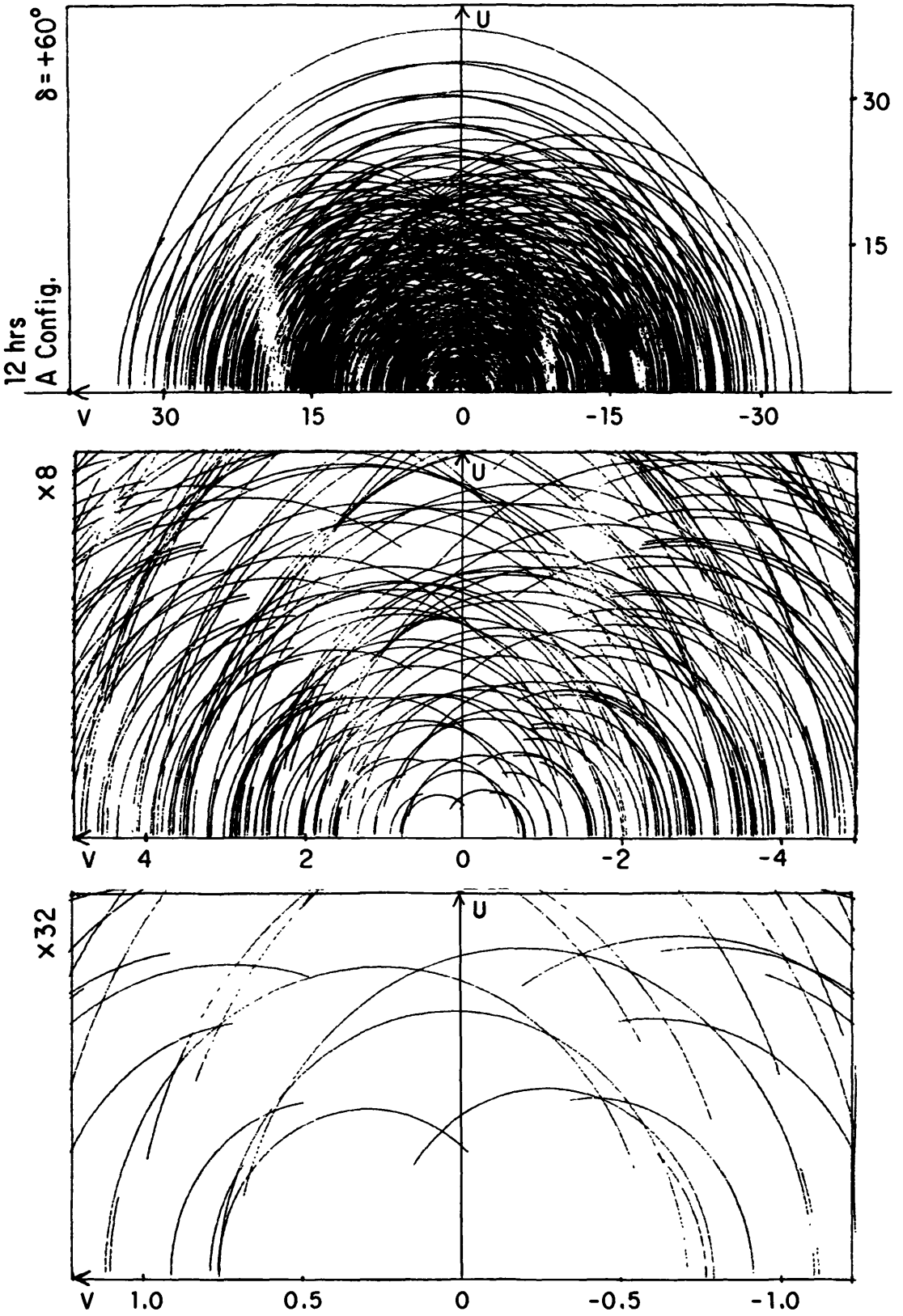


Figure 16–2. *u-v* coverage for $\delta = +60°$ in the A configuration (12-hour tracks).

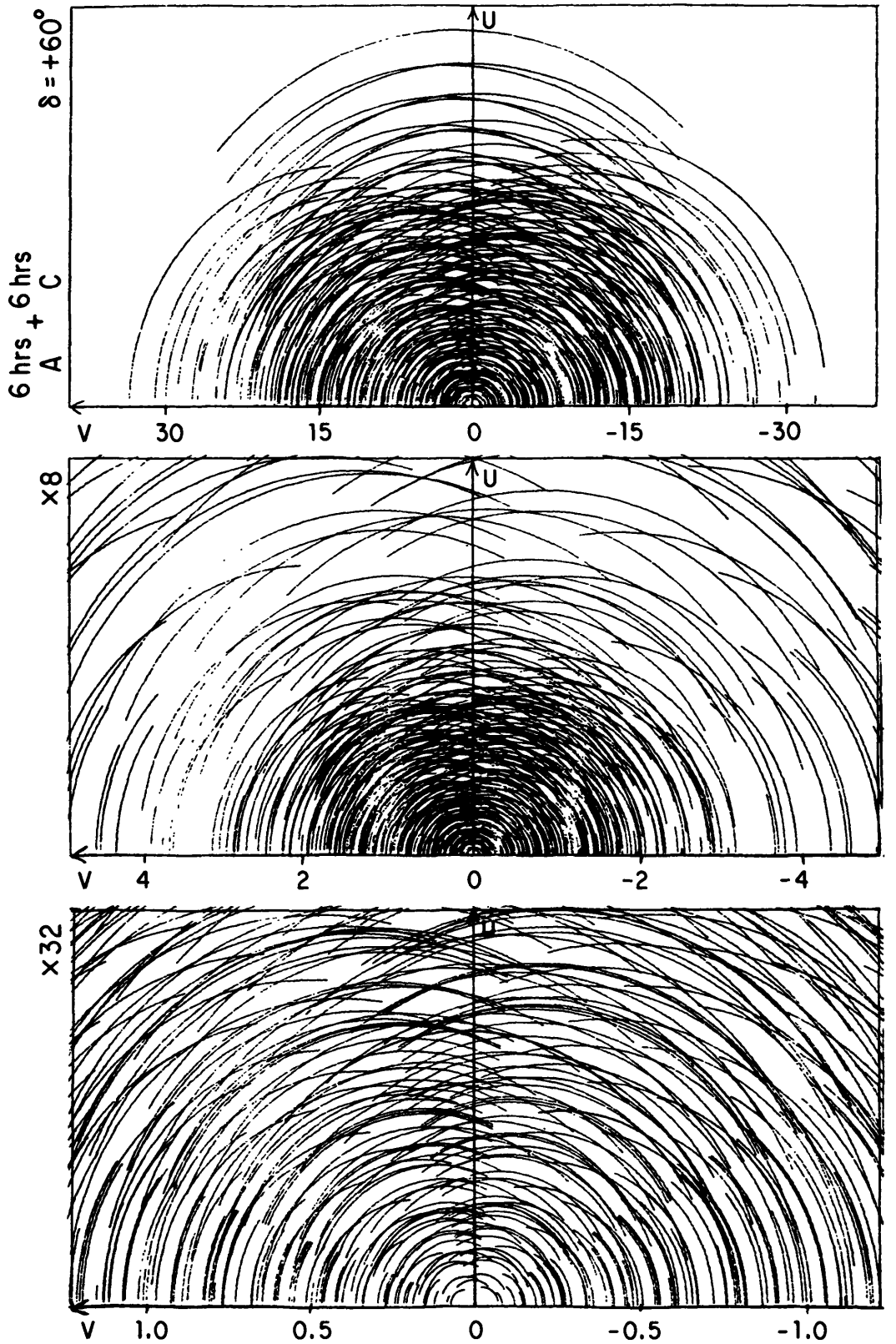


Figure 16–3. *u-v* coverage obtained by combining 6 hours of A configuration data with 6 hours of C configuration data at $\delta = +60°$. Note the superior coverage of the inner *u-v* plane, relative to Fig. 16–2.

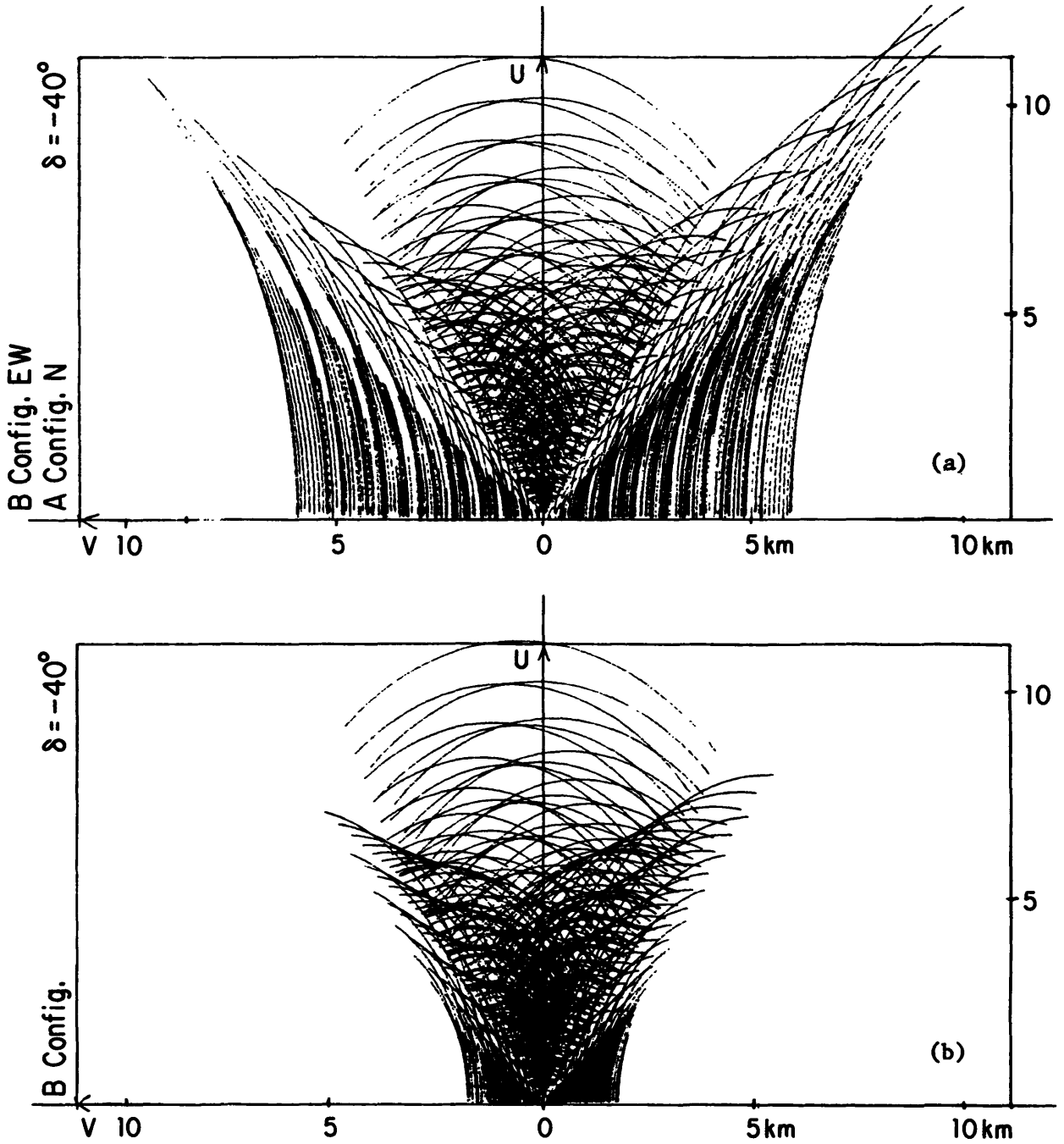


Figure 16–4. *u-v* coverage at $\delta = -40°$ with (a) *(top)* the VLA East and West arms in B configuration and the North arm in A configuration, and (b) *(bottom)* the entire VLA in B configuration.

In general, you get better *u-v* coverage by mixing data from two different standard configurations than you do from the same total time spent in *any* hybrid configuration, so no other hybrid configurations are regularly scheduled.

2.5. Sub-arrays.

"Sub-arrays" are nonstandard configurations obtained by dividing the VLA into as many as three smaller arrays that are then 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,

however. 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 will manifest itself in poorer sensitivity and u-v sampling in the sub-array data. The use of sub-arrays is therefore generally undesirable unless your program calls for *strictly* simultaneous observations of strong sources at several frequencies (e.g., instantaneous spectra of rapid variables) or for observations of a large number of compact sources with only modest demands on sensitivity and dynamic range in each image (e.g., astrometry of strong sources).

2.6. Interference and the detailed choice of frequency ν_0.

External interfering signals are partially rejected by interferometers 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 (strong interference may also degrade the noise performance). This rejection is better at the longer baselines, so the VLA's **A** and **B** configurations are less susceptible to external interfering signals than are its **C** and **D** configurations. (Delay rejection is not usually significant for narrow-band interfering signals).

Interference is rarely detected or suspected at C, U or K Bands ("6 cm", "2 cm" or "1.3 cm"). It is however a factor in choosing a continuum observing frequency within the VLA L Band (1340 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)[1]. Frequency allocations in the L band include aeronautical radio navigation, 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 anywhere at L Band outside the protected radio astronomy bands. (Note that use of the protected band at 1400 to 1427 MHz may itself be undesirable for some continuum observations, owing to the contribution of galactic neutral hydrogen line emission to the system temperature in this band).

There is also 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 may be a serious problem for spectral-line programs.

Before using a non-standard L Band frequency, consult with VLA scientific staff (particularly Pat Crane, the VLA frequency co-ordinator) for advice and lore based on recent observers' experiences.

3. Field of View Restrictions

Once you have settled on the resolution $\theta_{\rm HPBW}$ and observing frequency ν_0 for your program, the next level on the decision tree (Fig. 16–1) is the choice of IF bandwidth $\Delta\nu$ and averaging time τ_a. These must be made consistent with the field of view requirements of

[1]Spectral line observers do not, of course, have the same freedom to choose the center frequencies and bandwidths for their projects, so L band interference may determine whether a given spectral line experiment is possible.

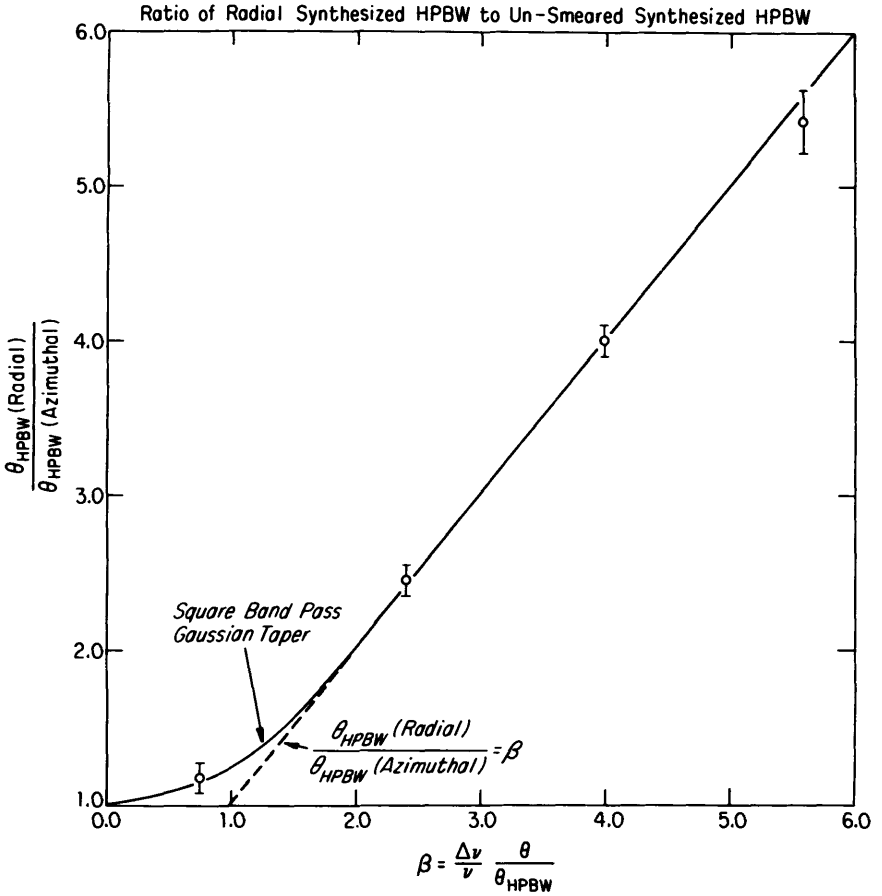


Figure 16–5. The ratio of radial to azimuthal beamwidth, resulting from finite IF bandwidth $\Delta\nu$, plotted as a function of the dimensionless parameter β. θ is the angular distance of the feature from the phase center, in the same units as the beamwidth $\theta_{\mathbf{HPBW}}$.

your program. The next step is therefore to consider the radius $\theta_{\mathbf{max}}$ (from the center of the field of view) over which you require the data to be minimally distorted by the bandwidth smearing and time-average smearing effects discussed in Lectures 2 and 8.

3.1. IF bandwidth $\Delta\nu$.

The choice of the IF bandwidth for VLA continuum observations is most important, as an unsuitable choice may lead (a) to irrecoverable distortion of the image if the bandwidth is too great, or (b) to loss of sensitivity if it is too small. As discussed in Lectures 2 and 8, observations made with finite bandwidth suffer both radial smearing and reduction in amplitude of the point source response away from the delay tracking center. These effects are discussed in detail by Perley (1981a), and their magnitudes are also graphed in Figures 16–5 and 16–6.

The first step in choosing the IF bandwidth for your observations is to ask over what field radius $\theta_{\mathbf{max}}$ (arcsec) you require either the radial smearing to be less than $n\%$ or the reduction in amplitude of a point source to be less than $m\%$, due to finite IF bandwidth. Then enter Figure 16–5 at ordinate $1 + n/100$, or Figure 16–6 at ordinate $1 - m/100$, and read the corresponding value of the normalized parameter β from the abscissa. Call this value $\beta_{\mathbf{max}}$. Then compute the maximum allowable IF bandwidth $\Delta\nu_{\mathbf{max}}$ (MHz) consistent with these constraints from the relation

$$\Delta\nu_{\mathbf{max}} = \frac{\beta_{\mathbf{max}}\nu_0\theta_{\mathbf{HPBW}}}{\theta_{\mathbf{max}}}, \qquad (16\text{–}2)$$

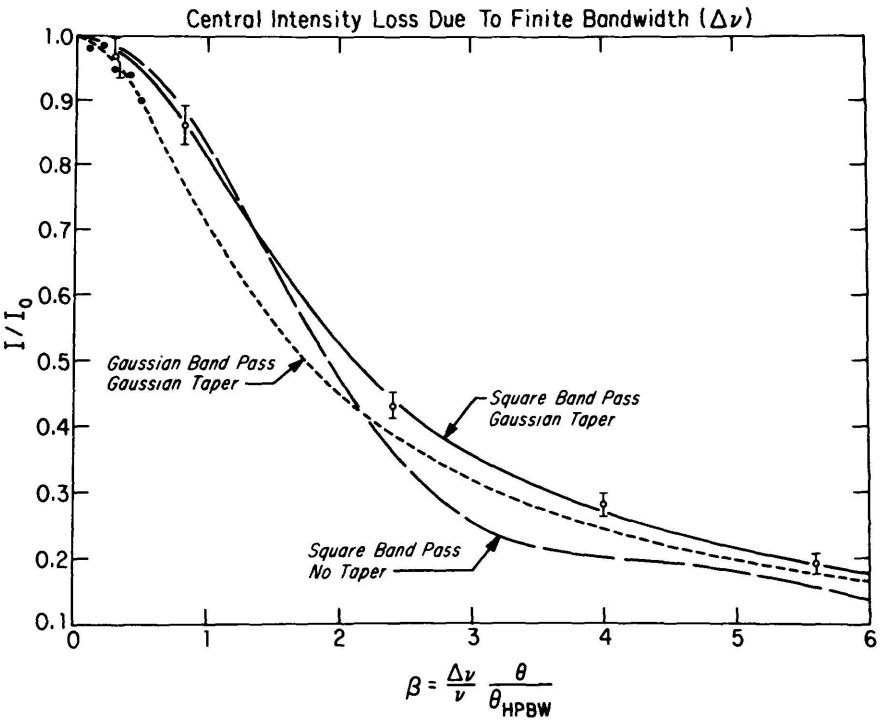


Figure 16–6. The central intensity loss, due to finite IF bandwidth $\Delta\nu$, plotted as a function of the dimensionless parameter β. θ is the angular distance of the feature from the phase center, in the same units as the beamwidth $\theta_{\rm HPBW}$.

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 VLA bandwidth that is *narrower* than the computed value $\Delta\nu_{\rm max}$. If you are prepared to relax it, choose the closest *wider* bandwidth.

For example, suppose you are prepared to tolerate an amplitude loss of 10% for a point source at 45″ from the image center in an **A** configuration observation at 1465 MHz. Entering Figure 16–6 at $I/I_0 = 0.9$ gives $\beta_{\rm max} = 0.8$, from which $\Delta\nu_{\rm max} = 0.8 \times 1465 \times 1''.25/45'' = 32$ MHz. You would then either choose $\Delta\nu = 25$ MHz, or relax the criterion and use $\Delta\nu = 50$ MHz.

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 'CLEAN' a confusing source's sidelobes from the region of interest. 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 configuration. 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" source 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 image center whenever high dynamic range is required. This strategy will minimize the total distortion of the image resulting from bandwidth, pointing, averaging time and u-v truncation effects involving the strong source (see Clark 1981).

For point source detection experiments the above criteria will normally select the 50

MHz bandwidth, unless the search position is exceptionally inaccurate or the field is known to be highly confused. The 50 MHz bandwidth is also normally required at 2 cm and 1.3 cm, because at these wavelengths the usable field of view is limited by the primary HPBW of the antennas for the narrower bandwidths, and because the system temperatures are greater than at 20 cm or 6 cm.

When deciding on the value of $\theta_{\max}$ that is appropriate for an image of an extended source, also 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.

Users of extremely narrow bandwidths should note 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. Note however that the VLA spectral-line system does not support polarimetry at present.

Spectral-line observers will normally choose their IF bandwidth from constraints other than those discussed above. For spectral-line imaging, bandwidth smearing is determined by the *channel* bandwidth, which will normally be set (to a small value) by determining the velocity resolution needed for the project, rather than by field of view requirements.

3.2. Visibility averaging time τ_a.

The choice of the visibility averaging time τ_a for VLA observations is less critical than the choice of IF bandwidth $\Delta\nu$, because the default 10-sec averaged visibilities (A and B configurations) and 30-sec averaged visibilities (C and D configurations) are preserved on the archive tape created by the VLA on-line system. If you change your mind about visibility averaging times, the off-line data base can be "refilled" from the archive tape with a changed value of τ_a. This is costly in CPU cycles, however, so should be avoided by choosing τ_a carefully when the off-line data base is first created.

The effects of finite averaging time τ_a were discussed in Lecture 2 (Section 11) and in Lecture 8 (Section 1.2). As τ_a is increased, phase winding of a feature at radius θ from the phase center causes both a smearing of the synthesized beam and a loss of the averaged intensity for a point source. The effect is worst on a given baseline when the feature is moving perpendicularly to the fringes produced by that interferometer and is zero when the feature is moving parallel to the fringes. The magnitude of the effect therefore depends on hour angle and declination, as noted in Lecture 2. For a point source at the north celestial pole however, the average reduction in amplitude $R_A = I/I_0$ varies as

$$\frac{I}{I_0} = 1 - \left(\frac{\pi\tau_a\omega_e\theta}{6\theta_{\mathrm{HPBW}}}\right)^2, \tag{16–3}$$

where ω_e is the angular velocity of the Earth's rotation, 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.

For the case of a square bandpass and Gaussian tapering in the u-v plane, which is closest to the VLA case, and in the regime $(0 < \beta \leq 1)$ where the amplitude reduction produced by bandwidth smearing $R_B = I/I_0 < 0.8$, the expression for bandwidth smearing (e.g., Lecture 8, Section 1.1) can be approximated by

$$\frac{I}{I_0} \approx 1 - \frac{\beta^2}{5} = 1 - \frac{1}{5}\left(\frac{\Delta\nu\theta}{\nu_0\theta_{\mathrm{HPBW}}}\right)^2. \tag{16–4}$$

The averaging time $\tau_{\Delta\nu}$ that produces the *same* intensity reduction for a source near the pole as does an IF bandwidth $\Delta\nu$ can therefore be approximated (for small intensity reductions) by

$$\tau_{\Delta\nu} \approx \frac{6\Delta\nu}{\sqrt{5}\pi\omega_e\nu_0} = 1.2 \times 10^4 \frac{\Delta\nu}{\nu_0} \text{ sec}. \qquad (16\text{–}5)$$

Equation 16–5 gives a reasonable criterion for the *maximum* averaging time τ_a which should be used with a given IF bandwidth $\Delta\nu$ at observing frequency ν_0. Notice that $\tau_{\Delta\nu}$ in Equation 16–5 does not depend on VLA configuration or on $\theta_{\max}$, owing to the first-order similarities between the bandwidth and time average smearing effects.

Note that you may often have to exceed the value of $\tau_{\Delta\nu}$ calculated from Equation 16–5 because the shortest available averaging time is the 1.67 seconds (two IFs), or 6.67 seconds (four IFs) set by the VLA's on-line computers. Also, note that the 'FILLER' program used to transport VLA data from the on-line computers to the off-line system requires the *same* averaging time for the source and calibrator observations. If the calibrator observations are only a few minutes in duration (as is often the case at the lower frequencies), averaging times longer than 30 seconds may be undesirable simply because they permit only crude editing of the calibrator data.

4. Total Integration Time t_{int}

Once you have determined the IF bandwidth $\Delta\nu$ from the field of view criteria, the next step in the decision tree (Fig. 16–1) is to estimate the total on-source integration time t_{int} required for given sensitivity on your final image[1]. Here you will use the expression for the r.m.s. noise ΔI_m on an image made with an N-antenna array:

$$\Delta I_m = F_w \Delta S \Big/ \sqrt{\frac{nN(N-1)}{2} \frac{t_{\text{int}}}{10} \frac{\Delta\nu}{46}}\ , \qquad (16\text{–}6)$$

where n is the number of independent IFs contributed to the image per antenna ($n = 2$ for images of Stokes I from both left and right circular polarized channels at one sky frequency, or for images of $P = \sqrt{Q^2 + U^2}$ at one sky frequency), t_{int} is in seconds, and $\Delta\nu$ is in MHz. In the numerator, $F_w = 1.0$ for natural weighting and ~ 1.5 for uniform weighting (see Lecture 6 for more details), while ΔS is the VLA single-interferometer sensitivity given in Table 6–3 of Lecture 6, namely 73 mJy at 92 cm, 28 mJy at 20 cm, 18 mJy at 6 cm, 52 mJy at 2 cm, and 180 mJy at 1.3 cm.

Table 16–1 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 and of more complete syntheses. (Interference

[1]Spectral-line observers should make this calculation for their channel images setting $\Delta\nu$ equal to the channel bandwidth.

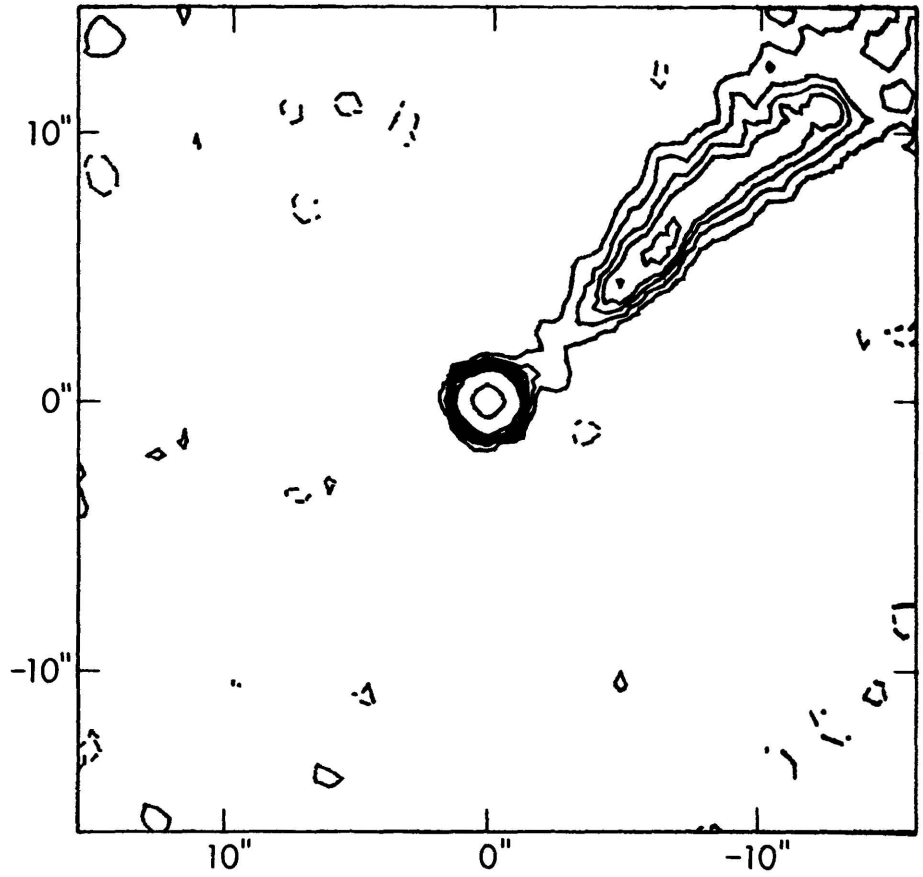


Figure 16–7(a). Contour plot of a 20 cm A configuration snapshot of the source 0055+300, made from 3 minutes of data at 50 MHz bandwidth. The contour levels are drawn at −2, 2, 4, 6, 8, 12, 20, 30, and 200 mJy/beam. The contour around the peak shows the HPBW. Compare with Fig. 16–7(b).

will normally restrict observations at 92 cm to a 3 MHz bandwidth).

Table 16–1.
R.m.s. Noise on Images Made with 27 VLA Antennas*

Band Designation:	92 cm P	20 cm L	6 cm C	2 cm U	1.3 cm K
Band Width $\Delta\nu$ (MHz)	3	50	50	50	50
r.m.s. noise in 5-min snapshot (mJy/beam)	2.0	0.19	0.12	0.36	1.24
r.m.s. noise in 12-hr integration (mJy/beam)	0.16	0.016	0.010	0.030	0.103

*For two IFs and natural weighting. For uniform weighting, multiply all entries by 1.5 (for a first approximation).

The sensitivity required for your observation will be determined by (a) the significance level you require for a detection in order to achieve your astronomical goals, and (b) whether the interesting emission is extended (see Section 2.1 above). If you are interested in polarimetry of the sources, calculate the sensitivity required for the polarization measurements first—this will normally drive the choice of total integration time for the experiment.

If the first estimate of $t_{\rm int}$ is significantly greater than 12 hours, consider carefully whether your choices of frequency and configuration are optimal. You may wish to re-enter the decision tree (Figure 16–1) with different starting parameters before considering the

proposal planning further. If the total integration time required is more than 4 hours, a full hour angle track is probably desirable.

If you estimate $t_{\rm int} < 4$ hours, your observing strategy should be determined by the need for dynamic range and by the availability of other sources to merge with the program. The u-v tracks on different VLA baselines begin to overlap after about 4 hours of observing. If you require high dynamic range, or wish to image an extended structure, with less than 4 hours integration time it is therefore best to fill in the u-v plane as uniformly as possible throughout a 4-hour range of hour angle around meridian transit. This can usually be done satisfactorily by distributing the observing over several short (e.g., $\sim$ 10-minute) scans spaced equally through this 4-hour range. Note however that the dynamic range achieved in a given observation is sensitive to atmospheric and ionospheric conditions, to the elevation angle range of the observation, and to your calibration strategy (Section 7 below), as well as to the u-v coverage.

If the total integration time required is much less than 1 hour, consider the use of "snapshot" mode (see the next Section).

5. Use of the VLA in "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 very brief observations if the sources to be studied 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!). To illustrate the power of snapshot made, compare the two 20 cm **A** configuration images of the source 0055+300 (NGC 315) shown as Figure 16–7(a) and 16–7(b). Contour map (a) is from a 3 minute snapshot at 50 MHz bandwidth, and has a signal-to-noise of about 200 : 1. Contour map (b) is from a 9 hour synthesis at 25 MHz bandwidth. It has a signal-to-noise of about 1500 : 1, limited by dynamic range. Apart from the obvious differences in signal-to-noise, the images show identical jet structures within 15″ of the 0.4 Jy unresolved peak.

In what follows, I consider a single "snapshot" to be an observation of about 1–5 minutes' duration. Snapshots < 1 minute long involve some risk because much of the data for a source could be lost if the instrument took unusually long to settle down after a drive from the previous source. Even shorter snapshots may 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.

5.1. Limitations of "snapshot" mode.

The clearest limitation of snapshot observing is sensitivity (see Table 16–1); it is suitable only for bright sources. At 20 cm, the high sidelobe levels of beams synthesized from snapshots exacerbate the problems created by confusing sources, so snapshots of fields near the galactic plane using the more compact VLA configurations will frequently be dominated by sidelobe clutter from confusing sources rather than by the noise that is quantified in Table 16–1. These problems are less severe at 6 cm and shorter wavelengths, because of the smaller primary beam and the typical source spectrum (see Section 6 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. 16–8) satisfies the sampling theorem and thus permits reconstruction of the correct sky brightness distribution. Table 16–2 codifies

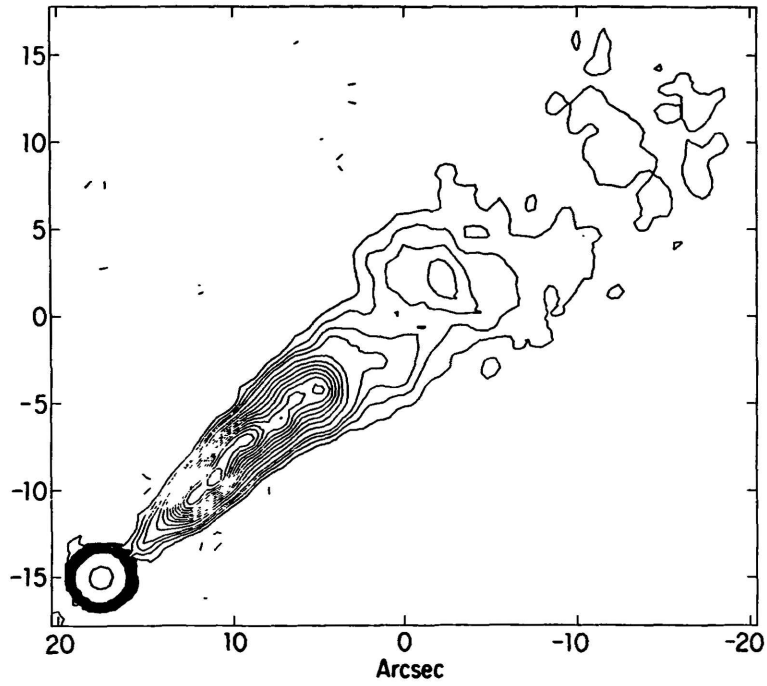


Figure 16–7(b). Contour plot of a 20 cm A configuration synthesis of the source 0055+300, made from 9 hours of data at 25 MHz bandwidth. The contour levels are drawn at −0.5, 0.5, 1, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, and 200 mJy/beam. The contour around the peak shows the HPBW. Compare with Fig. 16–7(a).

this limitation for the standard VLA configurations and frequencies.

Table 16–2.
Approximate Values of θ_{LAS} for a Single Snapshot*

	A	B	C	D
92 cm	170″	9′	30′	70′
20 cm	38″	2′	7′	15′
6 cm	10″	36″	2′	5′
2 cm	4″	10″	40″	90″
1.3 cm	2″	7″	27″	60″

*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 4, 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 20cm, 6cm and 2cm, and the default frequencies for 25 MHz and 12.5 MHz bandwidths at 6cm). Position angle calibration may also be difficult if the standard polarization calibrators (discussed in Lecture 4) 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.

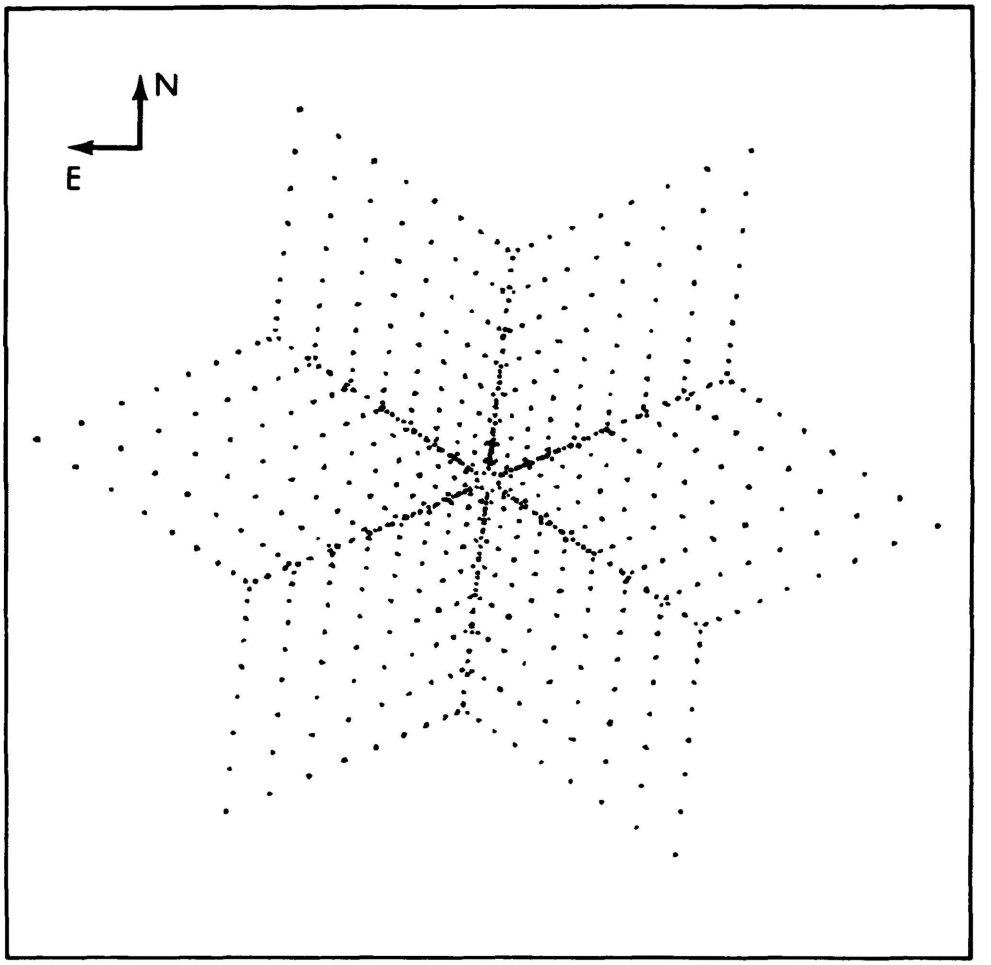


Figure 16–8. The u-v plane coverage for an instantaneous sampling of data for a source at $\delta = 30°$ and $H = 0$ by the 27-antenna VLA.

Snapshots are most effective when the sources are observed within about 2 hours of the meridian. At larger hour angles, foreshortening of the array will lead 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 source may be as large as a full synthesis image of the same source, snapshot programs also make heavy demands on disk storage. This can be especially true for snapshots made in the more compact configurations at 20 cm and 6 cm, which are particularly prone to degradation by sidelobe clutter from confusing sources (see Section 6 below). 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 source and beam images and u-v data sets whenever possible.

5.2. Multiple snapshots versus extended snapshots.

The question often arises of whether (for example) an observation requiring 15 minutes of integration time is best made as one continuous 15 minute observation or by combining

the data from three separate 5 minute snapshots. Under some circumstances, 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 might encounter different wedges and therefore combine to give an image with poorer final dynamic range. If the total time taken to acquire the data is longer than the time scale for significant changes in the phase screen in front of the region of sky being imaged, the dynamic range of the result will be degraded unless self-calibration (Lecture 9) can be used. In these circumstances, a single observation may be 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 18cm 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 however prove to be better for observations that must be made at low elevations, where phase "wedges" are more likely to arise, and in cases where 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.

6. 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

$$N(>S) = 0.032S^{-1.13} \tag{16–7}$$

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

$$N(>S) = 0.10S^{-0.9}. \tag{16–8}$$

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 16–7 by scaling flux densities to 6 cm with an effective mean spectrum of $\sim \nu^{-0.6}$.

Images made at 20 cm will therefore contain, on average, one extragalactic source of flux density 110 mJy closer to the field center than the 15′ HWHM of the primary beam of the VLA antennas. The 6 cm primary beam (4′.5 HWHM) will similarly contain, on average, one extragalactic source of flux density 2 mJy, the 2 cm beam (1′.85 HWHM) a source of < 0.1 mJy and the 1.3 cm beam (1′ HWHM) a source of < 0.01 mJy.

Individual pathological cases aside, confusion is thus unlikely to be a problem except at 20 cm and 6 cm in the VLA's more compact configurations. Confusion may have two effects on the interpretation of a synthesis image:

(1) degradation of the r.m.s. fluctuation level on the image by sidelobes or by aliasing of confusing sources, and

(2) identification of the wrong radio source as the target object in a detection experiment, or as part of the structure of an extended feature.

If you know you will be making observations near a bright confusing source, you may consider two strategies for reducing its effects on your final images. One is to plan to make wide-field images containing both the target source and the confusing source and subsequently to subtract or 'CLEAN' the confusing source and its sidelobes from the region containing the emission that is of interest. The ungridded subtraction technique[1] (Lecture 8, Section 1.3) helps this strategy considerably, as only the parts of the wide field that contain significant emission need to be computed and 'CLEAN'ed. This is probably the best technique if the angular separation of the confusing source from the region of interest is only one or two times the size of the field of view that you would otherwise have been interested in imaging. The confusing source may then be close enough that you do not require an unacceptably narrow bandwidth to include it in the minimally-distorted field around your target. If the confusing source is very strong you may want to displace the delay tracking center away from the target and towards the confusing source in order to minimize distortions of the response to the confusing source by bandwidth smearing and other effects.

This problem is likely to be encountered particularly often by snapshooters using the compact configurations at 20 cm and 6 cm, because the sidelobes resulting from the "snowflake" pattern of u-v coverage in a snapshot (Fig. 16–8) extend widely across the images. Snapshooters should therefore plan to reduce their data using the ungridded subtraction algorithm both because it permits imaging of multiple subfields and because it eliminates the effects of sidelobe aliasing.

A second approach, suitable for more distant confusing sources, is to choose your IF bandwidth and delay tracking and pointing centers so that the response to the confusing source is adequately reduced by the combined effects of bandwidth smearing and of primary beam attenuation. Because the attenuation produced by bandwidth smearing increases with baseline length $\sqrt{u^2 + v^2}$ (see Lectures 2 and 8), this attenuation does not filter confusing sources from the short-baseline data as effectively as it does from the long-baseline data. If a distant confusing source still dominates the data after attenuation by the primary beam, this approach may therefore leave 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 source. The most difficult case of all arises when the response to the confusing source is strong even after this stratagem has been adopted. Here, variable pointing errors and the rotation of the primary sidelobe pattern of the antennas on the sky (due to the VLA's altitude-azimuth antenna mounts) may make the confusing source appear to vary throughout a VLA observation; it is hard to make images of high dynamic range in this case (see also Lecture 8, Section 2.1).

If the confusing source lies in the target field itself, nothing need be done at the time of the observations, as the source and its sidelobe pattern can be 'CLEAN'ed as part of the normal data reduction. 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. In such cases the source count Equations (16–7 and 16–8) can be used to estimate the probability that the detected source occurs in the image by chance.

7. Calibration Strategy

Calibration sources should generally be chosen from the *VLA Calibrator List* maintained at the site by the NRAO staff, unless you are sure that a calibrator candidate is unresolved in the VLA configuration to be used, and has a position measured in the VLA

[1] coded in NRAO's Astronomical Image Processing System as the program 'MX'.

reference system to better than 0.1 arcsec. The basic issues to be decided by the observer are: how often to calibrate, and how close the calibrators should be to the target sources. Your strategy will depend on whether you attempt to calibrate only the instrumental fluctuations of the VLA, or these fluctuations plus the gain and phase variations introduced by the ionosphere and troposphere (see Lecture 4 for details).

7.1. Instrumental calibration.

The instrumental calibration should (a) detect grossly malfunctioning antennas so that faults might be corrected while the observations are in progress, and (b) monitor the overall amplitude and phase stability of the instrument sufficiently often that changes can be corrected for by interpolation throughout the run. Most instrumental fluctuations (apart from phase jumps) are slow, and observation of an unresolved strong calibrator every 20–60 minutes will normally be adequate for instrumental monitoring.

Bear in mind that if the instrumental calibration detects 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 phase jump can be located in the source data. If the source is strong enough, you may need to edit the data only between the gain table entries immediately before and after the phase jump—once a phase jump is localized, the gain table entries before and after it can be calibrated separately (from the earlier and later calibration observations, respectively). Of course, you you may not need to edit phase jumps in the data for strong sources at all if you will later use self-calibration to image such sources.

Calibrators for purely instrumental monitoring should be chosen primarily for their strength rather than for extreme closeness to the program source(s), particularly at 1.3 cm, where the VLA has degraded sensitivity. The interval between calibrations may vary with the total length of the program; very short programs should look at a calibrator at the beginning and the end to reassure the observer that no drastic changes have occurred during the run. It is always worth beginning a run with an observation of a calibration source, so that you can sample the data using the on-line display and come to a quick assessment of phase stability over the longer baselines, etc. Calibration of the instrumental effects more rapidly than every 30 minutes should hardly ever be necessary at 20 cm or 6 cm.

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 2 minutes at a time, however, as shorter calibrator scans may be lost as a result of unusually long settle-down times, etc. 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 calibration sources are weaker and therefore need to be observed for longer total integration times.

7.2. Atmospheric calibration.

It is more important, and also more difficult, to calibrate the amplitude and phase fluctuations resulting from changes in the propagation properties along the atmospheric path to the source. Unfortunately, ***no calibration based on observations of a reference source that is not in the same isoplanatic patch as the interesting source 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 calibration may or may not be successful. If the angular separation of the source and calibrator exceeds the scale size of the atmospheric cells responsible for the amplitude and phase variations, the fluctuations seen in the calibrator data may not be correlated with

those occurring in the source data. Corrections interpolated from the calibrator observations into the source data under these circumstances may then make the atmospheric amplitude and phase noise in the source data *worse* by a factor of $\approx \sqrt{2}$. At the other extreme, if the source and calibrator are typically within the *same* isoplanatic patch, the fluctuations observed in the calibrator will faithfully track those occurring in the source. Amplitude and phase corrections interpolated into the source 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 for your source will vary from day to day and even from hour to hour (as a function of the "weather" and of the position of your source 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.

The most reliable method for removing atmospheric fluctuations from the data is to use self-calibration, *if the source meets the basic criteria for use of this approach* (as discussed in Lecture 9). This means in practice that the source must produce sufficient signal to noise in the typical fluctuation time scale for the atmospheric phase screen over the baselines that will be used for the self-calibration.

External calibration is useful even when you know you will be able to self-calibrate your final images, for several reasons. 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 an unresolved calibrator near your source can also be used to estimate the coherence time of the atmosphere while your observations were in progress. This will enable you to judge a suitable averaging time τ_{sc} for the self-calibration (Lecture 9, Section 5.3). 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 then yield a good initial model of your source to help self-calibration of the whole data set converge quickly.

It is fortunate that 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, sources with weak extended structures around bright small-diameter components, as discussed in Lecture 11. 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, and for which external calibration is therefore still required. If you cannot, or do not wish to, rely on self-calibration to remove atmospheric effects from your data then you must choose your external calibrator(s) as close as possible to the source(s) you are observing, and hope that the amplitude and phase stability you observe on the calibrator scans meet the needs of your experiment. If the within-scan *and scan-to-scan* amplitude or phase fluctuations on a calibrator a few degrees from your source are small (less than 10% or 20°), it is unlikely that large fluctuations are occurring on your source. If you see large fluctuations on the calibrator, you are in trouble, which may or may not be mitigated by correcting the source data for the observed fluctuations. If you see *slow* drifts in the calibrator amplitude and phase, long-term ('BOXCAR') averaging of these and interpolating them as corrections into the source data should improve the output images. If you see rapid fluctuations, local point-to-point ('2POINT') interpolation of these may make matters better or make them worse. You then have little choice but to try making images from your data with both long-term averaging and with local interpolation of phase corrections from the calibrator data, to see empirically which approach gives better final images (using the final dynamic range, r.m.s. noise level, and/or any prior knowledge of the source properties to make this judgement).

Deletion of data from some or all baselines during periods of unusually bad phase

stability will usually improve the quality of images made by external calibration. If you cannot use self-calibration, imaging with a reduced amount of data of better amplitude and phase stability can give better results than imaging with a large amount of poor data, because the actual synthesized beam will be closer to the theoretical "dirty" beam in the former case. This allows deconvolution algorithms to do a better job, increasing the dynamic range of your final images. Note that tapering the final images is a way of down-weighting the data from the longer baselines where phase stability is poorer. Be ready to sacrifice resolution in favor of forming the theoretical beam more closely if the phase stability is poor, when your astronomical goals can still be met at lower resolution.

Significant atmospheric amplitude and phase fluctuations can occur on time scales of minutes, even at wavelengths of 6 cm and longer. At times of solar activity, *ionospheric* fluctuations will dominate at 18 cm and longer—they can also be rapid on the long baselines but are generally less troublesome near the minima of the sunspot cycle. It is completely impractical to adopt a *calibrator/source/calibrator* cycle that will guarantee following the fastest fluctuations of either kind. 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 not too far from the source being imaged. Keep in mind however that *no* external referencing, no matter how rapid, can be *guaranteed* to remove atmospheric fluctuations from the source data, and that time spent driving to and observing calibrators is time deleted from integration on your target source. You must decide for yourself how to play this particular roulette game during a given run.

Observers doing detection experiments will require such high dynamic range (and hence high phase stability) as observers imaging complex emission regions. (The loss of gain due to poor phase stability in a detection experiment can be estimated during the data reduction by calibrating with a > 2 hour 'BOXCAR' interpolation in the gain table, then imaging a calibrator source and determining its apparent flux density.)

The calibration done to monitor atmospheric fluctuations will, of course, calibrate the instrumental fluctuations also.

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} specified for the gain determination). The VLA default gain table interval of 10 minutes is adequate for a stable array and atmosphere, but shorter intervals are often appropriate if you will rely on external calibration.

7.3. Flux-density calibration.

If the LST range of your observing run permits, you should observe 3C 286 for a few minutes at each of the frequencies at which you have made source observations, as 3C 286 is the flux-density standard to which all VLA measurements are ultimately referred. Failing this, you should 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 most of these small-diameter sources are highly variable. The flux densities recorded in the *VLA Calibrator List* will rarely be sufficiently current to be useful in determining 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.

7.4. Polarization calibration.

This was previously discussed in Section 7 of Lecture 4, so only a brief recapitulation is given here.

To calibrate the *instrumental polarizations*, you should observe one unresolved source, 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 4). 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 synthesis calibrator. When determining the integration time for the instrumental polarization calibration, bear in mind that the leakage terms (the D's of Lecture 4) 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 source observations each time the array is pointing at the chosen 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 (see Section 5.1 above for the details). If you do not obtain an instrumental calibration, your ability to determine small degrees of polarization, and to 'CLEAN' polarized extended structures properly will be limited[1].

To calibrate the *polarization position angle* scale, observe 3C286 or 3C138 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 values is corrected later in the data reduction 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 the calibration in your program, so that the observations of 3C286 or 3C138 can be extended or rescheduled if necessary to prevent losing them due to an equipment failure. Note that this calibration is *essential* if you wish to make any use of your polarization position angle data.

At wavelengths of 18cm and longer, the position angle calibration may appear to be time variable because of fluctuations in the ionospheric Faraday rotation (Lecture 4, Section 7.3). If you will make use of the polarization position angle information at these long wavelengths, it is therefore a good idea to monitor one polarized calibrator *in the same part of the sky as your source(s)* throughout your observing run, to check whether 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

[1] Antenna-to-antenna polarization differences distort the polarization images in ways that do not satisfy the convolution theorem.

are small (< 20°), the success of this repair cannot be guaranteed, however. The observation of the polarized calibrator is best thought of as a "warning light" for the existence of ionospheric Faraday rotation problems, not necessarily as 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 at 6 cm, 2 cm or 1.3 cm, so this calibration is not required at these wavelengths.

8. Stormy Weather and What to Do About It

You can't tell the phase stability by looking out of the window.
— attributed to B. G. Clark

Some observing programs have frequency agility. When this is the case, on-site observers may wish to adjust their observation files to take account of the weather prevailing during their observing program—this is a prime reason for being on-site when your observations begin. The import of the above quote is that you have to *observe* to find out how good (or bad) the phase stability is. Clear blue skies do not guarantee good phase stability, particularly in spring and summer. Thunderstorms do however guarantee bad phase stability.

If your proposal has frequency agility, it is a good idea to monitor the VLA on-line computer's amplitude-phase ("D10") display over a long baseline as your observations start. Look at the phase on a strong calibrator for a few minutes. 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 any 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. This requires stability over the time scale of your calibration cycle (unless you are going to self-calibrate). You should therefore pay attention to the stability of the phase 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°, you might consider editing your observing file to achieve a faster calibration cycle. Users of 1.3 and 2 cm wavelengths might consider preparing several observing files with different calibration cycle times before the observations begin; this makes it easier to alter the strategy while they are in progress.

Snapshots require phase stability only for the duration of the individual snapshot. Instabilities over the calibration cycle but not on the time scale of the snapshots themselves may lead to snapshot images with fair dynamic range but uncalibrated position shifts.

In any case, the stability to be expected during a run is hard to assess in advance (unless it is very bad), and you must be prepared to observe for a while before making gross adjustments to your observing strategy.

9. The Observing Proposal

A few guidelines can be given for writing a VLA proposal to maximize its chances of being scheduled in the competition for observing time. Above all else, the project must be one whose scientific goals favorably impress the referees. A "highly-placed source who wishes to remain anonymous" notes that more concisely-written proposals are more likely to be received favorably by the referees, all else being equal. Before you begin writing a proposal, it is also worth checking whether any source you are interested in has previously been observed at the VLA—catalogs of the observed sources, with relevant instrumental parameters, can be obtained by writing to Teresa McBride at the VLA or by accessing

(18) Number of sources____ (If more than 10 sources please attach list. If more than 30 give only selection criteria and LST range(s).)

Name	Epoch 1950☐ 2000☐ RA hh mm Dec ±xx°x	Config.	Band (cm)	Band width (MHz)	Total Flux line (Jy)	Total Flux cont. (Jy)	Largest ang. size	Weakest signal (mJy/beam)	Required dynamic range	Possible LST range hh - hh	Time requested

Figure 16–9. A sample of Item 18 from the standard VLA proposal cover sheet.

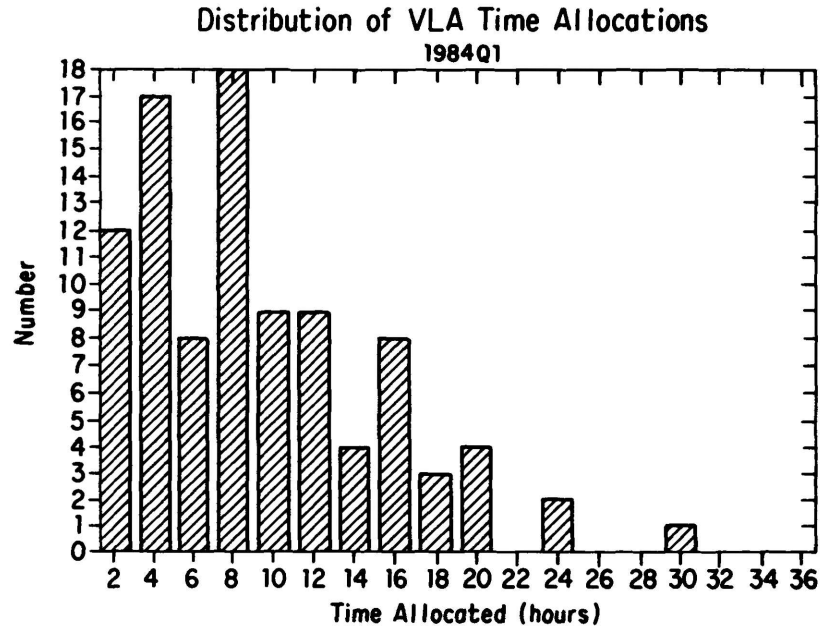


Figure 16–10. A histogram of durations of projects scheduled for VLA observations for the first two months of 1984, when the array was in the B configuration.

the relevant disk files on the VLA DEC–10 computer. See the *NRAO Newsletter* of July 1, 1985, p. 13, for details and for a brief description of NRAO policy regarding access to archived data sets.

The proposal cover sheet should be filled out in as much detail as possible. Filling out item 18 on the cover sheet (Fig. 16–9) fully for each source, or for typical sources, will lead you to consider the issues discussed in this Lecture. Your entries here should show the proposal referees and the VLA scheduling committee that the proposal is well suited to the VLA configuration(s) you are requesting.

The distribution of observing time allotted to successful proposals during the first two months of 1984, when the VLA was in the B configuration, is shown as a histogram in Figure 16–10. The median observing time scheduled is 7 hours, reflecting the large number of proposals for which less than full hour angle tracks are appropriate. 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!

Finally, submit your proposal to the NRAO Director in Charlottesville well before the

deadline given for your desired configuration(s). These deadlines and the VLA configuration schedule are published regularly in the *NRAO Newsletter* and in the *AAS Newsletter*. Proposals may be submitted between the deadline dates, and indeed 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 give you a chance to 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 time of the deadline.

ACKNOWLEDGMENTS

I thank Ron Ekers, Rick Perley and Fred Schwab for their perceptive comments on earlier versions of this Lecture, and Bob Hjellming for providing Figures 16–5 and 16–6 from the 1982 edition of the VLA "Green Book".

REFERENCES

Clark, B. G. (1981), "Orders of Magnitude of Some Instrumental Effects", VLA Scientific Memorandum No. 137.

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.", *Astron. J.*, **85**, 780.

Perley, R. A. (1981a), "The Effect of Bandwidth on the Synthesized Beam", VLA Scientific Memorandum No. 138.

Perley, R. A. (1981b), "VLA Hybrid Configuration *u-v* Plane Coverage", VLA Scientific Memorandum No. 139.

NATIONAL RADIO ASTRONOMY OBSERVATORY

OPERATED BY ASSOCIATED UNIVERSITIES, INC.
UNDER CONTRACT WITH THE NATIONAL SCIENCE FOUNDATION