INTERFEROMETRY AND SYNTHESIS IN RADIO ASTRONOMY

Second Edition

A. Richard Thompson
National Radio Astronomy Observatory

James M. Moran
Harvard-Smithsonian Center for Astrophysics

George W. Swenson, Jr.
University of Illinois at Urbana-Champaign

WILEY-VCH

WILEY-VCH Verlag GmbH & Co. KGaA

All books published by Wiley-VCH are carefully produced.
Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors.
Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.:
Applied for

British Library Cataloging-in-Publication Data:
A catalogue record for this book is available from the British Library

Bibliographic information published by
Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

© 2001 by John Wiley & Sons, Inc.
© 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

All rights reserved (including those of translation into other languages).
No part of this book may be reproduced in any form – nor transmitted or translated into machine language without written permission from the publishers.
Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Printed in the Federal Republic of Germany
Printed on acid-free paper

Printing and Bookbinding buch bücher dd ag, Birkach

ISBN-13: 978-0-471-25492-8
ISBN-10: 0-471-25492-4

To
Sheila, Barbara, Janice,
Sarah, Susan, and Michael

... truste wel that alle the conclusiouns that han ben founde, or elles possibly mighten be founde in so noble an instrument as an Astrolabie, ben un-knowe perfitly to any mortal man...

GEOFFREY CHAUCER
A Treatise on the Astrolabe
circa 1391

CONTENTS

Preface to the Second Edition xix

Preface to the First Edition xxi

1 Introduction and Historical Review 1

1.1 Applications of Radio Interferometry 1
1.2 Basic Terms and Definitions 3
Cosmic Signals 3
Source Positions and Nomenclature 9
Reception of Cosmic Signals 10
1.3 Development of Radio Interferometry 12
Evolution of Synthesis Techniques 12
Michelson Interferometer 13
Early Two-Element Radio Interferometers 16
Sea Interferometer 18
Phase-Switching Interferometer 18
Optical Identifications and Calibration Sources 21
Early Measurements of Angular Width 21
Survey Interferometers and the Mills Cross 24
Centimeter-Wavelength Solar Mapping 26
Measurements of Intensity Profiles 27
Spectral Line Interferometry 28
Earth-Rotation Synthesis Mapping 28
Development of Synthesis Arrays 31
Very-Long-Baseline Interferometry 33
VLBI Using Orbiting Antennas 37
1.4 Quantum Effect 39

2 Introductory Theory of Interferometry and Synthesis Imaging 50

2.1 Planar Analysis 50
2.2 Effect of Bandwidth 53

2.3 One-Dimensional Source Synthesis 57
Interferometer Response as a Convolution 58
Convolution Theorem and Spatial Frequency 60
Example of One-Dimensional Synthesis 61
2.4 Two-Dimensional Synthesis 64
Projection-Slice Theorem 65

3 Analysis of the Interferometer Response 68
3.1 Fourier Transform Relationship between Intensity and Visibility 68
3.2 Cross-Correlation and the Wiener–Khinchin Relation 77
3.3 Basic Response of the Receiving System 78
Antennas 78
Filters 79
Correlator 80
Response to the Incident Radiation 80
Appendix 3.1 Mathematical Representation of Noise-Like Signals 82
Analytic Signal 82
Truncated Function 84

4 Geometric Relationships and Polarimetry 86
4.1 Antenna Spacing Coordinates and (u, v) Loci 86
4.2 (u', v') Plane 90
4.3 Fringe Frequency 91
4.4 Visibility Frequencies 92
4.5 Calibration of the Baseline 93
4.6 Antenna Mounts 94
4.7 Beamwidth and Beam-Shape Effects 96
4.8 Polarimetry 97
Parameters Defining Polarization 97
Antenna Polarization Ellipse 99
Stokes Visibilities 102
Instrumental Polarization 105
Matrix Formulation 109
Calibration of Instrumental Polarization 112
Appendix 4.1 Conversion Between Hour Angle–Declination and Azimuth–Elevation Coordinates 117
Appendix 4.2 Leakage Parameters in Terms of the Polarization Ellipse 117
Linear Polarization 118
Circular Polarization 119

5 Antennas and Arrays **122**
5.1 Antennas 122
5.2 Sampling the Visibility Function 126
Sampling Theorem 126
Discrete Two-Dimensional Fourier Transform 128
5.3 Introductory Discussion of Arrays 129
Phased Arrays and Correlator Arrays 129
Spatial Sensitivity and the Spatial Transfer Function 132
Meter-Wavelength Cross and T Arrays 137
5.4 Spatial Transfer Function of a Tracking Array 138
Desirable Characteristics of the Spatial Transfer Function 140
Holes in the Spatial Frequency Coverage 141
5.5 Linear Tracking Arrays 142
5.6 Two-Dimensional Tracking Arrays 147
Open-Ended Configurations 148
Closed Configurations 150
VLBI Configurations 155
Orbiting VLBI Antennas 158
Planar Arrays 159
5.7 Conclusions on Antenna Configurations 161
5.8 Other Considerations 162
Sensitivity 162
Long Wavelengths 163
Millimeter Wavelengths 163

6 Response of the Receiving System **168**
6.1 Frequency Conversion, Fringe Rotation, and Complex Correlators 168
Frequency Conversion 168
Response of a Single-Sideband System 169
Upper-Sideband Reception 171
Lower-Sideband Reception 172
Multiple Frequency Conversions 173
Delay Tracking and Fringe Rotation 173
Simple and Complex Correlators 174
Response of a Double-Sideband System 175
Double-Sideband System with Multiple Frequency Conversions 178
Fringe Stopping in a Double-Sideband System 180
Relative Advantages of Double- and Single-Sideband Systems 181

Sideband Separation 181
6.2 Response to the Noise 183
Signal and Noise Processing in the Correlator 183
Noise in the Measurement of Complex Visibility 188
Signal-to-Noise Ratio in a Synthesized Map 189
Noise in Visibility Amplitude and Phase 192
Relative Sensitivities of Different Interferometer Systems 193
System Temperature Parameter α 199
6.3 Effect of Bandwidth 199
Mapping in the Continuum Mode 200
Wide-Field Mapping with a Multichannel System 204
6.4 Effect of Visibility Averaging 205
Visibility Averaging Time 205
Effect of Time Averaging 206
Appendix 6.1 Partial Rejection of a Sideband 208

7 Design of the Analog Receiving System 212
7.1 Principal Subsystems of the Receiving Electronics 212
Low-Noise Input Stages 212
Noise Temperature Measurement 214
Local Oscillator 217
IF and Signal Transmission Subsystems 218
Optical Fiber Transmission 218
Delay and Correlator Subsystems 220
7.2 Local Oscillator and General Considerations of Phase Stability 221
Round-Trip Phase Measuring Schemes 221
Swarup and Yang System 222
Frequency-Offset Round-Trip System 223
Automatically Correcting System 228
Fiberoptic Transmission of LO Signals 229
Phase-Locked Loops and Reference Frequencies 230
Phase Stability of Filters 232
Effect of Phase Errors 233
7.3 Frequency Responses of the Signal Channels 233
Optimum Response 233
Tolerances on Variation of the Frequency Response: Degradation of Sensitivity 235
Tolerances on Variation of the Frequency Response: Gain Errors 235

Delay-Setting Tolerances 238
Implementation of Bandpass Tolerances 239
7.4 Polarization Mismatch Errors 240
7.5 Phase Switching 240
Reduction of Response to Spurious Signals 240
Implementation of Phase Switching 241
Interaction of Phase Switching with Fringe Rotation and Delay Adjustment 246
7.6 Automatic Level Control and Gain Calibration 248
Appendix 7.1 Sideband-Separating Mixer 248
Appendix 7.2 Dispersion in Optical Fiber 249

8 Digital Signal Processing 254
8.1 Bivariate Gaussian Probability Distribution 255
8.2 Periodic Sampling 256
Nyquist Rate 256
Correlation of Sampled but Unquantized Waveforms 257
8.3 Sampling with Quantization 260
Two-Level Quantization 261
Four-Level Quantization 264
Three-Level Quantization 271
Quantization with Eight or More Levels 273
Quantization Correction 276
Comparison of Quantization Schemes 277
System Sensitivity 278
8.4 Accuracy in Digital Sampling 278
Principal Causes of Error 278
Tolerances in Three-Level Sampling 279
8.5 Digital Delay Circuits 282
8.6 Quadrature Phase Shift of a Digital Signal 283
8.7 Digital Correlators 283
Correlators for Continuum Observations 283
Principles of Digital Spectral Measurements 284
Lag (XF) Correlator 289
FX Correlator 290
Comparison of Lag and FX Correlators 293
Hybrid Correlator 297
Demultiplexing in Broadband Correlators 297
Appendix 8.1 Evaluation of $\sum_{q=1}^{\infty} R_{\infty}^2(q\tau_s)$ 298
Appendix 8.2 Probability Integral for Two-Level Quantization 299

Appendix 8.3 Correction for Four-Level Quantization 300

9 Very-Long-Baseline Interferometry 304

9.1 Early Development 304
9.2 Differences Between VLBI and Conventional Interferometry 306
9.3 Basic Performance of a VLBI System 308
Time and Frequency Errors 308
Retarded Baselines 315
Noise in VLBI Observations 316
Probability of Error in the Signal Search 319
Coherent and Incoherent Averaging 323
9.4 Fringe Fitting for a Multielement Array 326
Global Fringe Fitting 326
Relative Performance of Fringe Detection Methods 329
Triple Product, or Bispectrum 330
Fringe Searching with a Multielement Array 331
Multielement Array with Incoherent Averaging 331
9.5 Phase Stability and Atomic Frequency Standards 332
Analysis of Phase Fluctuations 332
Oscillator Coherence Time 340
Precise Frequency Standards 342
Rubidium and Cesium Standards 346
Hydrogen Maser Frequency Standard 348
Local Oscillator Stability 351
Phase Calibration System 352
Time Synchronization 353
9.6 Recording Systems 353
9.7 Processing Systems and Algorithms 357
Fringe Rotation Loss (η_R) 358
Fringe Sideband Rejection Loss (η_S) 361
Discrete Delay Step Loss (η_D) 363
Summary of Processing Losses 365
9.8 Bandwidth Synthesis 366
Burst Mode Observing 368
9.9 Phased arrays as VLBI Elements 369
9.10 Orbiting VLBI (OVLBI) 373

10 Calibration and Fourier Transformation of Visibility Data 383

10.1 Calibration of the Visibility 383
Corrections for Calculable or Directly Monitored Effects 384
Use of Calibration Sources 385
10.2 Derivation of Intensity from Visibility 387
Mapping by Direct Fourier Transformation 387
Weighting of the Visibility Data 388
Mapping by Discrete Fourier Transformation 392
Convolving Functions and Aliasing 394
Aliasing and the Signal-to-Noise Ratio 398
10.3 Closure Relationships 399
10.4 Model Fitting 401
Basic Considerations for Models 402
Cosmic Background Anisotropy 404
10.5 Spectral Line Observations 404
General Considerations 404
VLBI Observations of Spectral Lines 406
Variation of Spatial Frequency over the Bandwidth 409
Accuracy of Spectral Line Measurements 409
Presentation and Analysis of Spectral Line Observations 410
10.6 Miscellaneous Considerations 411
Interpretation of Measured Intensity 411
Errors in Maps 412
Hints on Planning and Reduction of Observations 413
Appendix 10.1 The Edge of the Moon as a Calibration Source 414
Appendix 10.2 Doppler Shift of Spectral Lines 417
Appendix 10.3 Historical Notes 421
Maps from One-Dimensional Profiles 421
Analog Fourier Transformation 422

11 Deconvolution, Adaptive Calibration, and Applications 426

11.1 Limitation of Spatial Frequency Coverage 426
11.2 The Clean Deconvolution Algorithm 427
CLEAN Algorithm 427
Implementation and Performance of the CLEAN Algorithm 429
11.3 Maximum Entropy Method 432
MEM Algorithm 432
Comparison of CLEAN and MEM 434
Other Deconvolution Procedures 435

11.4 Adaptive Calibration and Mapping With Amplitude Data Only 438
Hybrid Mapping 438
Self-Calibration 440
Mapping with Visibility Amplitude Data Only 444
11.5 Mapping With High Dynamic Range 445
11.6 Mosaicking 446
Methods of Producing the Mosaic Map 449
Some Requirements of Arrays for Mosaicking 451
11.7 Multifrequency Synthesis 453
11.8 Non-Coplanar Baselines 454
11.9 Further Special Cases of Image Analysis 459
Use of CLEAN and Self-Calibration with Spectral Line Data 459
Low-Frequency Mapping 459
Lensclean 461

12 Interferometer Techniques for Astrometry and Geodesy 467
12.1 Requirements for Astrometry 467
Reference Frames 469
12.2 Solution for Baseline and Source-Position Vectors 470
Connected-Element Systems 470
Measurements with VLBI Systems 472
Phase Referencing in VLBI 476
12.3 Time and the Motion of the Earth 480
Precession and Nutation 481
Polar Motion 482
Universal Time 482
Measurement of Polar Motion and UT1 484
12.4 Geodetic Measurements 485
12.5 Mapping Astronomical Masers 485
Appendix 12.1 Least-Mean-Squares Analysis 490

13 Propagation Effects 507
13.1 Neutral Atmosphere 508
Basic Physics 508
Refraction and Propagation Delay 513
Absorption 518
Origin of Refraction 524
Smith–Weintraub Equation 528
Phase Fluctuations 530

Kolmogorov Turbulence 534
Anomalous Refraction 539
Water Vapor Radiometry 541
13.2 Atmospheric Effects at Millimeter Wavelengths 543
Site Testing by Opacity Measurement 543
Site Testing by Direct Measurement of Phase Stability 546
Reduction of Atmospheric Phase Errors by Calibration 550
13.3 Ionosphere 554
Basic Physics 555
Refraction and Propagation Delay 559
Calibration of Ionospheric Delay 560
Absorption 562
Small- and Large-Scale Irregularities 562
13.4 Scattering Caused by Plasma Irregularities 564
Gaussian Screen Model 564
Power-Law Model 569
13.5 Interplanetary Medium 571
Refraction 571
Interplanetary Scintillation 574
13.6 Interstellar Medium 576
Dispersion and Faraday Rotation 576
Diffractive Scattering 579
Refractive Scattering 580

14 Van Cittert–Zernike Theorem, Spatial Coherence, and Scattering 594
14.1 Van Cittert–Zernike Theorem 594
Mutual Coherence of an Incoherent Source 596
Diffraction at an Aperture and the Response of an Antenna 597
Assumptions in the Derivation and Application of the Van Cittert–Zernike Theorem 600
14.2 Spatial Coherence 602
Incident Field 602
Source Coherence 603
Completely Coherent Source 606
14.3 Scattering and the Propagation of Coherence 607

15 Radio Interference 613
15.1 General Considerations 613
15.2 Short- and Intermediate-Baseline Arrays 615
Fringe-Frequency Averaging 616

Decorrelation of Broadband Signals 620
15.3 Very-Long-Baseline Systems 621
15.4 Interference From Airborne and Space Transmitters 624
Appendix 15.1 Regulation of the Radio Spectrum 625

16 Related Techniques **627**
16.1 Intensity Interferometer 627
16.2 Lunar Occultation Observations 632
16.3 Measurements on Antennas 636
16.4 Optical Interferometry 641
Modern Michelson Interferometer 642
Sensitivity of Direct Detection and Heterodyne Systems 644
Optical Intensity Interferometer 646
Speckle Imaging 647

Principal Symbols **655**

Author Index **667**

Subject Index **677**

PREFACE TO THE SECOND EDITION

Half a century of remarkable scientific progress has resulted from the application of radio interferometry to astronomy. Advances since 1986, when this book was first published, have resulted in the VLBA (Very Long Baseline Array) which is the first array fully dedicated to very-long-baseline interferometry (VLBI), the globalization of VLBI networks with the inclusion of antennas in orbit, increasing importance of spectral line observations, and improved instrumental performance at both ends of the radio spectrum. At the highest frequencies, millimeter-wavelength arrays of the Berkeley–Illinois–Maryland Association (BIMA), the Institut de Radio Astronomie Millimétrique (IRAM), Nobeyama Radio Observatory (NRO) and Owens Valley Radio Observatory (OVRO), which were in their infancy in 1986, have been greatly expanded in their capabilities. The Submillimeter Array (SMA), and the Atacama Large Millimeter Array (ALMA), which is a major international project at millimeter and submillimeter wavelengths, are under development. At low frequencies, with their special problems involving the ionosphere and wide-field mapping, the frequency coverage of the Very Large Array (VLA) has been extended down to 75 MHz, and the Giant Meter-wave Radio Telescope (GMRT), operating down to 38 MHz, has been commissioned. The Australia Telescope and an expanded Multielement Radio-linked Interferometer Network (MERLIN) have provided increased capability at centimeter wavelengths.

Such progress has led to this revised edition, the intent of which is not only to bring the material up to date but also to expand its scope and improve its comprehensibility and general usefulness. In a few cases symbols used in the first edition have been changed to follow the general usage that is becoming established in radio astronomy. Every chapter contains new material, and there are new figures and many new references. Material in the original Chapter 3 that was peripheral to the basic discussion has been condensed and moved to a later chapter. Chapter 3 now contains the essential analysis of the response of an interferometer. The section on polarization in Chapter 4 has been substantially expanded, and a brief introduction to antenna theory has been added to Chapter 5. Chapter 6 contains a discussion of the sensitivity for a wide variety of instrumental configurations. A discussion of spectral line observations is included in Chapter 10. Chapter 13 has been expanded to include a descrip-

tion of the new techniques for atmospheric phase correction, and site testing data and techniques at millimeter wavelengths. Chapter 14 has been added, and contains an examination of the van Cittert–Zernike theorem and discussions of spatial coherence and scattering, some of which is derived from the original Chapter 3.

Special thanks are due to a number of people for reviews or other help during the course of the revision. These include D. C. Backer, J. W. Benson, M. Birkinshaw, G. A. Blake, R. N. Bracewell, B. F. Burke, B. Butler, C. L. Carilli, B. G. Clark, J. M. Cordes, T. J. Cornwell, L. R. D'Addario, T. M. J. Dame, J. Davis, J. L. Davis, D. T. Emerson, R. P. Escoffier, E. B. Fomalont, L. J. Greenhill, M. A. Gurwell, C. R. Gwinn, K. I. Kellermann, A. R. Kerr, E. R. Keto, S. R. Kulkarni, S. Matsushita, D. Morris, R. Narayan, S.-K. Pan, S. J. E. Radford, R. Rao, M. J. Reid, A. Richichi, A. E. E. Rogers, J. E. Salah, F. R. Schwab, S. R. Spangler, E. C. Sutton, B. E. Turner, R. F. C. Vessot, W. J. Welch, M. C. Wiedner, and J.-H. Zhao. For major contributions to the preparation of the text and diagrams, we thank J. Heidenrich, G. L. Kessler, P. Smiley, S. Watkins, and P. Winn. For extensive help in preparation and editing we are especially indebted to P. L. Simmons. We are grateful to P. A. Vanden Bout, Director of the National Radio Astronomy Observatory, and to I. I. Shapiro, Director of the Harvard–Smithsonian Center for Astrophysics, for encouragement and support. The National Radio Astronomy Observatory is operated by Associated Universities, Inc. under contract with the National Science Foundation, and the Harvard–Smithsonian Center for Astrophysics is operated by Harvard University and the Smithsonian Institution.

A. RICHARD THOMPSON
JAMES M. MORAN
GEORGE W. SWENSON, JR.

Charlottesville, Virginia
Cambridge, Massachusetts
Urbana, Illinois
November 2000

PREFACE TO THE FIRST EDITION

The techniques of radio interferometry as applied to astronomy and astrometry have developed enormously in the past four decades, and the attainable angular resolution has advanced from degrees to milliarcseconds, a range of over six orders of magnitude. As arrays for synthesis mapping* have developed, techniques in the radio domain have overtaken those in optics in providing the finest angular detail in astronomical images. The same general developments have introduced new capabilities in astrometry and in the measurement of the earth's polar and crustal motions. The theories and techniques that underlie these advances continue to evolve, but have reached by now a sufficient state of maturity that it is appropriate to offer a detailed exposition.

The book is intended primarily for graduate students and professionals in astronomy, electrical engineering, physics, or related fields who wish to use interferometric or synthesis-mapping techniques in astronomy, astrometry, or geodesy. It is also written with radio systems engineers in mind and includes discussions of important parameters and tolerances for the types of instruments involved. Our aim is to explain the underlying principles of the relevant interferometric techniques but to limit the discussion of details of implementation. Such details of the hardware and the software are largely specific to particular instruments and are subject to change with developments in electronic engineering and computing techniques. With an understanding of the principles involved, the reader should be able to comprehend the instructions and instrumental details that are encountered in the user-oriented literature of most observatories.

The book does not stem from any course of lectures, but the material included is suitable for a graduate-level course. A teacher with experience in the techniques described should be able to interject easily any necessary guidance to emphasize astronomy, engineering, or other aspects as required.

The first two chapters contain a brief review of radio astronomy basics, a short history of the development of radio interferometry, and a basic discussion of the operation of an interferometer. Chapter 3 discusses the underlying relationships of interferometry from the viewpoint of the theory of partial coherence and may

*We define synthesis mapping as the reconstruction of images from measurements of the Fourier transforms of their brightness distributions. In this book the terms map, image, and brightness (intensity) distribution are largely interchangeable.

be omitted from a first reading. Chapter 4 introduces coordinate systems and parameters that are required to describe synthesis mapping. It is appropriate then to examine configurations of antennas for multielement synthesis arrays in Chapter 5. Chapters 6–8 deal with various aspects of the design and response of receiving systems, including the effects of quantization in digital correlators. The special requirements of very-long-baseline interferometry (VLBI) are discussed in Chapter 9. The foregoing material covers in detail the measurement of complex visibility and leads to the derivation of radio maps discussed in Chapters 10 and 11. The former presents the basic Fourier transformation method, and the latter the more powerful algorithms that incorporate both calibration and transformation. Precision observations in astrometry and geodesy are the subject of Chapter 12. There follow discussions of factors that can degrade the overall performance, namely, effects of propagation in the atmosphere, the interplanetary medium and the interstellar medium in Chapter 13, and radio interference in Chapter 14. Propagation effects are discussed at some length since they involve a wide range of complicated phenomena that place fundamental limits on the measurement accuracy. The final chapter describes related techniques including intensity interferometry, speckle interferometry, and lunar occultation observations.

References are included to seminal papers and to many other publications and reviews that are relevant to the topics of the book. Numerous descriptions of instruments and observations are also referenced for purposes of illustration. Details of early procedures are given wherever they are of help in elucidating the principles or origin of current techniques, or because they are of interest in their own right. Because of the diversity of the phenomena described, it has been necessary, in some cases, to use the same mathematical symbol for different quantities. A glossary of principal symbols and usage follows the final chapter.

The material in this book comes only in part from the published literature, and much of it has been accumulated over many years from discussions, seminars, and the unpublished reports and memoranda of various observatories. Thus we acknowledge our debt to colleagues too numerous to mention individually. Our special thanks are due to a number of people for critical reviews of portions of the book, or other support. These include D. C. Backer, D. S. Bagri, R. H. T. Bates, M. Birkinshaw, R. N. Bracewell, B. G. Clark, J. M. Cordes, T. J. Cornwell, L. R. D'Addario, J. L. Davis, R. D. Ekers, J. V. Evans, M. Faucherre, S. J. Franke, J. Granlund, L. J. Greenhill, C. R. Gwinn, T. A. Herring, R. J. Hill, W. A. Jeffrey, K. I. Kellermann, J. A. Klobuchar, R. S. Lawrence, J. M. Marcaide, N. C. Mathur, L. A. Molnar, P. C. Myers, P. J. Napier, P. Nisenson, H. V. Poor, M. J. Reid, J. T. Roberts, L. F. Rodriguez, A. E. E. Rogers, A. H. Rots, J. E. Salah, F. R. Schwab, I. I. Shapiro, R. A. Sramek, R. Stachnik, J. L. Turner, R. F. C. Vessot, N. Wax, and W. J. Welch. The reproduction of diagrams from other publications is acknowledged in the captions, and we thank the authors and the publishers concerned for permission to use this material. For major contributions to the preparation of the manuscript, we wish to thank C. C. Barrett, C. F. Burgess, N. J. Diamond, J. M. Gillberg, J. G. Hamwey, E. L. Haynes, G. L. Kessler, K. I. Maldonis, A. Patrick, V. J. Peterson, S. K. Rosenthal, A. W. Shepherd, J. F. Singarella, M. B. Weems, and C. H. Williams. We are grateful to M. S. Roberts and P. A. Vanden Bout, for-

mer Director and present Director of the National Radio Astronomy Observatory, and to G. B. Field and I. I. Shapiro, former Director and present Director of the Harvard–Smithsonian Center for Astrophysics, for encouragement and support. Much of the contribution by J. M. Moran was written while on sabbatical leave at the Radio Astronomy Laboratory of the University of California, Berkeley, and he is grateful to W. J. Welch for hospitality during that period. G. W. Swenson, Jr. thanks the Guggenheim Foundation for a fellowship during 1984–1985. Finally, we acknowledge the support of our home institutions: the National Radio Astronomy Observatory which is operated by Associated Universities, Inc. under contract with the National Science Foundation; the Harvard-Smithsonian Center for Astrophysics which is operated by Harvard University and the Smithsonian Institution; and the University of Illinois.

A. RICHARD THOMPSON
JAMES M. MORAN
GEORGE W. SWENSON, JR.

Charlottesville, Virginia
Cambridge, Massachusetts
Urbana, Illinois
January 1986

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

1 Introduction and Historical Review

The subject of this book can be broadly described as the principles of radio interferometry applied to the measurement of natural radio signals from cosmic sources. The uses of such measurements lie mainly within the domains of astronomy, astrometry, and geodesy. As an introduction we consider in this chapter the applications of the technique, some basic terms and concepts, and the historical development of the instruments and their uses.

1.1 APPLICATIONS OF RADIO INTERFEROMETRY

Radio interferometers and synthesis arrays, which are basically ensembles of two-element interferometers, are used to make measurements of the fine angular detail in the radio emission from the sky. The angular resolution of single radio antennas is insufficient for many astronomical purposes. Practical considerations limit the resolution to a few tens of arcseconds. For example, the beamwidth of a 100-m-diameter antenna at 7 mm wavelength is approximately 17 arcsec. In the optical range the diffraction limit of large telescopes (diameter ~8 m) is about 0.015 arcsec, but the angular resolution achievable from the ground by conventional techniques is limited to about one arcsec by turbulence in the troposphere. For progress in astronomy it is particularly important to measure the positions of radio sources with sufficient accuracy to allow identification with objects detected in the optical and other parts of the electromagnetic spectrum. It is also very important to be able to measure parameters such as intensity, polarization, and frequency spectrum with similar angular resolution in both the radio and optical domains. Radio interferometry enables such studies to be made.

Precise measurement of the angular positions of stars and other cosmic objects is the concern of astrometry. This includes the study of the small changes in celestial positions attributable to the parallax introduced by the earth's orbital motion, as well as those resulting from the intrinsic motions of the objects. Such measurements are an essential step in the establishment of the distance scale of the universe. Astrometric measurements have also provided a means to test the general theory of relativity and to establish the dynamical parameters of the solar system. In making astrometric measurements it is essential to establish a reference frame for celestial positions. A frame based on extremely distant large-mass

objects as position references is close to ideal. Radio measurements of distant, compact, extragalactic sources presently offer the best prospects for the establishment of such a system. Radio techniques provide an accuracy of the order of 10^{-3} arcsec for absolute positions and 10^{-5} arcsec or less for the relative positions of objects closely spaced in angle. Optical measurements of stellar images, as seen through the earth's atmosphere, allow the positions to be determined with a precision of about 0.05 arcsec. However, stellar positions have been measured to $\sim$1 milliarcsecond (mas) with the Hipparcos satellite, and optical measurements with the National Aeronautics and Space Administration (NASA) Space Interferometry Mission hold promise of position measurements to $\sim$4 μarcsec.

As part of the measurement process, astrometric observations include a determination of the orientation of the instrument relative to the celestial reference frame. Ground-based observations therefore provide a measure of the variation of the orientation parameters for the earth. In addition to the well-known precession and nutation of the direction of the axis of rotation, there are irregular shifts of the earth's axis relative to the surface. These shifts, referred to as *polar motion*, are attributed to the gravitational effects of the sun and moon on the equatorial bulge of the earth, and to dynamic effects in the earth's mantle, crust, oceans, and atmosphere. The same causes give rise to changes in the angular rotation velocity of the earth, which are manifest as corrections that must be applied to the system of universal time. Measurements of the orientation parameters are important in the study of the dynamics of the earth. During the 1970s it became clear that radio techniques could provide an accurate measure of these effects, and in the late 1970s the first radio programs devoted to the monitoring of universal time and polar motion were set up jointly by the U.S. Naval Observatory and the U.S. Naval Research Laboratory, and also by NASA and the National Geodetic Survey. Polar motion can also be studied by observation of satellites, in particular the Global Positioning System, but distant radio sources provide the best standard for measurement of earth rotation.

In addition to revealing angular changes in the motion and orientation of the earth, precise interferometer measurements entail an astronomical determination of the vector spacing between the antennas, which for spacings of $\sim$100 km or more, is usually more precise than can be obtained by conventional surveying techniques. Very-long-baseline interferometry (VLBI) involves antenna spacings of hundreds or thousands of kilometers, and the uncertainty with which these spacings can be determined has decreased from a few meters in 1967, when VLBI measurements were first made, to a few millimeters. Average relative motions of widely spaced sites on separate tectonic plates lie in the range 1–10 cm per year, and have been tracked extensively with VLBI networks. Interferometric techniques have also been applied to the tracking of vehicles on the lunar surface and the determination of the positions of spacecraft. In this book, however, we limit our concern mainly to measurements of natural signals from astronomical objects. The attainment of the highest angular resolution in the radio domain of the electromagnetic spectrum results in part from the ease with which radio frequency signals can be processed electronically. Also, the phase variations induced by the earth's neutral atmosphere are less severe than at shorter wavelengths. Fu-

ture technology will provide even higher resolution at infrared and optical wavelengths from observatories above the earth's atmosphere. However, radio waves will remain of vital importance in astronomy since they reveal objects that do not radiate in other parts of the spectrum, and they are able to pass through galactic dust clouds that obscure the view in the optical range.

1.2 BASIC TERMS AND DEFINITIONS

This section is written for readers who are unfamiliar with the basics of radio astronomy. It presents a brief review of some background information that is useful when approaching the subject of radio interferometry.

Cosmic Signals

The voltages induced in antennas by radiation from cosmic sources are generally referred to as *signals*, although they do not contain information in the usual engineering sense. Such signals are generated by natural processes and almost universally have the form of Gaussian random noise. That is to say, the voltage as a function of time at the terminals of a receiving antenna can be described as a series of very short pulses of random occurrence that combine as a waveform with Gaussian amplitude distribution. In a bandwidth $\Delta\nu$ the envelope of the radio frequency waveform has the appearance of random variations with duration of order $1/\Delta\nu$. For most radio sources the characteristics of the signals are invariant with time, at least on the scale of minutes or hours typical of the duration of a radio astronomy observation. Gaussian waveforms of this type are assumed to be identical in character to the noise voltages generated in resistors and amplifiers. Such waveforms are usually assumed to be stationary and ergodic, that is, ensemble averages and time averages converge to equal values.

Most of the power is in the form of *continuum radiation*, the power spectrum of which shows slow variation with frequency and may be regarded as constant over the receiving bandwidth of most instruments. Figure 1.1 shows continuum spectra of three radio sources. Radio emission from the radio galaxy Cygnus A and from the quasar 3C48 is generated by the synchrotron mechanism [see, e.g., Rybicki and Lightman (1979), Longair (1992)], in which high-energy electrons in magnetic fields radiate as a result of their orbital motion. The radiating electrons are generally highly relativistic, and under these conditions the radiation emitted by each one is concentrated in the direction of its instantaneous motion. An observer therefore sees pulses of radiation from those electrons whose orbital motion lies in, or close to, a plane containing the observer. The observed polarization of the radiation is mainly linear, and any circularly polarized component is generally very small. The overall linear polarization from a source, however, is seldom large, since it is randomized by the variation of the direction of the magnetic field within the source and by Faraday rotation. The power in the electromagnetic pulses from the electrons is concentrated at harmonics of the orbital frequency, and a continuous distribution of electron energies results in a contin-

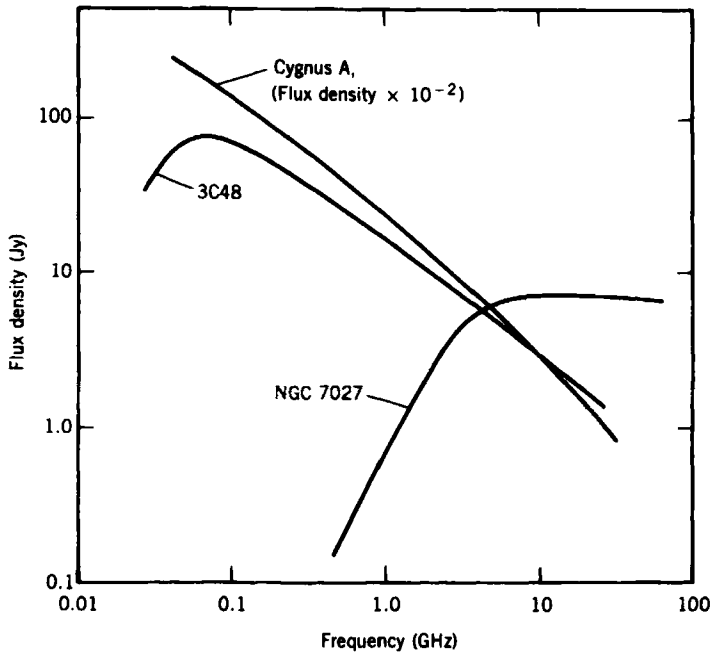


Figure 1.1 Continuum spectra of three discrete sources: Cygnus A, a radio galaxy; 3C48, a quasar; and NGC7027, an ionized nebula within our Galaxy. Data are from Conway, Kellermann, and Long (1963); Kellermann and Pauliny-Toth (1969); and Thompson (1974). [One jansky (Jy) = 10^{-26} W m^{-2} Hz^{-1}.]

uum radio spectrum. The individual pulses from the electrons are too numerous to be separable, and the electric field appears as a continuous random process with zero mean. The variation of the spectrum as a function of frequency is related to the slope of the energy distribution of the electrons. In the quasar in Fig. 1.1, which is a very much more compact object than the radio galaxy, the electron density and magnetic fields are high enough to produce self-absorption of the radiation at low frequencies.

NGC7027, the spectrum of which is shown in Fig. 1.1, is a planetary nebula within our Galaxy in which the gas is ionized by radiation from a central star. The radio emission is a thermal process and results from free–free collisions between unbound electrons and ions within the plasma. At the low-frequency end of the spectral curve the nebula is opaque to its own radiation and emits a blackbody spectrum, for which the Rayleigh–Jeans law is a valid approximation. As the frequency increases, the absorptivity, and hence the emissivity, decrease approximately as ν^{-2} [see, e.g., Rybicki and Lightman (1979)], where ν is the frequency. This behavior counteracts the ν^2 dependence of the Rayleigh–Jeans law, and thus the spectrum becomes flat when the nebula is no longer opaque to the radiation. Radiation of this type is unpolarized.

In contrast to continuum radiation, *spectral line radiation* is generated at specific frequencies by atomic and molecular processes. A fundamentally important

line is that of neutral atomic hydrogen at 1420.405 MHz, which results from the transition between two energy levels of the atom, the separation of which is related to the spin vector of the electron in the magnetic field of the nucleus. The natural width of the hydrogen line is negligibly small, but Doppler shifts caused by thermal motion of the atoms and large-scale motion of gas clouds spread the line radiation. The overall Doppler spread within our Galaxy covers several hundred kilohertz. Information on galactic structure is obtained by comparison of these velocities with those of models incorporating galactic rotation.

Our Galaxy and others like it also contain large molecular clouds at temperatures of 10–100 K in which new stars are continually forming. These clouds give rise to many atomic and molecular transitions in the radio and far-infrared ranges. Over 4500 molecular lines from approximately 80 molecular species have been measured (Lovas, Snyder, and Johnson 1979; Lovas 1992). A list of atomic and molecular lines is given by Rohlfs and Wilson (1996)—see bibliography. A few of the more important lines are given in Table 1.1. Note that this table contains less than 1% of the known lines in the frequency range below 1 THz. Figure 1.2 shows the spectrum of radiation of many molecular lines from the Orion nebula in the bands from 214 to 246 and from 329 to 360 GHz. Although the radio win-

TABLE 1.1 Some Important Radio Lines

Chemical Name	Chemical Formula	Transition	Frequency (GHz)
Deuterium	D	$^2S_{\frac{1}{2}}, F = \frac{3}{2} \rightarrow \frac{1}{2}$	0.327
Hydrogen	H1	$^2S_{\frac{1}{2}}, F = 1 \rightarrow 0$	1.420
Hydroxyl radical	OH	$^2\Pi_{\frac{3}{2}}, J = \frac{3}{2}, F = 1 \rightarrow 2$	1.612[a]
Hydroxyl radical	OH	$^2\Pi_{\frac{3}{2}}, J = \frac{3}{2}, F = 1 \rightarrow 1$	1.665[a]
Hydroxyl radical	OH	$^2\Pi_{\frac{3}{2}}, J = \frac{3}{2}, F = 2 \rightarrow 2$	1.667[a]
Hydroxyl radical	OH	$^2\Pi_{\frac{3}{2}}, J = \frac{3}{2}, F = 2 \rightarrow 1$	1.721[a]
Methyladyne	CH	$^2\Pi_{\frac{1}{2}}, J = \frac{1}{2}, F = 1 \rightarrow 1$	3.335
Hydroxyl radical	OH	$^2\Pi_{\frac{1}{2}}, J = \frac{1}{2}, F = 1 \rightarrow 0$	4.766[a]
Formaldehyde	H_2CO	$1_{10} - 1_{11}$, six F transitions	4.830
Hydroxyl radical	OH	$^2\Pi_{\frac{3}{2}}, J = \frac{5}{2}, F = 3 \rightarrow 3$	6.035[a]
Methanol	CH_3OH	$5_1 \rightarrow 6_0A^+$	6.668[a]
Helium	$^3He^+$	$^2S_{\frac{1}{2}}, F = 1 \rightarrow 0$	8.665
Methanol	CH_3OH	$2_0 \rightarrow 3_{-1}E$	12.179[a]
Formaldehyde	H_2CO	$2_{11} \rightarrow 2_{12}$, four F transitions	14.488
Cyclopropenylidene	C_3H_2	$1_{10} \rightarrow 1_{01}$	18.343
Water	H_2O	$6_{16} \rightarrow 5_{23}$, five F transitions	22.235[a]
Ammonia	NH_3	$1, 1 \rightarrow 1, 1$, eighteen F transitions	23.694
Ammonia	NH_3	$2, 2 \rightarrow 2, 2$, seven F transitions	23.723
Ammonia	NH_3	$3, 3 \rightarrow 3, 3$, seven F transitions	23.870
Methanol	CH_3OH	$6_2 \rightarrow 6_1, E$	25.018
Silicon monoxide	SiO	$v = 2, J = 1 \rightarrow 0$	42.821[a]
Silicon monoxide	SiO	$v = 1, J = 1 \rightarrow 0$	43.122[a]

TABLE 1.1 *(Continued)*

Chemical Name	Chemical Formula	Transition	Frequency (GHz)
Carbon monosulfide	CS	$J = 1 \rightarrow 0$	48.991
Silicon monoxide	SiO	$v = 1, J = 2 \rightarrow 1$	86.243[a]
Hydrogen cyanide	HCN	$J = 1 \rightarrow 0$, three F transitions	88.632
Formylium	HCO^+	$J = 1 \rightarrow 0$	89.189
Diazenylium	N_2H^+	$J = 1 \rightarrow 0$, seven F transitions	93.174
Carbon monosulfide	CS	$J = 2 \rightarrow 1$	97.981
Carbon monoxide	$^{12}C^{18}O$	$J = 1 \rightarrow 0$	109.782
Carbon monoxide	$^{13}C^{16}O$	$J = 1 \rightarrow 0$	110.201
Carbon monoxide	$^{12}C^{17}O$	$J = 1 \rightarrow 0$, three F transitions	112.359
Carbon monoxide	$^{12}C^{16}O$	$J = 1 \rightarrow 0$	115.271
Carbon monosulfide	CS	$J = 3 \rightarrow 2$	146.969
Water	H_2O	$3_{13} \rightarrow 2_{20}$	183.310[a]
Carbon monoxide	$^{12}C^{16}O$	$J = 2 \rightarrow 1$	230.538
Carbon monosulfide	CS	$J = 5 \rightarrow 4$	244.936
Carbon monosulfide	CS	$J = 7 \rightarrow 6$	342.883
Carbon monoxide	$^{12}C^{16}O$	$J = 3 \rightarrow 2$	345.796
Water	H_2O	$4_{14} \rightarrow 3_{21}$	380.197
Carbon monoxide	$^{12}C^{16}O$	$J = 4 \rightarrow 3$	461.041
Heavy water	HDO	$1_{01} \rightarrow 0_{00}$	464.925
Carbon	C1	$^3P_1 \rightarrow {}^3P_0$	492.162
Water	H_2O	$1_{10} \rightarrow 1_{01}$	556.936
Ammonia	NH_3	$1_0 \rightarrow 0_0$	572.498
Carbon monoxide	$^{12}C^{16}O$	$J = 6 \rightarrow 5$	691.473
Carbon monoxide	$^{12}C^{16}O$	$J = 7 \rightarrow 6$	806.652
Carbon	C1	$^3P_2 \rightarrow {}^3P_1$	809.340

[a] Strong maser transition.

dow in the earth's atmosphere ends above ~1 THz, sensitive submillimeter- and millimeter-wavelength arrays should be able to detect such lines as the $^2P_{3/2} \rightarrow {}^2P_{1/2}$ line of CII at 1.90054 THz (158 μm), which will be Doppler shifted into the radio window for redshifts (z) greater than ~2. Some of the lines, notably those of OH, H_2O, SiO, and CH_3OH, show very intense emission from sources of very small apparent angular diameter. This emission is believed to be generated by a maser process [see, e.g., Reid and Moran (1988), Elitzur (1992)].

The strength of the radio signal received from a discrete source is expressed as the *spectral flux density*, or *spectral power flux density*, and is measured in watts per square meter per hertz (W m^{-2} Hz^{-1}). For brevity, astronomers often refer to this quantity as *flux density*. The unit of spectral power flux density is the jansky (Jy); 1 Jy = 10^{-26}W m^{-2} Hz^{-1}. It is used for both spectral line and continuum radiation. The measure of radiation integrated in frequency over a spectral band has units of W m^{-2} and is referred to as *power flux density*, or simply *flux density*. In the standard definition of the IEEE (1977), power flux density is equal to the

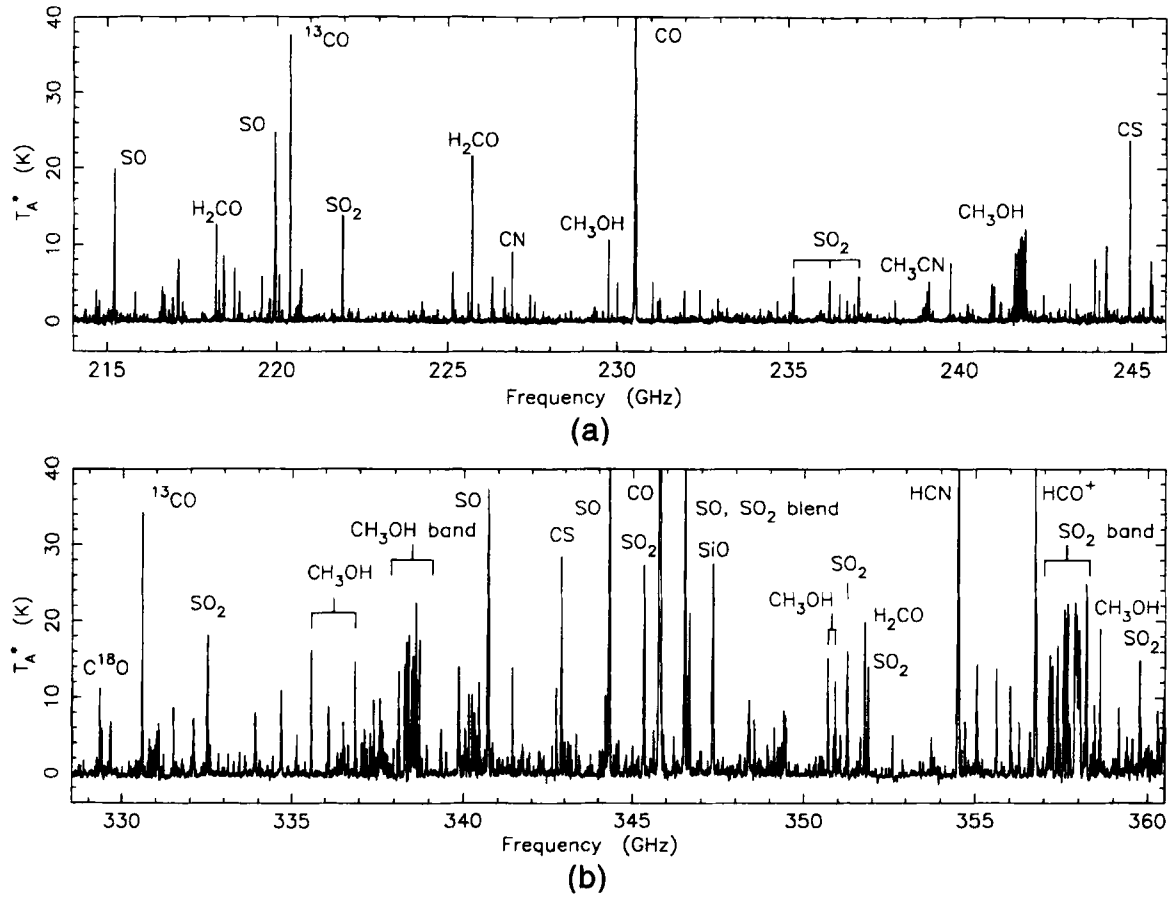


Figure 1.2 Spectrum of the Orion Nebula for 214–246 and 329–360 GHz. The ordinate is antenna temperature corrected for atmospheric absorption, which is proportional to the power received. The frequency scale has been corrected for motion of the earth with respect to the local standard of rest. The spectral resolution is 1 MHz. Note the higher density of lines in the higher frequency band. The measurements shown in panel (a) are from Blake et al. (1987), and those in panel (b) are from Schilke et al. (1997).

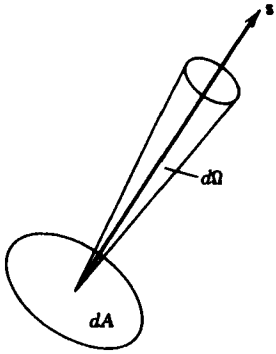


Figure 1.3 Elements of solid angle and surface area illustrating the definition of intensity. dA is normal to s.

time average of the Poynting vector of the wave. In producing a map or image* of a radio source the desired quantity is the spectral power flux density emitted per unit solid angle subtended by the radiating surface, which is measured in units of $\mathrm{W\ m^{-2}\ Hz^{-1}\ sr^{-1}}$. This quantity is variously referred to as the *intensity, specific intensity*, or *brightness* of the radiating surface. In radio astronomical mapping we can measure the intensity in only two dimensions on the surface of the celestial sphere, and the measured emission is the component normal to that surface, as seen by the observer.

In radiation theory the quantity intensity, or specific intensity, often represented by I_ν, is the measure of radiated energy flow per unit area, per unit time, per unit frequency bandwidth, and per unit solid angle. Thus in Fig. 1.3 the power flowing in direction s within solid angle $d\Omega$, frequency band $d\nu$, and area dA is $I_\nu(\mathbf{s})\, d\Omega\, d\nu\, dA$. This can be applied to emission from the surface of a radiating object, to propagation through a surface in space, or to reception on the surface of a transducer or detector. The last case applies to reception in an antenna and the solid angle then denotes the area of the celestial sphere from which the radiation emanates. Henceforth we use I to denote I_ν [note that in optical astronomy the specific intensity is usually defined as the intensity per unit bandwidth I_λ; see, e.g., Rybicki and Lightman (1979)].

For thermal radiation from a blackbody the intensity is related to the physical temperature T of the radiating matter by the Planck formula, for which

$$I = \frac{2kT\nu^2}{c^2}\left[\frac{\frac{h\nu}{kT}}{e^{h\nu/kT} - 1}\right], \tag{1.1}$$

where k is Boltzmann's constant, c is the velocity of light, and h is Planck's constant. When $h\nu \ll kT$, we can use the Rayleigh–Jeans approximation, in which case the expression in the square brackets is replaced by unity. The Rayleigh–

*The terms *map* and *image* are basically interchangeable as used in most places in this book. However, in the presentation there is some logic in using *map* for a contour depiction and *image* for one in gray scale or false color.

Jeans approximation requires ν (GHz) $\ll 20T$ (K), and is violated at high frequencies and low temperatures in many astronomical stiuations. However, for any radiation mechanism a brightness temperature T_B can be defined:

$$T_B = \frac{c^2 I}{2k\nu^2}. \tag{1.2}$$

In the Rayleigh–Jeans domain the brightness temperature T_B is that of a blackbody at physical temperature $T = T_B$. In the examples in Fig. 1.1, T_B is of the order of 10^4 K for NGC7027 and indicates the electron temperature. For Cygnus A and 3C48, T_B is of the order of 10^8 K or greater and is a measure of the energy density of the electrons and the magnetic fields, not a physical temperature. As a spectral line example, T_B for the carbon monoxide lines from molecular clouds is typically 10–100 K. In this case T_B is proportional to the excitation temperature associated with the energy levels of the transition, and is related to the temperature and density of the gas as well as to the temperature of the radiation field.

Source Positions and Nomenclature

The positions of radio sources are measured in the celestial coordinates *right ascension* and *declination*. On the celestial sphere these quantities are analogous, respectively, to longitude and latitude on the earth. The zero of right ascension is arbitrarily chosen as the point at which the sun crosses the celestial equator at the first point of Aries at a given epoch. In the international system of nomenclature (IAU 1974) radio sources are designated as follows. The first four characters give the hour and minutes of right ascension (RA); the fifth, the sign of the declination; and the remaining three, the degrees and truncated tenths of a degree of declination (Dec.) for the mean equator and equinox of 1950. For example, the source at RA $01^h 34^m 49.83^s$, Dec. $32° 54' 20.5''$ is designated 0134 + 329. Earlier nomenclature persists for many sources. Thus in Fig. 1.1, Cygnus A is the strongest radio source in the constellation of Cygnus, and 3C48 indicates source number 48 in the third Cambridge survey (Edge et al. 1959). NGC7027 is an optical designation and refers to the New General Catalog of nonstellar objects by Dreyer (1888), in which the majority of listings are galaxies.

Positions of objects in celestial coordinates vary as a result of precession and nutation of the earth's axis of rotation, aberration, and proper motion. Positions of radio sources are usually listed for the standard epochs 1950 or 2000. Procedures for the reduction of these positions to those for specific dates and times are given in Seidelmann (1992). A survey by Condon et al. (1998) using the Very Large Array (VLA) at 1.4 GHz contains approximately 2×10^6 sources. Most of the radio sources that have been detected are believed to be radio galaxies or quasars that lie far beyond our Galaxy. Another notable source list contains the 212 extragalactic sources with positional accuracy exceeding 1 mas that are used to define the International Celestial Reference Frame of the IAU, plus 396 sources with positions to a few milliarcseconds (Ma et al. 1998).

Reception of Cosmic Signals

The antennas used most commonly in radio astronomy are of the reflector type mounted to allow tracking over most of the sky. The exceptions are mainly instruments designed for meter or longer wavelengths. The collecting area A of a reflector antenna, for radiation incident in the center of the main beam, is equal to the geometric area multiplied by an aperture efficiency factor which is typically within the range 0.3–0.8. The received power P_A delivered by the antenna to a matched load in a bandwidth $\Delta\nu$, from a randomly polarized source of flux density S, assumed to be small compared to the beamwidth, is given by

$$P_A = \tfrac{1}{2} A S \, \Delta\nu. \tag{1.3}$$

Note that S is the intensity I integrated over the solid angle of the source. The factor $\frac{1}{2}$ takes account of the fact that the antenna responds to only one-half the power in the randomly polarized wave. It is often convenient to express random noise power, P, in terms of an effective temperature $T = P/k\,\Delta\nu$ where k is Boltzmann's constant. In the Rayleigh–Jeans domain, P is equal to the noise power delivered to a matched load by a resistor at physical temperature T (Nyquist 1928). In the general case, if we use the Planck formula, we can write $P = kT_{\text{Planck}}\,\Delta\nu$, where T_{Planck} is an effective radiation temperature, or noise temperature, of a load at physical temperature T, and is given by

$$T_{\text{Planck}} = T\left[\frac{\frac{h\nu}{kT}}{e^{h\nu/kT} - 1}\right]. \tag{1.4}$$

The noise power in a receiving system† can be specified in terms of the *system temperature* T_S associated with a matched resistive load that would produce an equal power level in an equivalent noise-free receiver when connected to the input terminals. T_S is defined as the power available from this load divided by $k\,\Delta\nu$. In terms of the Planck formula, the relation between T_S and the physical temperature, T, of such a load is given by replacing T_{Planck} by T_S in Eq. (1.4).

The system temperature consists of two parts: T_R, the *receiver temperature*, which represents the internal noise from the receiving amplifiers; and T_A', the *antenna temperature*, which represents the unwanted noise from the antenna produced by ground radiation, atmospheric attenuation, ohmic losses, and other sources.

It is important to note that the term *antenna temperature* is also used to refer to the component of the antenna output power that results from a source under study, which is the way it is most often used in this book. In that case the power received in an antenna from the source is

$$P_A = kT_A\,\Delta\nu, \tag{1.5}$$

†In radio astronomy the terms *receiver* and *receiving system* are broadly used and generally denote the electronic system following the output of the antenna(s) and may, or may not, include one or more detectors or correlators (which we define later) and subsequent processing and recording equipment.

and T_A is related to the flux density by Eqs. (1.3) and (1.5). It is useful to express this relation as $T_A(\mathrm{K}) = S(\mathrm{Jy}) \times A(\mathrm{m}^2)/2800$. Astronomers sometimes specify the performance of an antenna in terms of *janskys per kelvin*, that is, the flux density (in units of 10^{-26} W m^{-2} Hz^{-1}), of a point source that increases T_A by one kelvin. Thus this measure is equal to 2800/A(m^2) Jy K^{-1}.

Another term that may be encountered is the *system equivalent flux density*, S_E, which is an indicator of the combined sensitivity of both an antenna and receiving system. It is equal to the flux density of a point source in the main beam of the antenna that would cause the noise power in the receiver to be twice that of the system noise in the absence of a source. Equating P_A in Eq. (1.3) with $kT_S\,\Delta\nu$, we obtain

$$S_E = \frac{2kT_S}{A}. \tag{1.6}$$

The ratio of the signal power from a source to the noise power in the receiving amplifier is T_A/T_S. Because of the random nature of the signal and noise, measurements of the power levels made at time intervals separated by $(2\Delta\nu)^{-1}$ can be considered independent. A measurement in which the signal level is averaged for a time τ contains approximately $2\Delta\nu\tau$ independent samples. The signal-to-noise ratio $\mathcal{R}_{\mathrm{sn}}$ at the output of a power-measuring device attached to the receiver is increased in proportion to the square root of the number of independent samples and is of the form

$$\mathcal{R}_{\mathrm{sn}} = C\frac{T_A}{T_S}\sqrt{\Delta\nu\tau}, \tag{1.7}$$

where C is a constant. This result appears to have been first obtained by Dicke (1946). More detailed examination [see, e.g., Tiuri (1964), Tiuri and Räisänen (1986)] shows that $C = 1$ for a simple power-law receiver with a rectangular passband, and varies by factors ~2 for more complicated systems. Typical values of $\Delta\nu$ and τ are 50 MHz and 5 h, which result in a value of 10^6 for the factor $(\Delta\nu\tau)^{1/2}$. As a result, it is possible to detect a signal for which the power level is little more than 10^{-6} times the system noise. A particularly effective use of long averaging time is found in the observations with the Cosmic Background Explorer satellite, in which it was possible to measure structure at a brightness temperature level less than 10^{-7} of the system temperature (Smoot et al. 1990, 1992). The following calculation may help to illustrate the low energies involved in radio astronomy. Consider a large radio telescope with a total collecting area of 10^4 m^2 pointed toward a radio source of flux density 1 mJy(= 10^{-3} Jy) and accepting signals over a bandwidth of 50 MHz. In 10^3 years the total energy accepted is about 10^{-7} J (1 erg), which is comparable to a few percent of the kinetic energy in a single falling snowflake. To detect the source with the same telescope, and a system temperature of 50 K, would require an observing time of about 5 min, during which time the energy received would be about 10^{-15} J.

1.3 DEVELOPMENT OF RADIO INTERFEROMETRY

Evolution of Synthesis Techniques

This section presents a brief history of interferometry in radio astronomy. As an introduction, the following list indicates some of the more important steps in the progress from the Michelson stellar interferometer to the development of multi-element, synthesis mapping arrays and VLBI:

1. *Michelson stellar interferometer.* This instrument introduced the technique of using two spaced receiving apertures, and the measurement of fringe amplitude to determine angular width (1890–1921).
2. *First astronomical observations with a two-element radio interferometer.* Ryle and Vonberg (1946), solar observations.
3. *Phase-switching interferometer.* First implementation of the voltage multiplying action of a correlator, which is the device used to combine the signals from two antennas (1952).
4. *Astronomical calibration.* Gradual accumulation during the 1950s and 1960s of accurate positions for small-diameter radio sources from optical identifications and other means. Observations of such sources enabled accurate calibration of interferometer baselines and instrumental phases.
5. *Early measurements of angular dimensions of sources.* Use of variable baseline interferometers (~1952 onward).
6. *Solar arrays.* Development of multi-antenna arrays of centimeter-wavelength tracking antennas that provided detailed maps and profiles of the solar disk (mid-1950s onward).
7. *Arrays of tracking antennas.* General movement from meter-wavelength, non-tracking antennas to centimeter-wavelength, tracking antennas. Development of multielement arrays with a separate correlator for each baseline (~1960s).
8. *Earth-rotation synthesis.* Introduced by Ryle with some precedents from solar mapping. Development of computers to control receiving systems and perform Fourier transforms required in mapping was an essential component (1962).
9. *Spectral line capability.* Introduced into radio interferometry (~1962).
10. *Development of image processing techniques.* Based on phase closure, nonlinear deconvolution and other techniques, as described in Chapters 10 and 11 (~1974 onward).
11. *Very-long-baseline interferometry (VLBI).* First observations 1967.
12. *Millimeter-wavelength instruments (~100–300 GHz).* Major developments mid-1980s onward.
13. *Orbiting VLBI (OVLBI).* U.S. Tracking and Data Relay Satellite System (TDRSS) experiment, 1986–88. HALCA satellite, 1997.
14. *Submillimeter-wavelength instruments (300 GHz–1 THz).* JCMT-CSO interferometer (1992–1996). Submillimeter Array of the Smithsonian Astrophysical Observatory and Academica Sinica of Taiwan, ~2001. Atacama Large Millimeter Array (ALMA), first decade of the twenty-first century.

Michelson Interferometer

Interferometric techniques in astronomy date back to the optical work of Michelson (1890, 1920) and of Michelson and Pease (1921), who were able to obtain sufficiently fine angular resolution to measure the diameters of some of the nearer and larger stars such as Arcturus and Betelgeuse. The basic similarity of the theory of radio and optical radiation fields was recognized early by radio astronomers, and optical experience has provided valuable precedents to the theory of radio interferometry.

As shown in Fig. 1.4, beams of light from a star fall upon two apertures and are combined in a telescope. The resulting stellar image has a finite width and is shaped by effects that include atmospheric turbulence, diffraction at the mirrors, and the bandwidth of the radiation. Maxima in the light intensity resulting from interference occur at angles θ for which the difference Δ in the path lengths from the star to the point at which the light waves are combined is an integral number of wavelengths at the effective center of the optical passband. If the angular width of the star is small compared with the spacing in θ between adjacent maxima, the image of the star is crossed by alternate dark and light bands, known as interference fringes. If, however, the width of the star is comparable to the spacing between maxima, one can visualize the resulting image as being formed by the superposition of images from a series of points across the star. The maxima and minima of the fringes from different points do not coincide, and the fringe amplitude is attenuated as shown in Fig. 1.4b. As a measure of the relative amplitude of the fringes, Michelson defined the *fringe visibility*, $\mathcal{V}_M$, as

$$\mathcal{V}_M = \frac{\text{intensity of maxima} - \text{intensity of minima}}{\text{intensity of maxima} + \text{intensity of minima}}. \tag{1.8}$$

Note that with this definition the visibility is normalized to unity when the intensity at the minima is zero, that is, when the width of the star is small compared with the fringe width. If the fringe visibility is measurably less than unity the star is said to be *resolved* by the interferometer. Let $I(l, m)$ be the two-dimensional intensity of the star, or of a source in the case of a radio interferometer. (l, m) are coordinates on the sky, with l measured parallel to the aperture spacing vector and m normal to it. The fringes provide resolution in a direction parallel to the aperture spacing only. In the orthogonal direction the response is simply proportional to the intensity integrated over solid angle. Thus the interferometer measures the intensity projected onto the l direction, that is, the one-dimensional profile $I_1(l)$ given by

$$I_1(l) = \int I(l, m)dm\,. \tag{1.9}$$

As will be shown in later chapters, the fringe visibility is proportional to the modulus of the Fourier transform of $I_1(l)$ with respect to the spacing of the apertures measured in wavelengths. Figure 1.5 shows the integrated profile I_1 for three simple models of a star or radio source and the corresponding fringe visibility as a function of u, the spacing of the interferometer apertures in units of the

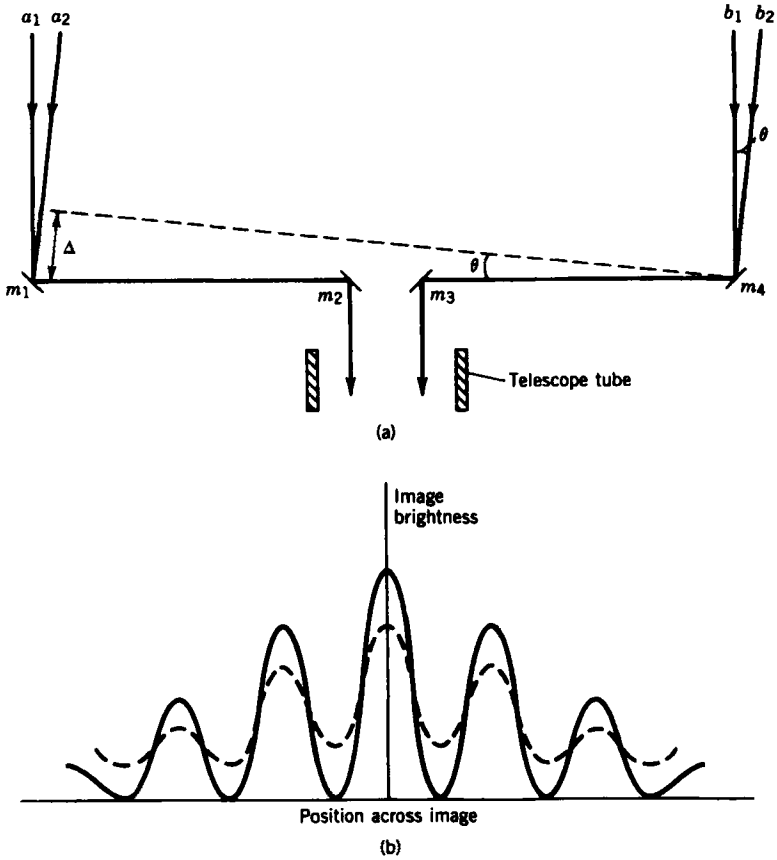


Figure 1.4 (a) Schematic diagram of the Michelson–Pease stellar interferometer. The incoming rays are guided into the telescope aperture by mirrors m_1 to m_4, of which the outer pair define the two apertures of the interferometer. Rays a_1 and b_1 traverse equal paths to the eyepiece at which the image is formed, but rays a_2 and b_2, which approach at an angle θ to the instrumental axis, traverse paths that differ by a distance Δ. (b) The intensity of the image as a function of position angle in a direction parallel to the spacing of the interferometer apertures. The solid line shows the fringe profiles for an unresolved star ($\mathcal{V}_M = 1.0$), and the broken line is for a partially resolved star for which $\mathcal{V}_M = 0.5$.

wavelength. At the top of the figure is a rectangular pillbox distribution, in the center a circular pillbox, and at the bottom a circular Gaussian function. The rectangular pillbox represents a uniformly bright rectangle on the sky with sides parallel to the l and m axes, and width a in the l direction. The circular pillbox represents a uniformly bright circular disk of diameter a. When projected onto the l axis the one-dimensional intensity function I_l has a semicircular profile. The Gaussian model is a circularly symmetric source with Gaussian taper of the intensity from the maximum at the center. The intensity is proportional to $\exp[-4 \ln 2\,(l^2 + m^2)/a^2]$, resulting in circular contours and a diameter a at the

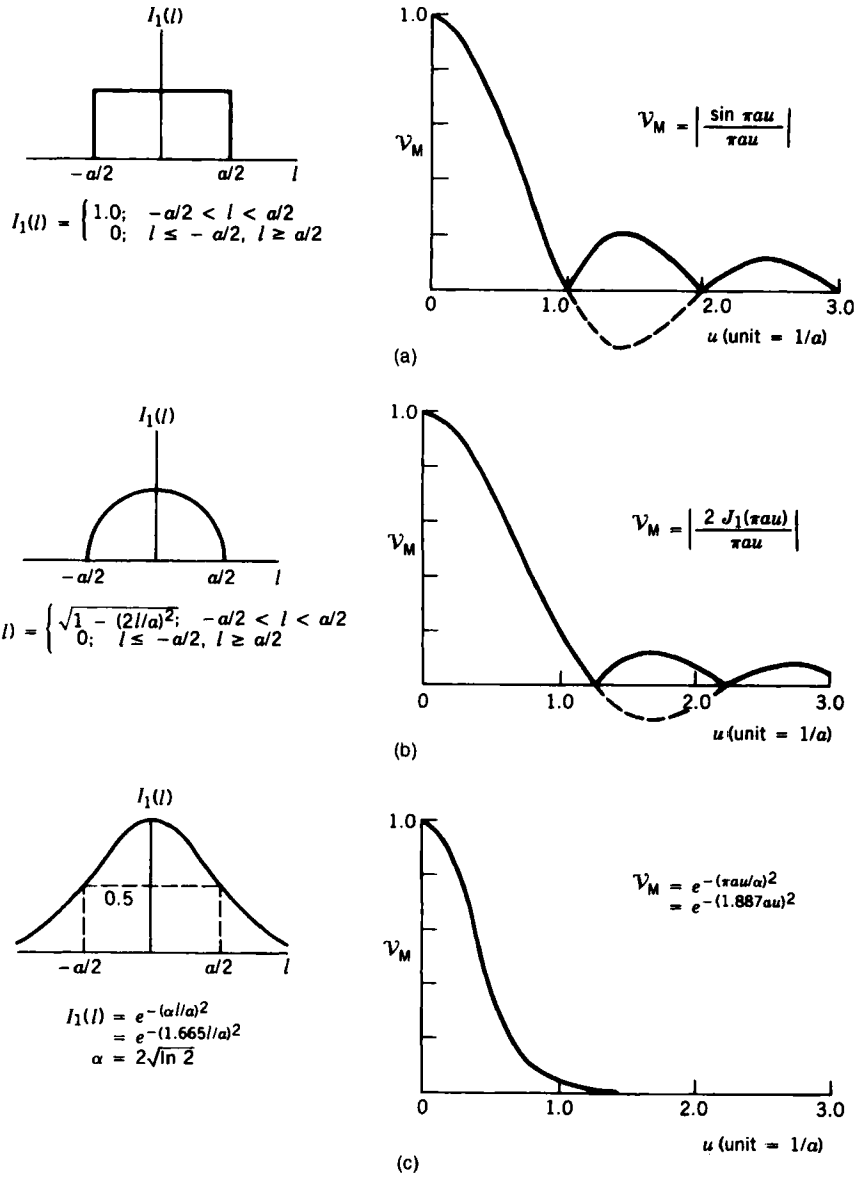


Figure 1.5 The one-dimensional intensity profiles $I_1(l)$ for three simple intensity models: (a) left, a uniform rectangular source; (b) left, a uniform circular source; (c) left, a circular Gaussian distribution. The corresponding Michelson visibility functions $\mathcal{V}_M$ are on the right. l is an angular variable on the sky, u is the spacing of the receiving apertures measured in wavelengths, and a is the characteristic angular width of the model. The solid lines in the curves of $\mathcal{V}_M$ indicate the modulus of the Fourier transform of $I_1(l)$, and the broken lines indicate negative values of the transform. See text for further explanation.

half-intensity level. Any slice through the model in a plane perpendicular to the (l, m) plane has a Gaussian profile with the same half-height width, a.

Michelson and Pease used mainly the circular disk model to interpret their observations and determined the stellar diameter by varying the aperture spacing of the interferometer to locate the first minimum in the visibility function. The adjustment of such an instrument and the visual estimation of $\mathcal{V}_M$ required great care, since the fringes were not stable but vibrated across the image in a random manner as a result of atmospheric fluctuations. The published results on stellar diameters measured with this method were never extended beyond the seven bright stars in Pease's (1931) list; for a detailed review see Hanbury Brown (1968). However, the use of electro-optical techniques now offers much greater instrumental capabilities in optical interferometry, as discussed in Section 16.4 of Chapter 16.

Early Two-Element Radio Interferometers

In 1946 Ryle and Vonberg constructed a radio interferometer to investigate cosmic radio emission, which had been discovered and verified by earlier investigators (Jansky 1933, Reber 1940, Appleton 1945, Southworth 1945). This interferometer used dipole antenna arrays at 175 MHz, with a baseline (i.e., the spacing between the antennas) that was variable between 10 and 140 wavelengths (17 and 240 m). A diagram of such an instrument and the type of record obtained are shown in Fig. 1.6. In this and most other meter-wavelength interferometers of the 1950s and 1960s, the antenna beams were pointed in the meridian and the rotation of the earth provided scanning in right ascension.

The receiver in Fig. 1.6 is sensitive to a narrow band of frequencies, and a simplified analysis of the response of the interferometer can be obtained in terms of monochromatic signals at the center frequency ν_0. We consider the signal from a radio source of very small angular diameter that is sufficiently distant that the incoming wavefront effectively lies in a plane. Let the signal voltage from the right-hand antenna in Fig. 1.6 be represented by $V \sin(2\pi \nu_0 t)$. The longer path length to the left-hand antenna introduces a time delay $\tau = (D/c) \sin\theta$, where D is the antenna spacing, θ is the angular position of the source, and c is the velocity of light. Thus, the signal from the left-hand antenna is $V \sin[2\pi \nu_0 (t - \tau)]$. The detector of the receiver generates the squared sum of the two signal voltages:

$$\{V \sin(2\pi \nu_0 t) + V \sin[2\pi \nu_0 (t - \tau)]\}^2. \tag{1.10}$$

The output of the detector contains a lowpass filter that removes any frequencies greater than a few hertz or tens of hertz, so in expanding (1.10) we can ignore terms in harmonics of $2\pi \nu_0 t$, which represent radio frequencies. The detector output is therefore

$$F = V^2[1 + \cos(2\pi \nu_0 \tau)]. \tag{1.11}$$

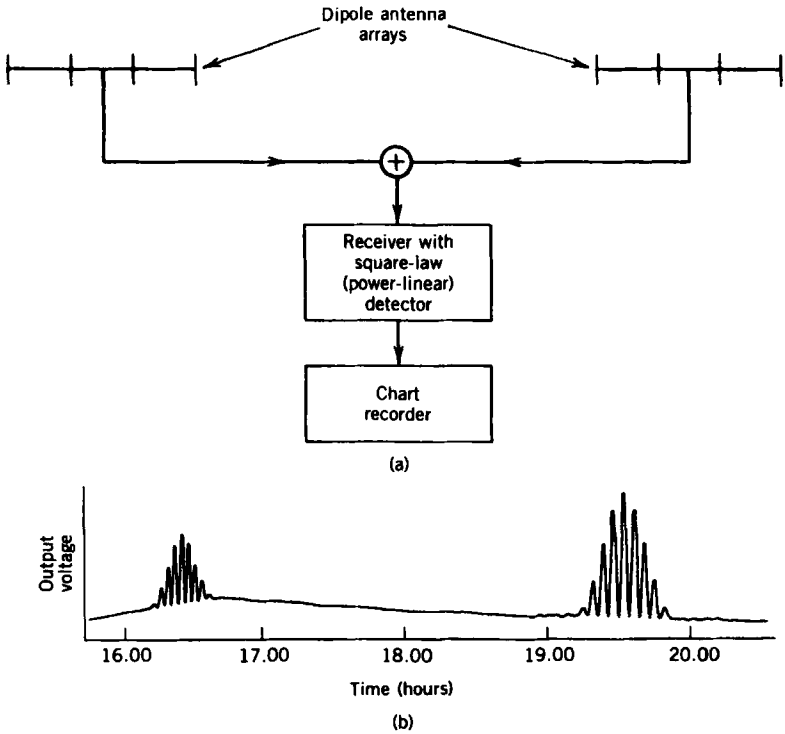


Figure 1.6 (a) Simple interferometer in which the signals are combined additively. (b) Record from such an interferometer with east–west antenna spacing. The ordinate is the total power received and the abscissa is time. The source at the left is Cygnus A and the one at the right Cassiopeia A. The increase in level near Cygnus A results from the galactic background radiation, which is concentrated toward the plane of our Galaxy but is completely resolved by the interferometer fringes. The record is from Ryle (1952).

Because τ varies only slowly as the earth rotates, the frequency represented by $\cos(2\pi \nu_0 \tau)$ is not filtered out. In terms of the source position, θ, we have

$$F = V^2 \left[1 + \cos\left(\frac{2\pi \nu_0 D \sin\theta}{c} \right) \right]. \tag{1.12}$$

Thus as the source moves across the sky, the output fringe pattern F varies between 0 and V^2, as shown by the sources in Fig. 1.6b. The response is modulated by the beam pattern of the antennas, of which the maximum is pointed in the meridian. The cosine function in Eq. (1.12) represents the Fourier component of the source brightness to which the interferometer responds. The angular width of the fringes is less than the angular width of the antenna beam by (approximately) the ratio of the width of an antenna to the baseline D, which in this example is about 1/10. The use of an interferometer instead of a single antenna results in a corresponding increase in precision in determining the time of transit of the

source. The form of the fringe pattern in Eq. (1.12) also applies to the Michelson interferometer in Fig. 1.4.

Sea Interferometer

A different implementation of interferometry, known as the sea interferometer, or Lloyd's mirror interferometer (Bolton and Slee 1953), was provided by a number of horizon-pointing antennas near Sydney, Australia. These had been installed for radar during World War II at several coastal locations, at elevations of 60–120 m above the sea. Radiation from sources rising over the eastern horizon was received both directly and by reflection from the sea, as shown in Fig. 1.7. The frequencies of the observations were in the range 40–400 MHz, the middle part of the range being the most satisfactory because of ionospheric effects at lower frequencies and sea roughness at higher frequencies. The sudden appearance of a rising source was useful in separating individual sources. Because of the reflected wave, the power received at the peak of a fringe was four times that for direct reception with the single antenna, and twice that of an adding interferometer (Fig. 1.6a) with two of the same antennas. Observations of the sun by McCready, Pawsey, and Payne-Scott (1947) using this system provided the first published record of interference fringes in radio astronomy. Observations of the source Cygnus A by Bolton and Stanley (1948) provided the first positive evidence of the existence of a discrete non-solar radio source. Thus the sea interferometer played an important part in early radio astronomy, but the effects of the long atmospheric paths and the roughness of the sea surface precluded further useful development.

Phase-Switching Interferometer

A problem with the interferometer systems in both Figs. 1.6 and 1.7 is that in addition to the signal from the source, the output of the receiver contains components from other sources of noise power such as the galactic background radiation, thermal noise from the ground picked up in the antenna sidelobes, and the noise generated in the amplifiers of the receiver. For all except the few strongest cosmic sources, the component from the source is several orders of magnitude less than the total noise power in the receiver. Thus a large offset has been removed from the records shown in Figs. 1.6b and 1.7b. This offset is proportional to the receiver gain, changes in which are difficult to eliminate entirely. The resulting drifts in the output level degrade the detectability of weak sources and the accuracy of measurement of the fringes. With the technology of the 1950s, the receiver output was usually recorded on a paper chart, and could be lost when baseline drifts caused the recorder pen to go off scale.

The introduction of *phase switching* by Ryle (1952), which removed the unwanted components of the receiver output leaving only the fringe oscillations, was the most important technical improvement in early radio interferometry. If V_1 and V_2 represent the signal voltages from the two antennas, the output from the simple adding interferometer is proportional to $(V_1 + V_2)^2$. In the phase-switching system, shown in Fig. 1.8, the phase of one of the signals is periodically reversed,

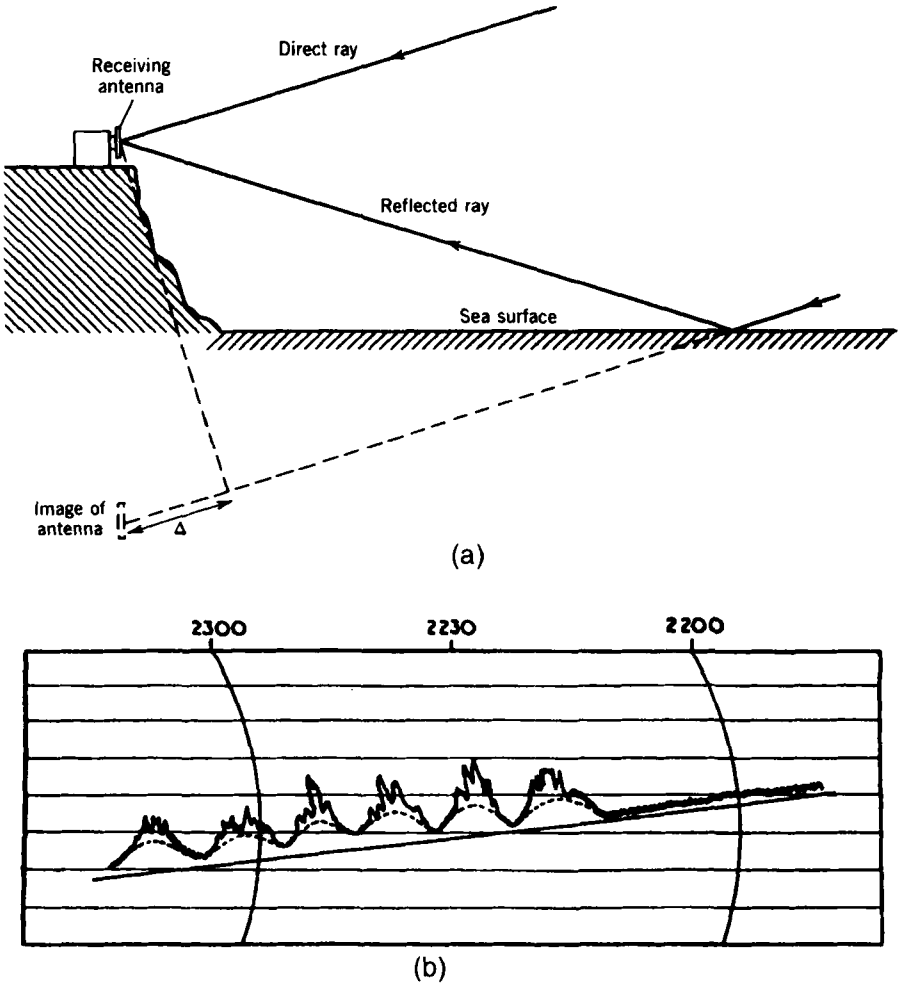


Figure 1.7 (a) Schematic diagram of a sea interferometer. The fringe pattern is similar to that which would be obtained with the actual receiving antenna and one at the position of its image in the sea. The reflected ray undergoes a phase change of 180° on reflection and travels an extra distance Δ in reaching the receiving antenna. (b) Sea interferometer record of the source Cygnus A at 100 MHz by Bolton and Stanley (1948). The source rose above the horizon at approximately 22.17. The broken line was inserted to show that the record could be interpreted in terms of a steady component and a fluctuating component of the source; the fluctuations were later shown to be of ionospheric origin. The fringe width was approximately 1.0° and the source is unresolved, that is, its angular width is small in comparison with the fringes. Part (b) is reprinted by permission from *Nature*, Vol. 161, No. 4087, p. 313; copyright ©1948 Macmillan Journals Limited.

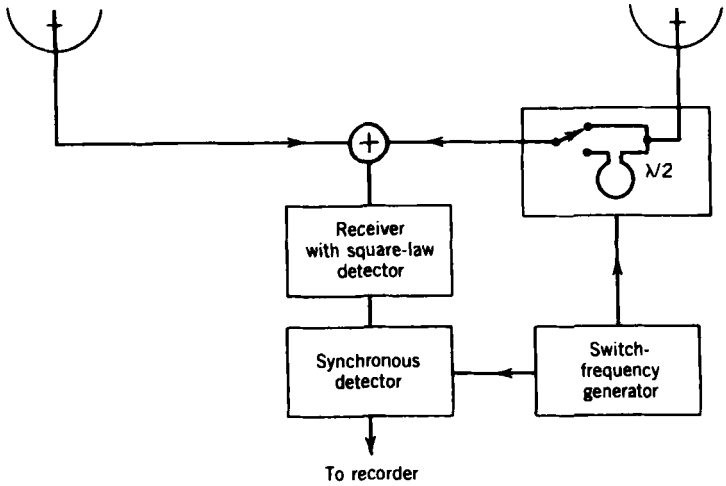


Figure 1.8 Phase-switching interferometer. The signal from one antenna is periodically reversed in phase, indicated here by switching an additional half wavelength of path into the transmission line.

so the output of the detector alternates between $(V_1 + V_2)^2$ and $(V_1 - V_2)^2$. The frequency of the switching is a few tens of hertz, and a synchronous detector takes the difference between the two output terms, which is proportional to $V_1 V_2$. Thus the output of a phase-switching interferometer is the time average of the product of the signal voltages; that is, it is proportional to the cross-correlation of the two signals. The circuitry that performs the multiplication and averaging of the signals in a modern interferometer is known as a *correlator*; a more general definition of a correlator will be given later. Comparison with the output of the system in Fig. 1.6 shows that if the signals from the antennas are multiplied instead of added and squared, then the constant term within the square brackets in Eq. (1.12) disappears and only the cosine term remains. The output consists of the fringe oscillations only, as shown in Fig. 1.9. With the reduction in the sensitivity

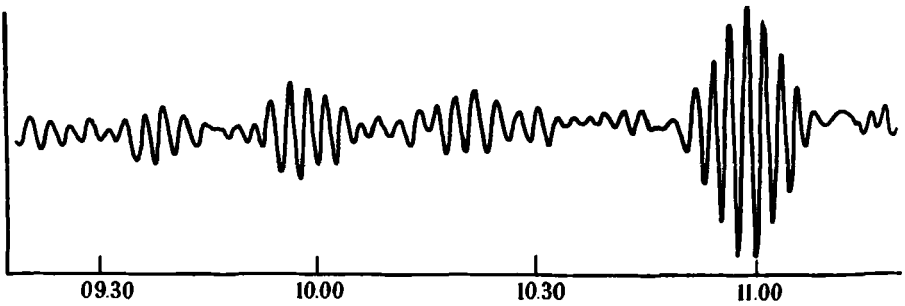


Figure 1.9 Output of a phase-switching interferometer as a function of time showing the response to a number of sources. From Ryle (1952).

to instrumental gain variation, it became practicable to install amplifiers at the antennas to overcome attenuation in the transmission lines. This advance resulted in the use of longer antenna spacings and larger arrays. Most interferometers from about 1950 onward incorporated phase switching, which provided the first means of implementing the action of a correlator. It is no longer necessary to use phase switching to obtain the voltage-multiplying action, but it is often included to help eliminate various instrumental imperfections, as described in Section 7.5.

Optical Identifications and Calibration Sources

Interferometer observations by Bolton and Stanley (1948), Ryle and Smith (1948), Ryle, Smith, and Elsmore (1950), and others provided evidence of numerous discrete sources. Identification of the optical counterparts of these required accurate measurement of radio positions. The principal method then in use for position measurement with interferometers was to determine the time of transit of the central fringe using an east–west baseline, and also the frequency of the fringe oscillations, which is proportional to the cosine of the declination. The measurement of position is only as accurate as the knowledge of the interferometer fringe pattern, which is determined by the relative locations of the electrical centers of the antennas. In addition, any inequality in the electrical path lengths in the cables and amplifiers from the antennas to the point where the signals are combined introduces an instrumental phase term, which offsets the fringe pattern. Smith (1952a) obtained positions for four sources with rms errors as small as ± 20 arcsec in right ascension and ± 40 arcsec in declination, and gave a detailed analysis of the accuracy that was attainable. The optical identification of Cygnus A and Cassiopeia A by Baade and Minkowski (1954a,b) was a direct result of improved radio positions by Smith (1951) and Mills (1952). Cygnus A proved to be a distant galaxy and Cassiopeia A a supernova remnant, but the interpretation of the optical observations was not fully understood at the time.

The need for absolute calibration of the antennas and receiving system rapidly disappeared after a number of compact radio sources were identified with optical objects. Optical positions accurate to ~ 1 arcsec could then be used, and observations of such sources enabled calibration of interferometer baseline parameters and fringe phases. Although it cannot be assumed that the radio and optical positions of a source coincide exactly, the offsets for different sources are randomly oriented. Thus errors were reduced as more calibration sources became available. Another important way of obtaining accurate radio positions during the 1960s and 1970s was by observation of occultation of sources by the moon, which is described in Section 16.2.

Early Measurements of Angular Width

Comparison of the angular widths of radio sources with the corresponding dimensions of their optical counterparts helped in some cases to confirm identifications, as well as to provide important data for physical models of the emission processes. In the simplest procedure, measurements of the fringe amplitude are interpreted in terms of intensity models such as those shown in Fig. 1.5. The peak-to-peak

fringe amplitude for a given spacing normalized to the same quantity when the source is unresolved provides a measure of the fringe visibility equivalent to the definition in Eq. (1.8).

Some of the earliest measurements were made by Mills (1953), who used an interferometer operating at 101 MHz, in which a small transportable array of Yagi elements could be located at distances up to 10 km from a larger antenna. The signal from this remote antenna was transmitted back over a radio link, and fringes were formed. Smith (1952b,c), at Cambridge, England, also measured the variation of fringe amplitude with antenna spacing, but used shorter baselines than did Mills and concentrated on precise measurements of small changes in the fringe amplitude. Results by both investigators provided dimensions of a number of the strongest sources: Cassiopeia A, the Crab nebula, NGC4486 (Virgo A), and NGC5128 (Centaurus A).

A third early group working on angular widths at the Jodrell Bank Experimental Station,[‡] England, used a different technique: *intensity interferometry* (Jennison and Das Gupta 1953, 1956; Jennison 1994). Hanbury Brown and Twiss (1954) had shown that if the signals received by two spaced antennas are passed through square-law detectors, the fluctuations in the intensity that result from the Gaussian fluctuations in the received field strength are correlated. The degree of correlation varies in proportion to the square of the visibility that would be obtained in a conventional interferometer in which signals are combined before detection. The intensity interferometer has the advantage that it is not necessary to preserve the radio-frequency phase of the signals in bringing them to the location at which they are combined. This simplifies the use of long baselines, which in this case extended up to 10 km. A VHF radio link was used to transmit the detected signal from the remote antenna, for measurement of the correlation. The disadvantage of the intensity interferometer is that it requires a high signal-to-noise ratio, and even for Cygnus A and Cassiopeia A, the two highest flux density sources in the sky, it was necessary to construct large arrays of dipoles, which operated at 125 MHz. The intensity interferometer is discussed further in Section 16.1, but it has been of only limited use in radio astronomy because of its lack of sensitivity.

The most important result of these intensity interferometer measurements was the discovery that for Cygnus A the fringe visibility for the east–west intensity profile falls close to zero and then increases to a secondary maximum as the antenna spacing is increased. Two symmetric source models were consistent with the visibility values derived from the measurements. These were a two-component model in which the phase of the fringes changes by 180° in going through the minimum, and a three-component model in which the phase does not change. The intensity interferometer gives no information on the fringe phase, so a subsequent experiment was made by Jennison and Latham (1959) using conventional interferometry. Because the instrumental phase of the equipment was not stable enough to permit calibration, three antennas were used and three sets

[‡]Later known as the Nuffield Radio Astronomy Laboratories, and since 1999 as the Jodrell Bank Observatory.

of fringes for the three pair combinations were recorded simultaneously. If ϕ_{mn} is the phase of the fringe pattern for antennas m and n, it is easy to show that at any instant the combination

$$\phi_{123} = \phi_{12} + \phi_{23} + \phi_{31} \tag{1.13}$$

is independent of instrumental and atmospheric phase effects and is a measure of the corresponding combination of fringe phases (Jennison 1958). By moving one antenna at a time it was found that the phase does indeed change by approximately 180° at the visibility minimum, and therefore that the two-component model in Fig. 1.10 is the appropriate one. The use of combinations of simultaneous visibility measurements typified by Eq. (1.13), now referred to as *closure relationships*, became important about 20 years later in image processing techniques. Closure relationships and the conditions under which they apply are discussed in Section 10.3.

The results on Cygnus A demonstrated that the simple models of Fig. 1.5 are not generally satisfactory for representation of radio sources. To determine even the most basic structure, it is necessary to measure the fringe visibility at spacings well beyond the first minimum of the visibility function to detect multiple components, and to make such measurements at a number of position angles across the source.

An early interferometer aimed at achieving high angular resolution with high sensitivity was developed by Hanbury Brown, Palmer, and Thompson (1955) at the Jodrell Bank Experimental Station, England. This interferometer used an offset local oscillator technique that took the place of a phase switch and also enabled the frequency of the fringe pattern to be slowed down to within the response time of the chart recorder used to record the output. A radio link was used to bring the signal from the distant antenna. Three sources were found to have diameters less than 12 arcsec using spacings up to 20 km at 158 MHz observing frequency (Morris, Palmer, and Thompson 1957). During the 1960s this instrument was extended to achieve resolution of less than 1 arcsec and greater sensitivity (Elgaroy,

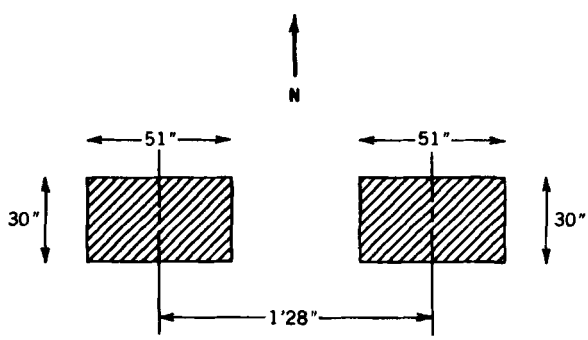


Figure 1.10 Two-component model of Cygnus A derived by Jennison and Das Gupta (1953) using the intensity interferometer. Reprinted by permission from *Nature*, Vol. 172, No. 4387, p. 996; copyright ©1953 Macmillan Journals Limited.

Morris, and Rowson 1962). The program later led to the development of a multielement, radio-linked interferometer known as the MERLIN array (Thomasson 1986).

Survey Interferometers and the Mills Cross

In the middle 1950s the thrust of much of the work was toward cataloging larger numbers of sources with positions of sufficient accuracy to allow optical identification. The instruments operated mainly at meter wavelengths, where the spectrum was then much less heavily crowded with man-made emissions. A large interferometer at Cambridge used four antennas located at the corners of a rectangle 580 m east–west by 49 m north–south (Ryle and Hewish 1955). This arrangement provided both east–west and north–south fringe patterns for measurement of right ascension and declination.

A different type of survey instrument was developed by Mills et al. (1958) at Fleurs, near Sydney, consisting of two long, narrow antenna arrays in the form of a cross, as shown in Fig. 1.11. Each array produced a *fan beam*, that is, a beam that is narrow in a plane containing the long axis of the array and wide in the orthogonal direction. The outputs of these two arrays were combined in a phase-switching receiver, and the voltage-multiplying action produced a power response pattern equal to the product of the voltage responses of the two arrays. This combined response had the form of a narrow *pencil beam*. The two arrays had a common electrical center, so there were no interferometer fringes. The arrays were 457 m long, and the cross produced a beam of width 49 arcmin and approximately circular cross section at 85.5 MHz. The beam pointed in the meridian and could be steered in elevation by adjusting the phase of the dipoles in the north–south arm. The sky survey made with this instrument provided a list of over 2200 sources.

A comparison of the source catalogs from the Mills cross with those from the Cambridge interferometer, which initially operated at 81.5 MHz (Shakeshaft et al. 1955), showed poor agreement between the source lists for a common area of sky (Mills and Slee 1957). The discrepancy was found to result principally from the occurrence of *source confusion* in the Cambridge observations. When two or more sources are simultaneously present within the antenna beams, they

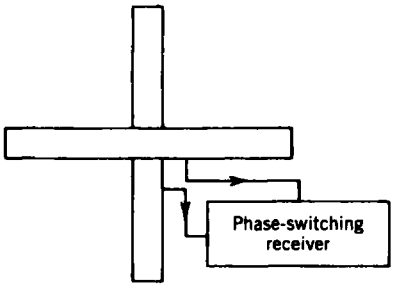


Figure 1.11 Simplified diagram of the Mills cross radio telescope. The cross-shaped area represents the apertures of the two antennas.

produce fringe oscillations with slightly different frequencies, resulting from differences in the source declinations. Maxima in the fringe amplitude, which occur when the fringe components happen to combine in phase, can mimic responses to sources. This was a serious problem because the beams of the interferometer antennas were too wide, a problem that did not arise in the Mills cross, which was designed to provide the required resolution for accurate positions in the single pencil beam. The frequency of the Cambridge interferometer was later increased to 159 MHz, thereby reducing the solid angles of the beams by a factor of 4, and a new list of 471 sources was rapidly compiled (Edge et al. 1959). This was the 3C survey (source numbers are preceded by 3C, indicating the third Cambridge catalog), a revised version of which (Bennett 1962) became a cornerstone of radio astronomy for the following decade. To avoid confusion problems with these types of instruments, some astronomers subsequently recommended that the density of sources cataloged should not, on average, exceed one in about twenty times the solid angle of the antenna beams (Pawsey 1958, Hazard and Walsh 1959).

In the 1960s a generation of new and larger survey instruments began to appear. Two such instruments developed at Cambridge are shown in Fig. 1.12. One was an interferometer with one antenna elongated in the east–west direction and the other north–south, and the other was a large T-shaped array which had characteristics similar to those of a cross, as explained in Section 5.3. In each of these instruments the north–south element was not constructed in full, but the response with such an aperture was synthesized by using a small antenna that was moved in steps to cover the required aperture; a different position was used for each 24-h scan in right ascension (Ryle, Hewish, and Shakeshaft 1959; Ryle and Hewish 1960). The records from the various positions were combined by computer to synthesize the response with the complete north–south aperture. An analysis of these instruments is given by Blythe (1957). The large interferometer produced

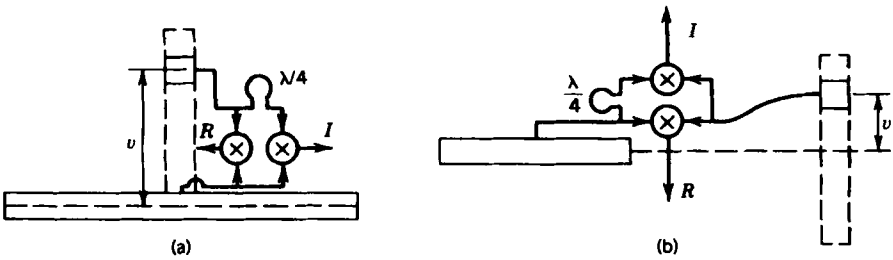


Figure 1.12 Schematic diagrams of two instruments, in each of which a small antenna is moved to different positions between successive observations to synthesize the response that would be obtained with a full aperture corresponding to the rectangle shown by the broken line. The arrangement of two signal-multiplying correlators producing real (R) and imaginary (I) outputs is explained in Section 6.1 under *Simple and Complex Correlators*. Instruments of both types, the T-array (a), and the two-element interferometer (b), were constructed at the Mullard Radio Astronomy Observatory, Cambridge, England.

the 4C (fourth Cambridge) catalog containing over 4800 sources (Gower, Scott, and Wills 1967). At Molonglo in Australia a larger Mills cross (Mills et al. 1963) was constructed with arrays 1 mile long producing a beam of 2.8 arcmin width at 408 MHz. The development of the Mills cross is described in papers by Mills and Little (1953), Mills (1963), and Mills et al. (1958, 1963). Crosses of comparable dimensions located in the northern hemisphere included one at Bologna, Italy (Braccesi et al. 1969), and one at Serpukhov in the Russian Federation (Vitkevich and Kalachev 1966).

Centimeter-Wavelength Solar Mapping

A number of instruments have been designed specifically for mapping the sun. The antennas are usually paraboloid reflectors mounted to track the sun, but since the sun is a strong radio source, the apertures do not have to be very large. Figure 1.13a shows an array of antennas from which the signals at the receiver input are aligned in phase when the angle θ between the direction of the source and a plane normal to the line of the array is such that $\ell_\lambda \sin\theta$ is an integer, where ℓ_λ is the unit antenna spacing measured in wavelengths. This type of array is sometimes referred to as a grating array, since it forms a series of fan-shaped beams, narrow in the θ direction, in a manner analogous to the response of an optical diffraction grating. Christiansen and Warburton (1955) obtained a two-dimensional map of the quiet sun at 21-cm wavelength using both east–west and north–south grating arrays. These arrays consisted of 32 (east–west) and 16 (north–south) uniformly spaced, paraboloid antennas. As the sun moved through the sky it was scanned at different angles by the different beams, and a two-

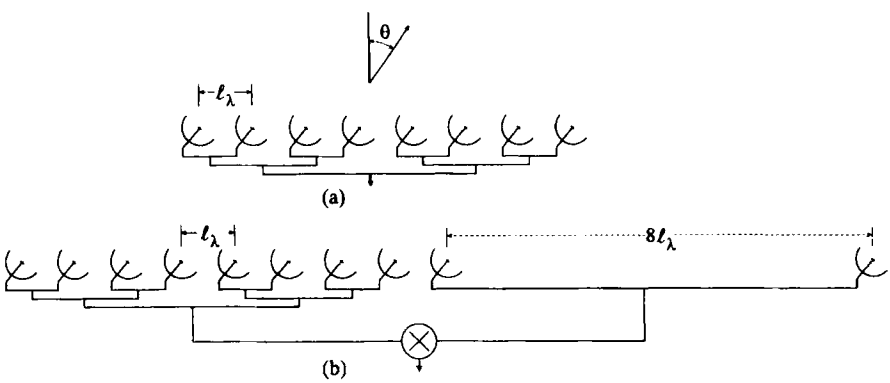


Figure 1.13 (a) A linear array of eight equally spaced antennas connected by a branching network in which the electrical path lengths from the antennas to the receiver input are equal. This arrangement is sometimes referred to as a grating array, and in practice there are usually 16 or more antennas. (b) An eight-element grating array combined with a two-element array to enhance the angular resolution. A phase-switching receiver, indicated by the multiplication symbol, is used to form the product of the signal voltages from the two arrays. The receiver output contains the simultaneous responses of antenna pairs with 16 different spacings. Systems of this general type are known as compound interferometers.

dimensional map could be synthesized by Fourier analysis of the scan profiles. To obtain a sufficient range of scan angles, observations extending over 8 months were used. In later instruments for solar mapping it was generally necessary to be able to make a complete map within a day to study the variation of enhanced solar emission associated with active regions. Several instruments used grating arrays, typically containing 16 or 32 antennas, and crossed in the manner of a Mills cross. Crossed grating arrays produce a rectangular matrix pattern of beams on the sky, and the rotation of the earth enables sufficient scans to be obtained to provide daily maps of active regions and other features. Instruments of this type included crosses at 21-cm wavelength at Fleurs, Australia (Christiansen and Mullaly 1963), and at 10-cm wavelength at Stanford, California (Bracewell and Swarup 1961), and a T-shaped array at 1.9-m wavelength at Nançay, France (Blum, Boischot, and Ginat 1957; Blum 1961). These were the earliest mapping arrays with large numbers ($\sim$16 or more) antennas.

Figure 1.13b illustrates the principle of a configuration known as a compound interferometer (Covington and Broten 1957), which was used to enhance the performance of a grating array or other antenna with high angular resolution in one dimension. The system shown consists of the combination of a grating array with a two-element array. An examination of Fig. 1.13b shows that pairs of antennas, chosen one from the grating array and one from the two-element array, can be found for all spacings from one to sixteen times the unit spacing ℓ_λ. In comparison, the grating array alone provides only one to seven times the unit spacing, so the number of different spacings simultaneously contributing to the response is increased by a factor of more than 2 by the addition of two more antennas. Arrangements of this type were used to increase the angular resolution of one-dimensional scans of strong sources (Picken and Swarup 1964, Thompson and Krishnan 1965). By combining a grating array with a single larger antenna it was also possible to reduce the number of grating responses on the sky (Labrum et al. 1963). Both the crossed grating arrays and the compound interferometers were originally operated with phase-switching receivers to combine the outputs of the two subarrays. In modern implementations the signal from each antenna is converted to an intermediate frequency (IF), and a separate voltage-multiplying correlator is used for each spacing. This allows further possibilities in arranging the antennas to maximize the number of different antenna spacings, as discussed in Section 5.5.

Measurements of Intensity Profiles

Continuing measurements of the structure of sources indicated that in general the intensity profiles are not symmetrical, so their Fourier transforms, and hence the visibility functions, are complex. This will be explained in detail in later chapters, but at this point we note that it means that the phase of the fringe pattern, as well as the amplitude, varies with antenna spacing and must be measured to allow the intensity profiles to be recovered. To accommodate both fringe amplitude and phase, visibility is expressed as a complex quantity. Measurement of the fringe phase became possible in the 1960s and 1970s, by which time a number of com-

pact sources with well-determined positions, suitable for calibration of the fringe phase, were available. Electronic phase stability had also improved, and computers were available for recording and processing the output data. Improvements in antennas and receivers enabled measurements to be made at wavelengths in the centimeter range (frequencies greater than ~1 GHz), using tracking antennas.

An interferometer at the Owens Valley Radio Observatory, California (Read 1961), provides a good example of one of the earliest instruments used extensively for determining radio structure. It consists of two 27.5-m-diameter paraboloid antennas on equatorial mounts with a rail track system that allows the spacing between them to be varied by up to 490 m in both the east–west and north–south directions. It has been used mainly at frequencies from 960 MHz to a few gigahertz. Studies by Maltby and Moffet (1962) and Fomalont (1968) illustrate the use of this instrument for measurement of intensity distributions, an example of which is shown in Fig. 1.14.

Spectral Line Interferometry

By the early 1960s the interferometer at the Owens Valley and several others had been fitted with spectral line receiving systems. The passband of each receiver is divided into a number of channels by a filter bank, usually in the IF (intermediate frequency) stages, and for each channel the signals from the two antennas go to a separate correlator. Alternatively, the IF signals are digitized and the filtering is performed digitally as described in Section 8.7. The width of the channels should ideally be less than that of the line to be observed so that the line profile can be studied. Spectral line interferometry allows the distribution of the line emission across a radio source to be examined. Roger et al. (1973) describe an array in Canada built specifically for observations in the 1420 MHz (21-cm wavelength) line of neutral hydrogen.

Spectral lines can also be observed in absorption, especially in the case of the neutral hydrogen line. At the line frequency the gas absorbs the continuum radiation from any more distant source that is observed through it. Comparison of the emission and absorption spectra of neutral hydrogen yields information on its temperature and density. Measurement of absorption spectra of sources can be made using single antennas, but in such cases the antenna also responds to the broadly distributed emitting gas within the antenna beam. The absorption spectra for weak sources are difficult to separate from the line emission. With an interferometer, the broad emission features on the sky are almost entirely resolved and the absorption spectrum can be observed directly. For examples of hydrogen line absorption, see Clark, Radhakrishnan, and Wilson (1962) and Hughes, Thompson, and Colvin (1971), and for absorption in the 4.8-GHz formaldehyde line, Moore and Marscher (1995).

Earth-Rotation Synthesis Mapping

A very important step in the development of synthesis imaging was the use of the variation of the antenna baseline provided by the rotation of the earth. Fig-

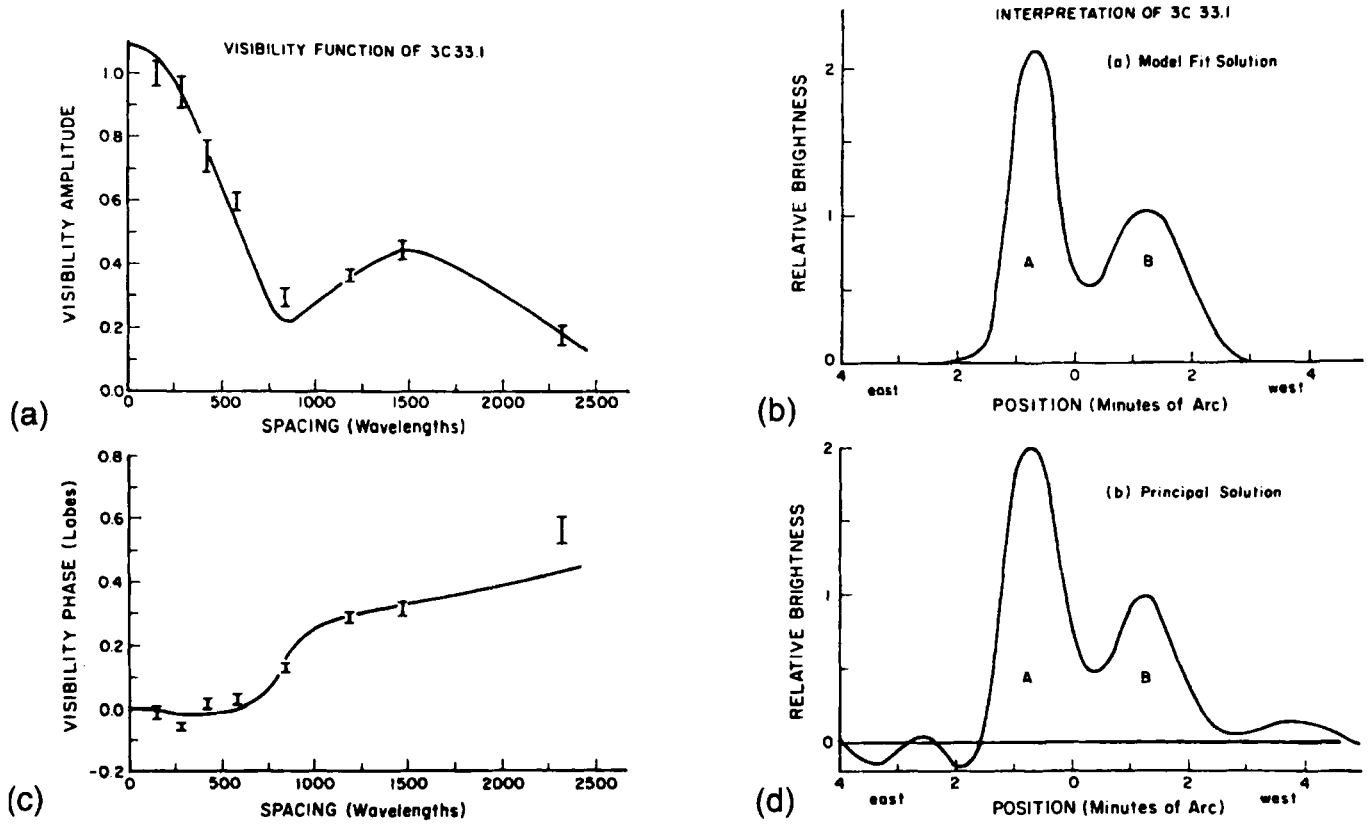


Figure 1.14 Example of interferometer measurements of one-dimensional intensity (brightness): the east–west profile of source 3C33.1 as determined by Fomalont (1968) using the interferometer at the Owens Valley Radio Observatory at 1425 MHz. (a, c) The points show the measured amplitude and phase of the visibility. (b) The profile was obtained by fitting Gaussian components to the visibility data, as shown by the curves through the measured visibility points. (d) The profile was obtained by Fourier transformation of the observed visibility values. The unit of visibility phase is 2π radians.

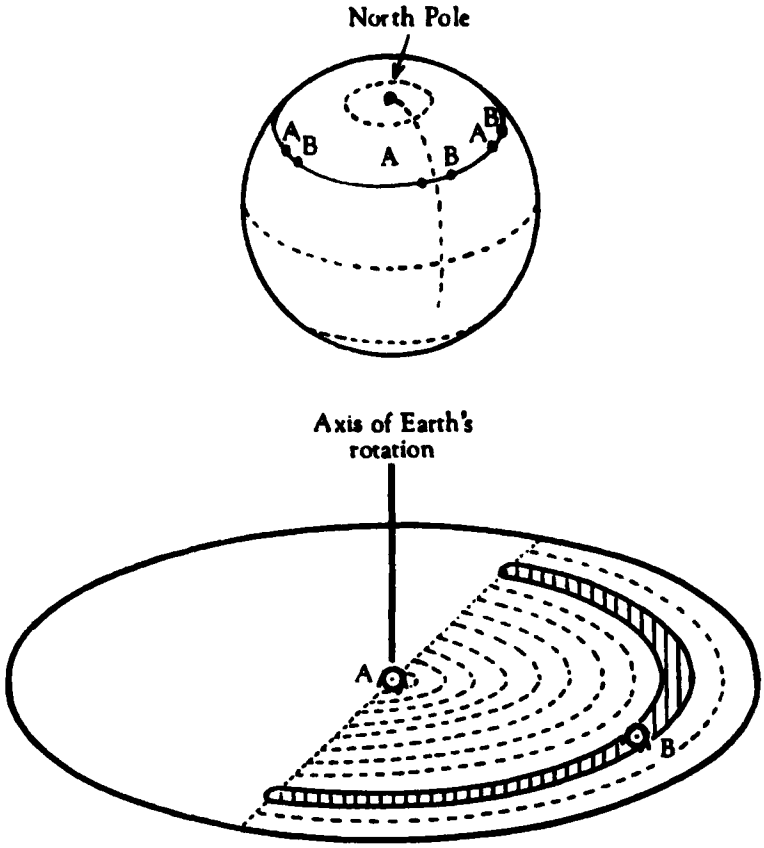


Figure 1.15 Use of earth rotation in synthesis mapping as explained by Ryle (1962). The antennas A and B are spaced on an east–west line. By varying the distance between the antennas from one day to another, and observing for 12 h with each configuration, it is possible to encompass all the spacings from the origin to the elliptical outer boundary of the lower diagram. Only 12 h of observing are required, since during the other 12 h the spacings covered are identical but the positions of the antennas are effectively interchanged. Reprinted by permission from *Nature*, Vol. 194, No. 4828, p. 517; copyright ©1962 Macmillan Journals Limited.

ure 1.15 illustrates this principle as described by Ryle (1962). For a source at a high declination, the position angle of the baseline projected onto a plane normal to the direction of the source rotates through 180° in 12 h. Thus if the source is tracked across the sky for a series of 12-h periods, each one with a different antenna spacing, the required two-dimensional visibility data can be collected while the antenna spacing is varied in one dimension only.

The Cambridge One-Mile Radio Telescope was the first instrument designed to exploit fully the earth-rotation technique and apply it to a large number of radio sources. The use of earth rotation was not a sudden development in radio astronomy, and had been used in solar studies for a number of years. As

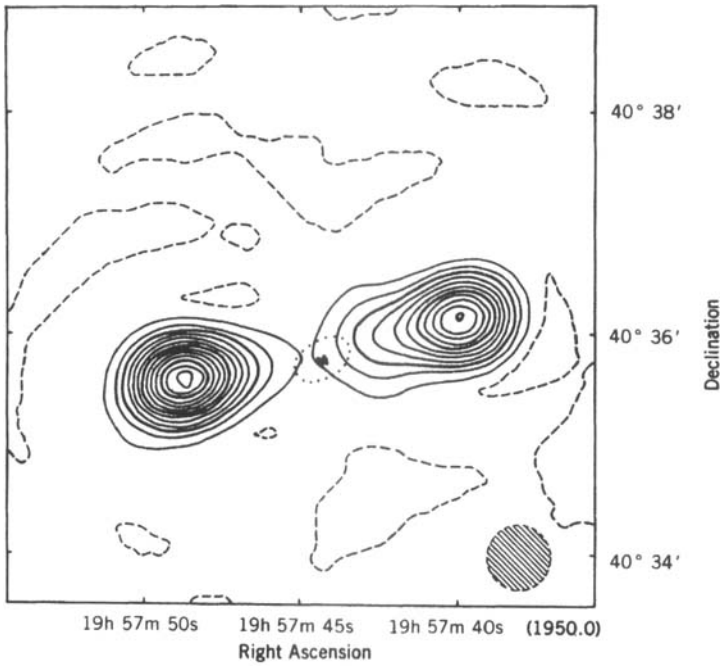


Figure 1.16 Map of the source Cygnus A, which was one of the first results (Ryle, Elsmore, and Neville 1965) from the Cambridge One-Mile Telescope using the earth-rotation principle shown in Fig. 1.15. The frequency is 1.4 GHz. The map has been scaled in declination so that the half-power beam contour is circular, as shown by the shaded area in the lower right corner. The dotted ellipse shows the outer boundary of the optical source, and its central structure is also indicated. Reprinted by permission from *Nature*, Vol. 205, No. 4978, p. 1260; ©1965 Macmillan Journals Limited.

noted earlier, Christiansen and Warburton (1955) had obtained a two-dimensional map of the sun, using tracking antennas in two grating arrays. At Jodrell Bank, Rowson (1963) had used a two-element interferometer with tracking antennas to map strong non-solar sources. Also, Ryle and Neville (1962) had mapped the north polar region using earth rotation to demonstrate the technique. However, the first maps published from the Cambridge One-Mile telescope, those of the strong sources Cassiopeia A and Cygnus A (Ryle, Elsmore, and Neville 1965), exhibited a degree of structural detail unprecedented in earlier studies and heralded the development of synthesis mapping. The map of Cygnus A is shown in Fig. 1.16.

Development of Synthesis Arrays

Following the success of the Cambridge One-Mile Telescope, interferometers such as the NRAO instrument at Green Bank, West Virginia (Hogg et al. 1969), were rapidly adapted for synthesis mapping. Several large arrays designed to pro-

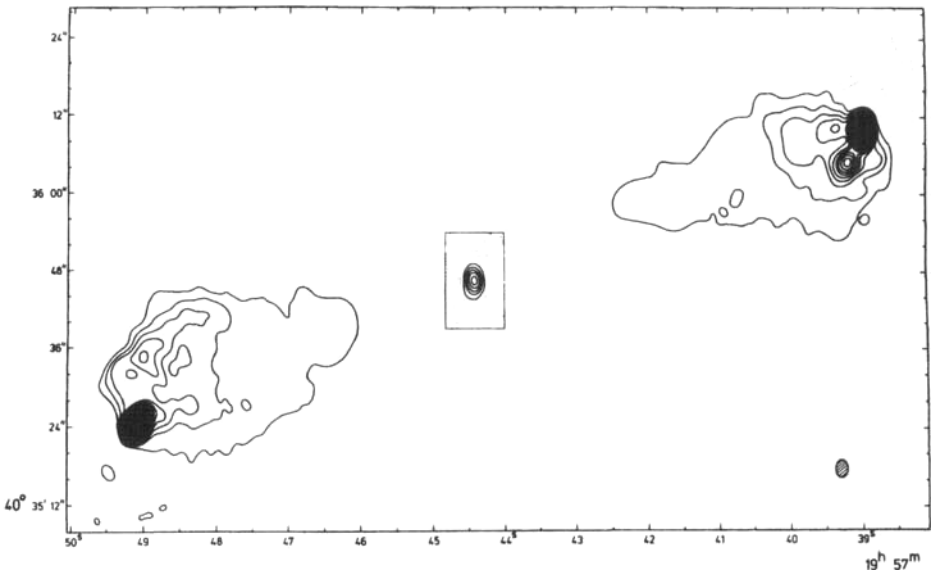


Figure 1.17 Map of the source Cygnus A by Hargrave and Ryle (1974) using the Cambridge Five-Kilometer Telescope at 5 GHz. This map showed for the first time the radio nucleus associated with the central galaxy and the high intensity at the outer edges of the radio lobes. Reprinted by permission of the Royal Astronomical Society.

vide increased mapping speed, sensitivity, and angular resolution were brought into operation during the 1970s. Prominent among these were the Five-Kilometer Radio Telescope at Cambridge, England (Ryle 1972), the Westerbork Synthesis Radio Telescope in the Netherlands (Baars et al. 1973), and the Very Large Array (VLA) in New Mexico (Thompson et al. 1980; Napier, Thompson, and Ekers 1983). These instruments permit mapping of radio sources with a resolution of less than one arcsec at centimeter wavelengths. By using n_a antennas, where n_a varies up to 27 in the arrays mentioned, as many as $n_a(n_a - 1)/2$ simultaneous baselines are obtained. If the array is designed to avoid redundancy in the antenna spacings, the speed with which the visibility function is measured is approximately proportional to n_a^2. Maps of Cygnus A obtained with two of the arrays mentioned above are shown in Figs. 1.17 and 1.18. A review of the development of synthesis instruments at Cambridge is given in the Nobel lecture by Ryle (1975). An array with large collecting area, the Giant Meter-Wave Radio Telescope (GMRT), which operates at frequencies from 38 to 1420 MHz, was completed in 1998 near Pune, India (Swarup et al. 1991). Current advances in broadband antenna technology and large-scale integrated circuits should enable further large increases in collecting area in the future, for example, the Square Kilometer Array (SKA) (Hopkins et al. 1999, Smolders and van Harlem 1999).

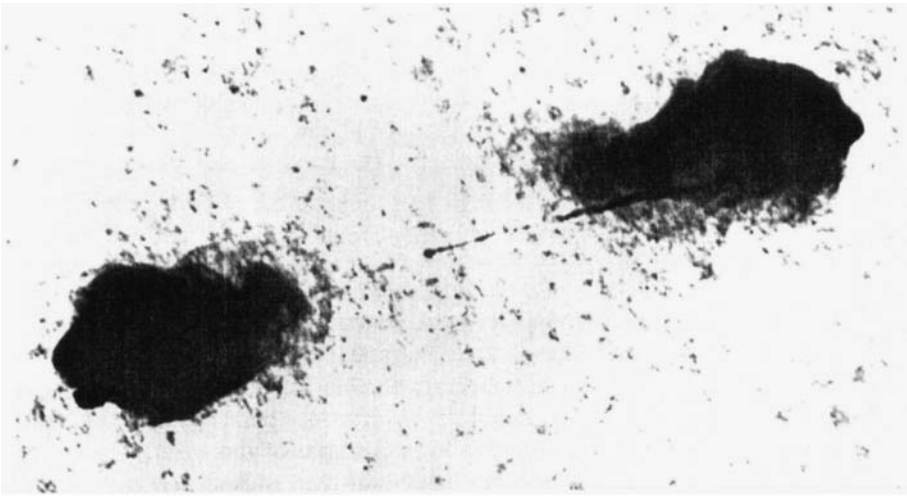


Figure 1.18 Radio image of Cygnus A made with the VLA at 4.9 GHz by Perley, Dreher, and Cowan (1984). Observations with four configurations of the array were combined and the resolution is 0.4 arcsec. The display of the image shown here involves a nonlinear process to enhance the contrast of the fine structure. This emphasizes the jet from the central galaxy to the northwestern lobe (top right) and the filamentary structure in the main lobes. Comparison with other records of Cygnus A in this chapter illustrates the technical advances made during three decades. Reproduced by permission of NRAO/AUI.

During the 1980s and 1990s synthesis arrays operating at short millimeter wavelengths (frequencies of 100 GHz or greater) were developed. Spectral lines are particularly numerous at these frequencies. Several considerations are more important at millimeter wavelengths than at centimeter wavelengths. Because the wavelengths are much shorter, any irregularity in the atmospheric path length results in a proportionately greater effect on the signal phase. Attenuation in the neutral atmosphere is much more serious at millimeter wavelengths. Also, the beams of the individual antennas become narrower at shorter wavelengths, and maintenance of a sufficiently wide field of view is one reason why the antenna diameter tends to decrease with increasing frequency. Thus, to obtain the necessary sensitivity, larger numbers of antennas are required than at centimeter wavelengths. Arrays for millimeter wavelengths include those at Hat Creek, California (Welch 1994); Owens Valley, California (Scoville et al. 1994); Nobeyama, Japan (Morita et al. 1994); the Plateau de Bure, France (Guilloteau 1994); Mauna Kea, Hawaii (Moran 1998a); and Chajnantor, Chile (Brown 1998).

Very-Long-Baseline Interferometry

Investigation of the angular diameters of quasars and other objects that appear nearly pointlike in structure presented an important challenge throughout the early years of radio astronomy. An advance that led to an immediate increase

of an order of magnitude in resolution, and subsequently to several orders more, was the use of independent local oscillators and signal recorders. By using local oscillators at each antenna that are controlled by high-precision frequency standards, it is possible to preserve the coherence of the signals for time intervals long enough to measure interference fringes. The received signals are converted to an intermediate frequency low enough that they can be recorded directly on magnetic tape, and the tapes are subsequently brought together and played into a correlator. This technique became known as *very-long-baseline interferometry* (VLBI), and the early history of its development is discussed by Moran (1998b). The technical requirements for VLBI were widely discussed in the early 1960s [see, e.g., Matveenko, Kardashev, and Sholomitskii (1965)].

A successful early experiment was performed in January 1967 by a group at the University of Florida who detected fringes from the burst radiation of Jupiter at 18 MHz (Brown, Carr, and Block 1968). Because of the strong signals and low frequency, the required recording bandwidth was only 2 kHz and the frequency standards were crystal oscillators. Much more sensitive and precise VLBI systems, which used wider bandwidths and atomic frequency standards, were developed by three other groups. In Canada an analog recording system was developed, with a bandwidth of 1 MHz based on television tape recorders (Broten et al. 1967). Fringes were obtained at a frequency of 448 MHz on baselines of 183 and 3074 km on several quasars in April 1967. In the United States, another group from the National Radio Astronomy Observatory and Cornell University developed a computer-compatible digital recording system with a bandwidth of 360 kHz (Bare et al. 1967). They obtained fringes at 610 MHz on a baseline of 220 km on several quasars in May 1967. A third group from MIT joined in the development of the NRAO–Cornell system in early 1967 and obtained fringes at a frequency of 1665 MHz on a baseline of 845 km on several OH-line masers, with spectroscopic analysis, in June 1967 (Moran et al. 1967).

The initial experiments used signal bandwidths of less than a megahertz, but by the 1980s systems capable of recording signals with bandwidths greater than 100 MHz were available, with corresponding improvements in sensitivity. Real-time linking of the signals from remote telescopes to the correlator via a geostationary satellite has been demonstrated (Yen et al. 1977). Also, experiments were performed in which the local oscillator signal was distributed over a satellite link (Knowles et al. 1982). These developments have lessened the distinction between VLBI and more conventional forms of interferometry. However, there are many technical peculiarities of VLBI, which are described in Chapter 9.

An early example of the extremely high angular resolution that can be achieved with VLBI is provided by a measurement by Burke et al. (1972), who obtained a resolution of 2×10^{-4} arcsec using antennas in Westford, Massachusetts, and Simeiz in the Crimea, operating at a wavelength of 1.3 cm. Early results, obtained using a few baselines only, were generally interpreted in terms of the simple models in Fig. 1.5. During the mid-1970s several groups of astronomers began to combine their facilities to obtain measurements over ten or more baselines simultaneously. In the United States the Network Users' Group included the following observatories: Haystack Observatory in Massachusetts (NEROC); Green

Bank, West Virginia (NRAO); Vermilion River Observatory in Illinois (Univ. of Ill.); North Liberty in Iowa (Univ. of Iowa); Fort Davis, Texas (Harvard College Observatory); Hat Creek Observatory, California (Univ. of Calif.); Owens Valley Radio Observatory, California (Caltech); Maryland Point in Maryland (U.S. Naval Observatory); and Goldstone in California (JPL). Observations by such groups led to more complex models [see, e.g., Cohen et al. (1975)]. Important results were the discovery and investigation of superluminal (apparently faster-than-light) motions in quasars (Whitney et al. 1971, Cohen et al. 1971), as shown in Fig. 1.19, and the measurement of proper motion in H_2O line masers (Genzel et al. 1981). The first array of antennas built specifically for astronomical measurements by VLBI, the Very-Long-Baseline Array (VLBA) of the U.S. National Radio Astronomy Observatory (NRAO), was brought into operation in 1994. It consists of ten 25-m-diameter antennas, one in the U.S. Virgin Islands, eight in the continental United States, and one in Hawaii (Napier et al. 1994). The VLBA is often linked with additional antennas to form even larger arrays.

A problem in VLBI observations is that the use of nonsynchronized local oscillators complicates the calibration of the phase of the fringes. To overcome this problem, the phase closure relationship of Eq. (1.13) was first applied to VLBI data by Rogers et al. (1974). The technique rapidly developed into a method to obtain images known as hybrid mapping. For examples of hybrid mapping, see Figs. 1.19 and 1.20. This and related procedures are also used in mapping with connected-element[§] arrays and are discussed in Chapter 11. For some spectral line observations when the source consists of spatially isolated masers, the signals from which are separated by their individual Doppler shifts, it is possible to map the masers with phase referencing techniques (Reid et al. 1980).

The great potential of VLBI in astrometry and geodesy was immediately recognized [see, e.g., Gold (1967)]. Its use in these applications developed rapidly during the 1970s and 1980s; see, for example, Whitney et al. (1976) and Clark et al. (1985). In the United States, NASA and several other federal agencies set up a cooperative program of geodetic measurements in the early 1980s. This work evolved in part from the use of deep-space communications facilities for VLBI observations. The program includes the use of transportable antennas for periodic monitoring of the positions of many different sites. Astrometry with submilliarcsecond accuracy has opened up new possibilities in astronomy, for example, the detection of the motion of the sun around the Galactic center from the proper motion of Sagittarius A* (Reid et al. 1999, Backer and Sramek 1999). The International Celestial Reference Frame, adopted by the International Astronomical Union, is based on VLBI measurements of 212 extragalactic sources (Ma et al. 1998).

The combination of VLBI with spectral line processing is particularly effective in the study of problems that involve both astrometry and dynamical analysis of astronomical systems. The galaxy NGC4258, which exhibits an active galactic

[§]The term *connected-element*, or *linked-element*, is used to describe arrays of the conventional type in which the signals are brought to the correlators in real time, usually by transmission lines or radio links, in contrast to systems in which the IF signals are recorded for subsequent correlation.

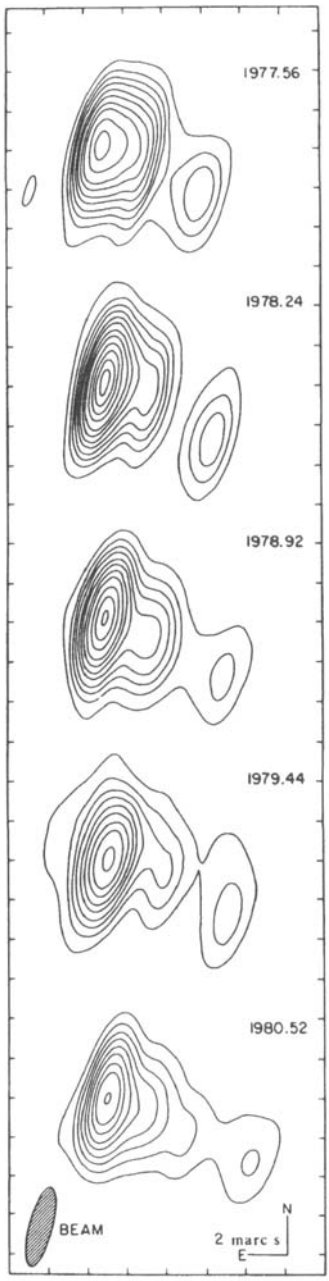


Figure 1.19 VLBI maps of the quasar 3C273 at five epochs, showing the relative positions of two components. From the distance of the object, deduced from the optical red shift, the apparent relative velocity of the components exceeds the velocity of light, but this can be explained by relativistic and geometric effects. The observing frequency is 10.65 GHz. An angular scale of 2 mas is shown in the lower right corner. From Pearson et al. (1981). Reprinted from *Nature*, Vol. 290, No. 5805, p. 366; copyright ©1981 Macmillan Journals Limited.

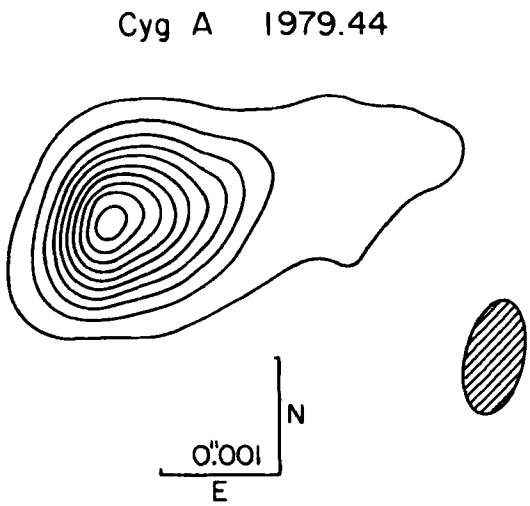


Figure 1.20 VLBI map of the small central component of Cygnus A by Linfield (1981), made using four antennas at 10.65 GHz. The half-power contour of the synthesized beam, shown by the shaded ellipse, has dimensions of 0.5 × 1 mas. The contour interval for the source is 8×10^8 K in brightness temperature. The major axis of the source is in the same direction as the jet in Fig. 1.17.

nucleus, has been found to contain a number of small regions that emit strongly in the 22.235 GHz water line as a result of maser processes. VLBI observations have provided an angular resolution of 2×10^{-4} arcsec, an accuracy of a few microarcseconds in the relative positions of the masers, and measurements of Doppler shifts to an accuracy of 0.1 km s^{-1} in radial velocity. NGC4258 is fortuitously aligned so that the disk is almost edge-on as viewed from the earth. The orbital velocities of the masers are accurately known as a function of radius from the center of motion because they obey Kepler's law. Hence the distance can be found by comparing the linear and angular motions. The angular motions are about 30 μarcsec per year. These results provide a value for the central mass of 3.9×10^7 times the mass of the sun (Miyoshi et al. 1995), and 7.2 ± 0.3 Mpc for the distance (Herrnstein et al. 1999). It is believed that the central mass cannot be explained in terms of a dense cluster of stars, but provides strong evidence of a black hole at the center of NGC4258. Analysis of the maser orbits indicates that the disk surrounding the black hole has a slightly warped profile, as shown in Fig. 1.21. The uncertainty of 4% in the distance of an extragalactic object, measured directly, is without precedent and is likely to be improved by continuing studies of this and similar galaxies.

VLBI Using Orbiting Antennas

The use of spaceborne antennas in VLBI observations is referred to as the OVLBI (orbiting VLBI) technique. The first observations of this type were made in 1986

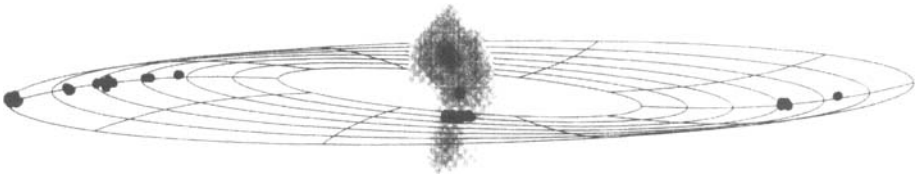


Figure 1.21 The warped annular disk surrounding the central mass of NGC4258, modeled to the maser positions, velocities, and accelerations observed using VLBI with spectral processing. The black spot at the center marks the dynamical center of the disk. The diameter of the disk is 0.6 pc (1 parsec = 3.1×10^{16} m). The continuum emission at 1.3 cm wavelength is shown as the gray-scale feature near the center. The position of the continuum is registered with respect to the masers to an accuracy of a few μarcsec. From Moran et al. (1999), *J. Astrophys. Astr.*, published by the Indian Academy of Sciences, Bangalore.

using a satellite of the U.S. Tracking and Data Relay Satellite System (TDRSS). These satellites are in geostationary orbit at a height of approximately 36,000 km, and are used to relay data from low-earth-orbit spacecraft to earth. They carry two 4.9-m antennas used to communicate with other satellites at 2.3 and 15 GHz and a smaller antenna for the space-to-earth link. In this experiment one of the 4.9-m antennas was used to observe a radio source and the other received a reference signal from a hydrogen maser on the ground (Levy et al. 1989). The received signals were transmitted to the ground and recorded on a VLBI tape system for correlation with signals from ground-based antennas. The numbers of sources detected were 23 and 11 at 2.3 and 15 GHz, respectively (Linfield et al. 1989, 1990). At 15 GHz the fringe width was of order 0.3 mas, and interpretation of the results in terms of circular Gaussian models indicated brightness temperatures as high as 2×10^{12} K.

VLBI observations using a satellite in a non-geostationary orbit were first made in 1997 following the launch of the HALCA satellite of Japan (Hirabayashi et al. 1998), designed specifically for VLBI observations. This was equipped with an antenna of 8 m diameter, and observations were made at 1.6 and 5 GHz. The orbital period was approximately 6.6 h and the apogee, 21,000 km. The rapid orbiting of such a satellite provides greater variation in the baseline vectors to terrestrial antennas than is provided by a geosynchronous satellite, and thus more effective measurement of source structure. However, correction of the phase data for the satellite motion requires very accurate orbit modeling.

The achievement of very long baselines by reflection from the moon, a natural satellite of the earth, has been discussed by Hagfors, Phillips, and Belcora (1990). Reflection from the surface of the moon could provide baselines up to a length approaching the radius of the lunar orbit. An antenna of 100 m aperture, or larger, would be used to track the moon and receive the reflected signal from the source under study, and a smaller antenna could be used for the direct signal. It is estimated that the sensitivity would be about three orders of magnitude less than would be obtained by observing the source directly with both antennas. Further complications result from the roughness of the lunar surface and from libration.

The technique could be useful for special observations requiring very high angular resolution of strong sources, for example, for the burst radiation from Jupiter. In the future, VLBI using a station on the moon is also a possibility.

1.4 QUANTUM EFFECT

The development of VLBI introduced a new facet into the apparent paradox in the quantum-mechanical description of interferometry (Burke 1969). The radio interferometer is the analog of Young's two-slit interference experiment. It is well known (Louden 1973) that a single photon creates an interference pattern, but that any attempt to determine which slit the photon entered will destroy the interference pattern; otherwise the uncertainty principle would be violated. Consideration of VLBI suggests that it might be possible to determine at which antenna a particular photon arrived, since its signature is captured on the tape as well as in the fringe pattern generated during correlation. However, in the radio frequency range, the input stages of receivers used as the measurement devices consist of amplifiers or mixers which conserve the received phase in their outputs. This allows formation of the fringes in subsequent stages. The response of such devices must be consistent with the uncertainty principle, $\Delta E\,\Delta t \simeq h/2\pi$, where ΔE and Δt are the uncertainties in signal energy and measurement time. This principle can be written in terms of uncertainty in photon number, ΔN, and phase, $\Delta\phi$, as

$$\Delta N\,\Delta\phi \simeq 1\,, \tag{1.14}$$

where $\Delta E = h\nu\,\Delta N$ and $\Delta\phi = 2\pi\nu\,\Delta t$. To preserve phase, $\Delta\phi$ must be small, so ΔN must be correspondingly large and there must be an uncertainty of at least one photon per unit bandwidth per unit time in the output of the receiving amplifier. Hence the signal-to-noise ratio is less than unity in the single-photon limit, and it is impossible to determine at which antenna a single photon entered. An alternative but equivalent statement is that the output of any receiving system must contain a noise component that is not less than an equivalent input power approximately equal to $h\nu$ per unit bandwidth.

The individual photons that constitute a radio signal arrive at an antenna at random times, but with an average rate that is proportional to the signal strength. For phenomena of this type, the number of events that occur in a given time interval τ varies statistically in accordance with the Poisson distribution. For a signal power $P_{\rm sig}$, the average number of photons that arrive within time τ is $\overline{N} = P_{\rm sig}\tau/h\nu$. The rms deviation of the number arriving during a series of intervals τ is, for Poisson statistics, given by $\Delta N = \sqrt{\overline{N}}$. From Eq. (1.14) the resulting uncertainty in the signal phase is

$$\Delta\phi \simeq \frac{1}{\sqrt{\overline{N}}} = \sqrt{\frac{h\nu}{P_{\rm sig}\tau}}\,. \tag{1.15}$$

We can also express the uncertainty in the measurement of the signal phase in terms of the noise that is present in the receiving system. The minimum noise power, P_{noise}, is approximately equal to the thermal noise from a matched resistive load at temperature $h\nu/k$, that is, $P_{\text{noise}} = h\nu\,\Delta\nu$. The uncertainty in the phase, as measured with an averaging time τ, becomes

$$\Delta\phi = \sqrt{\frac{P_{\text{noise}}}{P_{\text{sig}}\tau\,\Delta\nu}}\,. \tag{1.16}$$

Note that $\Delta\phi$ is the accuracy with which the phase of the amplified, received signal from one antenna can be measured: for example, in Doppler tracking of a spacecraft (Cannon 1990). This is not to be confused with the accuracy of measurement of the fringe phase of an interferometer. For a frequency $\nu = 1$ GHz, the effective noise temperature $h\nu/k$ is equal to 0.048 K. Thus for frequencies up to some tens of gigahertz the quantum effect noise makes only a small contribution to the receiver noise. At 900 GHz, which is generally considered to be about the high-frequency limit for ground-based radio astronomy, $h\nu/k = 43$ K, and the contribution to the system noise is becoming important. In the optical region $\nu \approx 500$ THz, $h\nu/k \approx 30,000$ K, and heterodyne systems are hardly practical, as discussed in Section 16.4. However, in the optical region it is possible to build "direct detection" devices that detect power without conserving phase, so $\Delta\phi$ in Eq. (1.16) effectively tends to infinity, and there is no constraint on the measurement accuracy of the number of photons. Thus most optical interferometers form fringes directly from the light received, and measure the resulting patterns of light intensity to determine the fringe parameters.

For further reading on the general subject of thermal and quantum noise, see, for example, Oliver (1965) and Kerr, Feldman, and Pan (1997). Nityananda (1994) compares quantum issues in the radio and optical domains, and a discussion of basic concepts is given by Radhakrishnan (1999).

BIBLIOGRAPHY

Alder, B., S. Fernbach, and M. Rotenberg, Eds., *Methods in Computational Physics*, Vol. 14, Academic Press, New York, 1975.

Berkner, L. V., Ed., *IRE Trans. Antennas Propag.*, Special Issue on Radio Astronomy, **AP-9**, No. 1, 1961.

Biraud, F., Ed., *Very Long Baseline Interferometry Techniques*, Cepadues, Toulouse, France, 1983.

Bracewell, R. N., Ed., Paris Symposium on Radio Astronomy, *IAU Symp. 9*, Stanford Univ. Press, Stanford, CA, 1959.

Bracewell, R. N., Radio Astronomy Techniques, in *Handbuch Der Physik*, Vol. 54, S. Flugge, Ed., Springer-Verlag, Berlin, 1962.

Burke, B. F. and F. Graham-Smith, *An Introduction to Radio Astronomy*, Cambridge Univ. Press, Cambridge, UK, 1997.

Christiansen, W. N. and J. A. Högbom, *Radiotelescopes*, Cambridge Univ. Press, Cambridge, UK, 1969 (2nd ed. 1985).

Cornwell, T. J. and R. A. Perley, Eds., *Radio Interferometry: Theory, Techniques, and Applications*, Astron. Soc. Pacific Conf. Ser., **19**, 1991.

Findlay, J. W., Ed., *Proc. IEEE*, Special Issue on Radio and Radar Astronomy, **61**, No. 9, 1973.

Frater, R. H. and J. W. Brooks, Eds., *Proc. IREE Australia*, Special Issue on the Australia Telescope, **12**, No. 2, 1992.

Goldsmith, P. F., Ed., *Instrumentation and Techniques for Radio Astronomy*, IEEE Press, New York, 1988.

Haddock, F. T., Ed., *Proc. IRE*, Special Issue on Radio Astronomy, **46**, No. 1, 1958.

Ishiguro, M. and W. J. Welch, Eds., *Astronomy with Millimeter and Submillimeter Wave Interferometry*, IAU Colloquium 140, Astron. Soc. Pacific. Conf. Ser., **59**, 1994.

Kraus, J. D., Ed., *IEEE Trans. Mil. Electron.*, Special Issue on Radio and Radar Astronomy, **Mil-8**, Nos. 3 and 4, 1964, also issued by *IEEE Trans. Antennas Propag.*, **AP-12**, No. 7, 1964.

Kraus, J. D., *Radio Astronomy*, McGraw-Hill, New York, 1966, 2nd. ed., Cygnus-Quasar, Powell, OH, 1986.

Lovell, B. and J. A. Clegg, *Radio Astronomy*, Chapman and Hall, London, 1952.

Meeks, M. L., Ed., *Methods of Experimental Physics*, Vol. 12, Parts B and C, Academic Press, New York, 1976.

Pawsey, J. L., Ed., *Proc. IRE Aust.*, Special Issue on Radio Astronomy, **24**, No. 2, 1963.

Pawsey, J. L. and R. N. Bracewell, *Radio Astronomy*, Oxford Univ. Press, Oxford, UK, 1955.

Perley, R. A., F. R. Schwab, and A. H. Bridle, Ed., *Synthesis Imaging in Radio Astronomy*, Astron. Soc. Pacific Conf. Ser., **6**, 1989.

Raimond, E. and R. Genee, Eds., *The Westerbork Observatory, Continuing Adventure in Radio Astronomy*, Kluwer, Dordrecht, 1996.

Rohlfs, K. and T. L. Wilson, *Tools of Radio Astronomy*, Springer-Verlag, Berlin, 1986, 1996.

Shklovsky, I. S., *Cosmic Radio Waves*, trans. R. B. Rodman and C. M. Varsavsky, Harvard Univ. Press, Cambridge, MA, 1960.

Sullivan, W. T., III, Ed., *The Early Years of Radio Astronomy*, Cambridge Univ. Press, Cambridge, UK, 1984.

Taylor G. B., C. L. Carilli, and R. A. Perley, Eds., *Synthesis Imaging in Radio Astronomy II*, Astron. Soc. Pacific Conf. Ser., **180**, 1999.

Wild, J. P., Ed., *Proc. IREE Aust.*, Special Issue on the Culgoora Radioheliograph, **28**, No. 9, 1967.

Wohlleben, R., H. Mattes, and T. Krichbaum, *Interferometry in Radioastronomy and Radar Techniques*, Kluwer, Dordrecht, 1991.

Yen, J. L., Image Reconstruction in Synthesis Radio Telescope Arrays, in *Array Signal Processing*, S. Haykin Ed., Prentice-Hall, Englewood Cliffs, NJ, 1985, pp. 293–350.

REFERENCES

Appleton, E. V., Departure of Long-Wave Solar Radiation from Black-Body Intensity, *Nature*, **156**, 534–535, 1945.

Baade, W. and R. Minkowski, Indentification of the Radio Sources in Cassiopeia, Cygnus A, and Puppis A, *Astrophys. J.*, **119**, 206–214, 1954a.

Baade, W. and R. Minkowski, On the Indentification of Radio Sources, *Astrophys. J.*, **119**, 215–231, 1954b.

Baars, J. W. M., J. F. van der Brugge, J. L. Casse, J. P. Hamaker, L. H. Sondaar, J. J. Visser, and K. J. Wellington, The Synthesis Radio Telescope at Westerbork, *Proc. IEEE*, **61**, 1258–1266, 1973.

Backer, D. C. and R. A. Sramek, Proper Motion of the Compact, Nonthermal Radio Source in the Galactic Center, Sagittarius A*, *Astrophys. J.*, **524**, 805–815, 1999.

Bare, C., B. G. Clark, K. I. Kellermann, M. H. Cohen, and D. L. Jauncey, Interferometer Experiment with Independent Local Oscillators, *Science*, **157**, 189–191, 1967.

Bennett, A. S., The Revised 3C Catalog of Radio Sources, *Mem. R. Astron. Soc.*, **68**, 163–172, 1962.

Blake, G. A., E. C. Sutton, C. R. Masson, and T. G. Phillips, Molecular Abundances in OMC-1: The Chemical Composition of Interstellar Molecular Clouds and the Influence of Massive Star Formation, *Astrophys. J.*, **315**, 621–645, 1987.

Blum, E. J., Le Réseau Nord-Sud à Multiples, *Ann. Astrophys.*, **24**, 359–366, 1961.

Blum, E. J., A. Boischot, and M. Ginat, Le Grand Interféromètre de Nancay, *Ann. Astrophys.*, **20**, 155–164, 1957.

Blythe, J. H., A New Type of Pencil Beam Aerial for Radio Astronomy, *Mon. Not. R. Astron. Soc.*, **117**, 644–651, 1957.

Bolton, J. G. and O. B. Slee, Galactic Radiation at Radio Frequencies, V. The Sea Interferometer, *Aust. J. Phys.*, **6**, 420–433, 1953.

Bolton, J. G. and G. J. Stanley, Variable Source of Radio Frequency Radiation in the Constellation of Cygnus, *Nature*, **161**, 312–313, 1948.

Braccesi, A., M. Ceccarelli, G. Colla, R. Fanti, A. Ficarra, G. Gelato, G. Greuff, and G. Sinigaglia, The Italian Cross Radio Telescope, III. Operation of the Telescope, *Nuovo Cimento B*, **62**, 13–19, 1969.

Bracewell, R. N. and G. Swarup, The Stanford Microwave Spectroheliograph Antenna, a Microsteradian Pencil Beam Interferometer, *IRE Trans. Antennas Propag.*, **AP-9**, 22–30, 1961.

Broten, N. W., T. H. Legg, J. L. Locke, C. W. McLeish, R. S. Richards, R. M. Chisholm, H. P. Gush, J. L. Yen, and J. A. Galt, Observations of Quasars Using Interferometer Baselines up to 3,074 km, *Nature*, **215**, 38, 1967.

Brown, G. W., T. D. Carr, and W. F. Block, Long Baseline Interferometry of S-Bursts from Jupiter, *Astrophys. Lett.*, **1**, 89–94, 1968.

Brown, R. L., Technical Specification of the Millimeter Array, in *SPIE Con. Advanced Technology MMW, Radio, and Terahertz Telescopes*, Kona, Hawaii, March 1998, T. G. Phillips, Ed., Proc. SPIE **3357**, 231–237, 1998.

Burke, B. F., Quantum Interference Paradox, *Nature*, **223**, 389–390, 1969.

Burke, B. F., K. J. Johnston, V. A. Efanov, B. G. Clark, L. R. Kogan, V. I. Kostenko, K. Y. Lo, L. I. Matveenko, I. G. Moiseev, J. M. Moran, S. H. Knowles, D. C. Papa, G. D. Papadopoulos, A. E. E. Rogers, and P. R. Schwartz, Observations of Maser Radio Source with an Angular Resolution of 0.″0002, *Soviet Astron.-AJ*, **16**, 379–382, 1972.

Cannon, W. H., Quantum Mechanical Uncertainty Limitations on Deep Space Navigation by Doppler Tracking and Very Long Baseline Interferometry, *Radio Sci.*, **25**, 97–100, 1990.

Christiansen, W. N. and R. F. Mullaly, Solar Observations at a Wavelength of 20 cm with a Cross-Grating Interferometer, *Proc. IRE Aust.*, **24**, 165–173, 1963.

Christiansen, W. N. and J. A. Warburton, The Distribution of Radio Brightness over the Solar Disk at a Wavelength of 21 cm, III. The Quiet Sun—Two Dimensional Observations, *Aust. J. Phys.*, **8**, 474–486, 1955.

Clark, B. G., V. Radhakrishnan, and R. W. Wilson, The Hydrogen Line in Absorption, *Astrophys. J.*, **135**, 151–174, 1962.

Clark, T. A. and Twenty Coauthors, Precision Geodesy Using the Mark-III Very-Long-Baseline Interferometer System, *IEEE Trans. Geosci. Remote Sens.*, **GE-23**, 438–449, 1985.

Cohen, M. H., W. Cannon, G. H. Purcell, D. B. Shaffer, J. J. Broderick, K. I. Kellermann, and D. L. Jauncy, The Small-Scale Structure of Radio Galaxies and Quasi-Stellar Sources, at 3.8 Centimeters, *Astrophys. J.*, **170**, 207–217, 1971.

Cohen, M. H., A. T. Moffet, J. D. Romney, R. T. Schilizzi, D. B. Shaffer, K. I. Kellermann, G. H. Purcell, G. Grove, G. W. Swenson Jr., J. L. Yen, I. I. K. Pauliny-Toth, E. Preuss, A. Witzel, and D. Graham, Observations with a VLBI Array, I. Introduction and Procedures, *Astrophys. J.*, **201**, 249–255, 1975.

Condon, J. J., W. D. Cotton, E. W. Greisen, Q. F. Yin, R. A. Perley, G. B. Taylor, and J. J. Broderick, The NRAO VLA Sky Survey, *Astron. J.*, **115**, 1693–1716, 1998.

Conway, R. G., K. I. Kellermann, and R. J. Long, The Radio Frequency Spectra of Discrete Radio Sources, *Mon. Not. R. Astron. Soc.*, **125**, 261–284, 1963.

Covington, A. E. and N. W. Broten, An Interferometer for Radio Astronomy with a Single-Lobed Radiation Pattern, *Proc. IRE Trans. Antennas Propag.*, **AP-5**, 247–255, 1957.

Dicke, R. H., The Measurement of Thermal Radiation at Microwave Frequencies, *Rev. Sci. Instrum.*, **17**, 268–275, 1946.

Dreyer, J. L. E., New General Catalog of Nebulae and Clusters of Stars, *Mem. R. Astron. Soc.*, **49**, Part 1, 1888 (repr. *R. Astron. Soc. London*, 1962).

Edge, D. O., J. R. Shakeshaft, W. B. McAdam, J. E. Baldwin, and S. Archer, A Survey of Radio Sources at a Frequency of 159 Mc/s, *Mem. R. Astron. Soc.*, **68**, 37–60, 1959.

Elgaroy, O., D. Morris, and B. Rowson, A Radio Interferometer for Use with Very Long Baselines, *Mon. Not. R. Astron. Soc.*, **124**, 395–403, 1962.

Elitzur, M., *Astronomical Masers*, Kluwer, Dordrecht, 1992.

Fomalont, E. B., The East-West Structure of Radio Source at 1425 MHz, *Astrophys. J. Suppl.*, **15**, 203–274, 1968.

Genzel, R., M. J. Reid, J. M. Moran, and D. Downes, Proper Motions and Distances of H_2O Maser Sources, I. The Outflow in Orion-KL, *Astrophys. J.*, **244**, 884–902, 1981.

Gold, T., Radio Method for the Precise Measurement of the Rotation Period of the Earth, *Science*, **157**, 302–304, 1967.

Gower, J. F. R., P. F. Scott, and D. Wills, A Survey of Radio Sources in the Declination Ranges −07° to 20° and 40° to 80°, *Mem. R. Astron. Soc.*, **71**, 49–144, 1967.

Guilloteau, S., The IRAM Interferometer on Plateau de Bure, in *Astronomy with Millimeter and Submillimeter Wave Interferometry*, M. Ishiguro, and W. J. Welch, Eds., Astron. Soc. Pacific. Conf. Ser., **59**, 27–34, 1994.

Hagfors, T., J. A. Phillips, and L. Belcora, Radio Interferometry by Lunar Reflections, *Astrophys. J.*, **362**, 308–317, 1990.

Hanbury Brown, R., Measurement of Stellar Diameters, *Ann. Rev. Astron. Astrophys.*, **6**, 13–38, 1968.

Hanbury Brown, R., H. P. Palmer, and A. R. Thompson, A Rotating-Lobe Interferometer and its Application to Radio Astronomy, *Philos. Mag.*, Ser. 7, **46**, 857–866, 1955.

Hanbury Brown, R. and R. Q. Twiss, A New Type of Interferometer for Use in Radio Astronomy, *Philos. Mag.*, Ser. 7, **45**, 663–682, 1954.

Hargrave, P. J. and M. Ryle, Observations of Cygnus A with the 5-km Radio Telescope, *Mon. Not. R. Astron. Soc.*, **166**, 305–327, 1974.

Hazard, C. and D. Walsh, A Comparison of an Interferometer and Total-Power Survey of Discrete Sources of Radio Frequency Radiation, in *The Paris Symposium on Radio Astronomy*, R. N. Bracewell, Ed., Stanford Univ. Press, Stanford, CA, 1959, pp. 477–486.

Herrnstein, J. R., J. M. Moran, L. J. Greenhill, P. J. Diamond, M. Inoue, N. Nakai, M. Miyoshi, C. Henkel, and A. Riess, A Geometric Distance to the Galaxy NGC4258 from Orbital Motions in a Nuclear Gas Disk, *Nature*, **400**, 539–841, 1999.

Hirabayashi, H. and 52 coauthors, Overview and Initial Results of the Very Long Baseline Interferometry Space Observatory Programme, *Science*, **281**, 1825–1829, 1998.

Hogg, D. E., G. H. Macdonald, R. G. Conway, and C. M. Wade, Synthesis of Brightness Distribution in Radio Sources, *Astron. J.*, **74**, 1206–1213, 1969.

Hopkins, A., R. Ekers, C. Jackson, L. Cram, A. Green, D. Manchester, L. Staveley-Smith, and R. Norris, Summary of the "Sub-microjansky Radio Sky" Workshop, *Pub. Astron. Soc. Aust.*, **16**, 152-159, 1999.

Hughes, M. P., A. R. Thompson, and R. S. Colvin, An Absorption-line Study of Galactic Neutral Hydrogen at 21 cm Wavelength, *Astrophys. J. Suppl.*, **23**, 232–367, 1971.

IAU, *Trans. Int. Astron. Union*, **15B**, 142, 1974.

IEEE, Standard Definitions of Terms for Radio Wave Propagation, Std. 211–1977, The Institute of Electrical and Electronics Engineers, New York, 1977.

Jansky, K. G., Electrical Disturbances Apparently of Extraterrestrial Origin, *Proc. IRE*, **21**, 1387–1398, 1933.

Jennison, R. C., A Phase Sensitive Interferometer Technique for the Measurement of the Fourier Transforms of Spatial Brightness Distributions of Small Angular Extent, *Mon. Not. R. Astron. Soc.*, **118**, 276–284, 1958.

Jennison, R. C., High Resolution Imaging Forty Years Ago, in *Very High Angular Resolution Imaging, IAU Symp. 158*, J. G. Robertson and W. J. Tango, Eds., Kluwer, Dordrecht, 1994, pp. 337–341.

Jennison, R. C. and M. K. Das Gupta, Fine Structure in the Extra-terrestrial Radio Source Cygnus 1, *Nature*, **172**, 996–997, 1953.

Jennison, R. C. and M. K. Das Gupta, The Measurement of the Angular Diameter of Two Intense Radio Sources, Parts I and II, *Philos. Mag.*, Ser. 8, **1**, 55–75, 1956.

Jennison, R. C. and V. Latham, The Brightness Distribution Within the Radio Sources Cygnus A (19N4A) and Cassiopeia (23N5A), *Mon. Not. R. Astron. Soc.*, **119**, 174–183, 1959.

Kellermann, K. I. and I. I. K. Pauliny-Toth, The Spectra of Opaque Radio Sources, *Astrophys. J.*, **155**, L71–L78, 1969.

Kerr, A. R., Feldman, M. J., and Pan, S.-K., Receiver Noise Temperature, the Quantum Noise Limit, and the Role of Zero-Point Fluctuations, *Proc. 8th Int. Symp. Space Terahertz Technology*, March 25–27, 1997, also available as *MMA Memorandum 161*, NRAO, Socorro, NM 1997.

Knowles, S. H., W. B. Waltman, J. L. Yen, J. Galt, D. N. Fort, W. H. Cannon, D. Davidson, W. Petrachenko, and J. Popelar, A Phase-Coherent Link via Synchronous Satellite Developed for Very Long Baseline Radio Interferometry, *Radio Sci.*, **17**, 1661–1670, 1982.

Labrum, N. R., E. Harting, T. Krishnan, and W. J. Payten, A Compound Interferometer with a 1.5 Minute of Arc Fan Beam, *Proc. IRE Aust.*, **24**, 148–155, 1963.

Levy, G. S. and 31 coauthors, VLBI Using a Telescope in Earth Orbit. II. The Observations, *Astrophys. J.*, **336**, 1089–1104, 1989.

Linfield, R., VLBI Observations of Jets in Double Radio Galaxies, *Astrophys. J.*, **244**, 436–446, 1981.

Linfield, R. P. and 14 coauthors, VLBI Using a Telescope in Earth Orbit. II. Brightness Temperatures Exceeding the Inverse Compton Limit, *Astrophys. J.*, **336**, 1105–1112, 1989.

Linfield, R. P. and 27 coauthors, 15 GHz Space VLBI Observations Using an Antenna on a TDRSS Satellite, *Astrophys J.*, **358**, 350–358, 1990.

Longair, M. S., *High Energy Astrophysics, (2 vols.)*, Cambridge Univ. Press, Cambridge, UK, 1992.

Loudon, R., *The Quantum Theory of Light*, Oxford Univ. Press, London, 1973, p. 229.

Lovas, F. J., Recommended Rest Frequencies for Observed Interstellar Molecular Microwave Transitions—1991 Revision, *J. Phys. and Chem. Ref. Data*, **21**, 181–272, 1992.

Lovas, F. J., L. E. Snyder, and D. R. Johnson, Recommended Rest Frequencies for Observed Interstellar Molecular Transitions, *Astrophys. J. Suppl.*, **41**, 451–480, 1979.

Ma, C., E. F. Arias, T. M. Eubanks, A. L. Fey, A.-M. Gontier, C. S. Jacobs, O. J. Sovers, B. A. Archinal, and P. Charlot, The International Celestial Reference Frame as Realized by Very Long Baseline Interferometry, *Astron. J.*, **116**, 516–546, 1998.

McCready, L. L., J. L. Pawsey, and R. Payne-Scott, Solar Radiation at Radio Frequencies and Its Relation to Sunspots, *Proc. R. Soc. A*, **190**, 357–375, 1947.

Maltby, P. and A. T. Moffet, Brightness Distribution in Discrete Radio Sources, *Astrophys. J. Suppl.*, **7**, 93–163, 1962.

Matveenko, L. I., N. S. Kardashev, and G. B. Sholomitskii, Large Base-Line Radio Interferometers, Radiofizika, **8**, 651–654, 1965; Engl. transl. in *Soviet Radiophys.*, **8**, 461–463, 1965.

Michelson, A. A., On the Application of Interference Methods to Astronomical Measurements, *Philos. Mag.*, Ser. 5, **30**, 1–21, 1890.

Michelson, A. A., On the Application of Interference Methods to Astronomical Measurements, *Astrophys. J.*, **51**, 257–262, 1920.

Michelson, A. A. and F. G. Pease, Measurement of the Diameter of α Orionis with the Interferometer, *Astrophys. J.*, **53**, 249–259, 1921.

Mills, B. Y., The Positions of the Six Discrete Sources of Cosmic Radio Radiation, *Aust. J. Sci. Res.*, **A5**, 456–463, 1952.

Mills, B. Y., The Radio Brightness Distribution Over Four Discrete Sources of Cosmic Noise, *Aust. J. Phys.*, **6**, 452–470, 1953.

Mills, B. Y., Cross-Type Radio Telescopes, *Proc. IRE Aust.*, **24**, 132–140, 1963.

Mills, B. Y., R. E. Aitchison, A. G. Little, and W. B. McAdam, The Sydney University Cross-Type Radio Telescope, *Proc. IRE Aust.*, **24**, 156–165, 1963.

Mills, B. Y. and A. G. Little, A High Resolution Aerial System of a New Type, *Aust. J. Phys.*, **6**, 272–278, 1953.

Mills, B. Y., A. G. Little, K. V. Sheridan, and O. B. Slee, A High-Resolution Radio Telescope for Use at 3.5 m, *Proc. IRE*, **46**, 67–84, 1958.

Mills, B. Y. and O. B. Slee, A Preliminary Survey of Radio Sources in a Limited Region of the Sky at a Wavelength of 3.5 m, *Aust. J. Phys.*, **10**, 162–194, 1957.

Miyoshi, M., J. Moran, J. Herrnstein, L. Greenhill, N. Nakal, P. Diamond, and M. Inoue, Evidence for a Black Hole from High Rotation Velocities in a Sub-parsec Region of NGC4258, *Nature*, **373**, 127–129, 1995.

Moore, E. M. amd A. P. Marscher, Observational Probes of the Small-Scale Structure of Molecular Clouds, *Astrophys. J.*, **452**, 671–679, 1995.

Moran, J. M., The Submillimeter Array, in *SPIE Conf. Advanced Technology MMW, Radio, and Terahertz Telescopes*, Kona, Hawaii, March 1998, T. G. Phillips, Ed., Proc. SPIE, **3357**, 208–219, 1998a.

Moran, J. M., Thirty Years of VLBI: Early Days, Successes, and Future, in *Radio Emission from Galactic and Extragalactic Compact Sources*, Astron Soc. Pacific Conf. Ser., **144**, J. A. Zensus, G. B. Taylor, and J. M. Wrobel, Eds., 1–10, 1998b.

Moran, J. M., P. P. Crowther, B. F. Burke, A. H. Barrett, A. E. E. Rogers, J. A. Ball, J. C. Carter, and C. C. Bare, Spectral Line Interferometer with Independent Time Standards at Stations Separated by 845 Kilometers, *Science*, **157**, 676–677, 1967.

Moran, J. M., L. J. Greenhill, and J. R. Herrnstein, Observational Evidence for Massive Black Holes in the Centers of Active Galaxies, *J. Astrophys. Astr. (Indian Acad. Sci.)*, **20**, 165–185, 1999.

Morita, K.-I., The Nobeyama Millimeter Array, in *Astronomy with Millimeter and Submillimeter Wave Interferometry*, M. Ishiguro and W. J. Welch, Eds., Astron. Soc. Pacific. Conf. Ser., **59**, 18–26, 1994.

Morris, D., H. P. Palmer, and A. R. Thompson, Five Radio Sources of Small Angular Diameter, *Observatory*, **77**, 103–106, 1957.

Napier, P. J., A. R. Thompson, and R. D. Ekers, The Very Large Array: Design and Performance of a Modern Synthesis Radio Telescope, *Proc. IEEE*, **71**, 1295–1322, 1983.

Napier, P. J., D. S. Bagri, B. G. Clark, A. E. E. Rogers, J. D. Romney, A. R. Thompson, and R. C. Walker, The Very Long Baseline Array, *Proc. IEEE*, **82**, 658–672, 1994.

Nityananda, R., Comparing Optical and Radio Quantum Issues, in *Very High Resolution Imaging, IAU Symp. 158*, J. G. Robertson and W. J. Tango, Eds., Kluwer, Dordrecht, 1994, pp. 11–18.

Nyquist, H., Thermal Agitation of Electric Charge in Conductors, *Phys. Rev.*, **32**, 110–113, 1928.

Oliver, B. M., Thermal and Quantum Noise, *Proc. IEEE*, **53**, 436–454, 1965.

Pawsey, J. L., Sydney Investigations and Very Distant Radio Sources, *Pub. Astron. Soc. Pac.*, **70**, 133–140, 1958.

Pearson, T. J., S. C. Unwin, M. H. Cohen, R. P. Linfield, A. C. S. Readhead, G. A. Seielstad, R. S. Simon, and R. C. Walker, Superluminal Expansion of Quasar 3C273, *Nature*, **290**, 365–368, 1981.

Pease, F. G., Interferometer Methods in Astronomy, *Ergeb. Exakten Naturwiss.*, **10**, 84–96, 1931.

Perley, R. A., J. W. Dreher, and J. J. Cowan, The Jet and Filaments in Cygnus, *Astrophys. J.*, **285**, L35–L38, 1984.

Picken, J. S. and G. Swarup, The Stanford Compound-Grating Interferometer, *Astron. J.*, **69**, 353–356, 1964.

Radhakrishnan, V., Noise and Interferometry, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 671–688, 1999.

Read, R. B., Two-Element Interferometer for Accurate Position Determinations at 960 Mc, *IRE Trans. Antennas Propag.*, **AP-9**, 31–35, 1961.

Reber, G., Cosmic Static, *Astrophys. J.*, **91**, 621–624, 1940.

Reid, M. J., A. D. Haschick, B. F. Burke, J. M. Moran, K. J. Johnston, and G. W. Swenson, Jr., The Structure of Interstellar Hydroxyl Masers: VLBI Synthesis Observations of W3(OH), *Astrophys. J.*, **239**, 89–111, 1980.

Reid, M. J. and J. M. Moran, Astronomical Masers, in *Galactic and Extragalactic Radio Astronomy*, 2nd ed., G. L. Verschuur and K. I. Kellermann, Eds., Springer-Verlag, Berlin, 1988, pp. 255–294.

Reid, M. J., A. C. S. Readhead, R. C. Vermuelen, and R. N. Treuhaft, The Proper Motion of Sagittarius A*. I. First VLBA Results, *Astrophys. J.*, **524**, 816–823, 1999.

Roger, R. S., C. H. Costain, J. D. Lacey, T. L. Landaker, and F. K. Bowers, A Supersynthesis Radio Telescope for Neutral Hydrogen Spectroscopy at the Dominion Radio Astrophysical Observatory, *Proc. IEEE*, **61**, 1270–1276, 1973.

Rogers, A. E. E., H. F. Hinteregger, A. R. Whitney, C. C. Counselman, I. I. Shapiro, J. J. Wittels, W. K. Klemperer, W. W. Warnock, T. A. Clark, L. K. Hutton, G. E. Marandino, B. O. Rönnäng, O. E. H. Rydbeck, and A. E. Niell, The Structure of Radio Sources 3C273B and 3C84 Deduced from the Closure Phases and Visibility Amplitudes Observed with Three-Element Interferometers, *Astrophys. J.*, **193**, 293–301, 1974.

Rowson, B., High Resolution Observations with a Tracking Radio Interferometer, *Mon. Not. R. Astron. Soc.*, **125**, 177–188, 1963.

Rybicki, G. B. and A. P. Lightman, *Radiative Processes in Astrophysics*, Wiley-Interscience, New York, 1979 (reprinted 1985).

Ryle, M., A New Radio Interferometer and Its Application to the Observation of Weak Radio Stars, *Proc. R. Soc. A*, **211**, 351–375, 1952.

Ryle, M., The New Cambridge Radio Telescope, *Nature*, **194**, 517–518, 1962.

Ryle, M., The 5-km Radio Telescope at Cambridge, *Nature*, **239**, 435–438, 1972.

Ryle, M., Radio Telescopes of Large Resolving Power, *Science*, **188**, 1071–1079, 1975.

Ryle, M., B. Elsmore, and A. C. Neville, High Resolution Observations of Radio Sources in Cygnus and Cassiopeia, *Nature*, **205**, 1259–1262, 1965.

Ryle, M. and A. Hewish, The Cambridge Radio Telescope, *Mem. R. Astron. Soc.*, **67**, 97–105, 1955.

Ryle, M. and A. Hewish, The Synthesis of Large Radio Telescopes, *Mon. Not. R. Astron. Soc.*, **120**, 220–230, 1960.

Ryle, M., A. Hewish, and J. R. Shakeshaft, The Synthesis of Large Radio Telescopes by the Use of Radio Interferometers, *IRE Trans. Antennas Propag.*, **7**, S120–S124, 1959.

Ryle, M. and A. C. Neville, A Radio Survey of the North Polar Region with a 4.5 Minute of Arc Pencil-Beam System, *Mon. Not. R. Astron. Soc.*, **125**, 39–56, 1962.

Ryle, M. and F. G. Smith, A New Intense Source of Radio Frequency Radiation in the Constellation of Cassiopeia, *Nature*, **162**, 462–463, 1948.

Ryle, M., F. G. Smith, and B. Elsmore, A Preliminary Survey of the Radio Stars in the Northern Hemisphere, *Mon. Not. R. Astron. Soc.*, **110**, 508–523, 1950.

Ryle, M. and D. D. Vonberg, Solar Radiation at 175 Mc/s, *Nature*, **158**, 339–340, 1946.

Scoville, N., J. Carlstrom, S. Padin, A. Sargent, S. Scott, and D. Woody, The Owens Valley Millimeter Array, in *Astronomy with Millimeter and Submillimeter Wave Interferometry*, M. Ishiguro and W. J. Welch, Eds., Astron. Soc. Pacific. Conf. Ser., **59**, 10–17, 1994.

Schilke, P., T. Groesbeck, G. A. Blake, and T. G. Phillips, A 325 to 360 GHz Emission Line Survey of the Orion KL Region, *Astrophys. J. Suppl.*, **108**, 301–337, 1997.

Seidelmann, P. K., Ed., *Explanatory Supplement to the Astronomical Almanac*, University Science Books, Mill Valley, CA, 1992.

Shakeshaft, J. R., M. Ryle, J. E. Baldwin, B. Elsmore, and J. H. Thomson, A Survey of Radio Sources Between Declinations −38° and +83°, *Mem. R. Astron. Soc.*, **67**, 106–154, 1955.

Smith, F. G., An Accurate Determination of the Positions of Four Radio Stars, *Nature*, **168**, 555, 1951.

Smith, F. G., The Determination of the Position of a Radio Star, *Mon. Not. R. Astron. Soc.*, **112**, 497–513, 1952a.

Smith, F. G., The Measurement of the Angular Diameter of Radio Stars, *Proc. Phys. Soc. B.*, **65**, 971–980, 1952b.

Smith, F. G., Apparent Angular Sizes of Discrete Radio Sources—Observations at Cambridge, *Nature*, **170**, 1065, 1952c.

Smolders, A. B., and M. P. van Haarlem, Eds., *Perspectives on Radio Astronomy: Technologies for Large Antenna Arrays*, ASTRON, Dwingeloo, Netherlands, 1999.

Smoot, G. and 27 coauthors, COBE Differential Microwave Radiometers: Instrument Design and Implementation, *Astrophys. J.*,**360**, 685–695, 1990.

Smoot, G. F. and 27 coauthors, Structure in the COBE Differential Microwave Radiometer First-Year Maps, *Astrophys. J.*,**396**, L1–L5, 1992.

Southworth, G. C., Microwave Radiation from the Sun, *J. Franklin Inst.*, **239**, 285–297, 1945.

Swarup, G., S. Ananthakrishnan, V. K. Kapahi, A. P. Rao, C. R. Subrahmanya, and V. K. Kulkarni, The Giant Meter-wave Radio Telescope, *Current Sci.*, **60**, 95–105, 1991.

Thomasson, P., MERLIN, *Quat. J. R. Astron. Soc.*, **27**, 413–431, 1986.

Thompson, A. R., The Planetary Nebulae as Radio Sources, in *Vistas in Astronomy*, Vol. 16, A. Beer, Ed., Pergamon Press, Oxford, 1974, pp. 309–328.

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

Thompson, A. R. and T. Krishnan, Observations of the Six Most Intense Radio Sources with a 1.0′ Fan Beam, *Astrophys. J.*, **141**, 19–33, 1965.

Tiuri, M. E., Radio Astronomy Receivers, *IEEE Trans. Antennas Propag.*, **AP-12**, 930–938, 1964.

Tiuri, M. E. and A. V. Räisänen, Radio-Telescope Receivers, in *Radio Astronomy*, 2nd ed., J. D. Kraus, Cygnus-Quasar Books, Powell, OH, 1986, Ch. 7.

Vitkevich, V. V. and P. D. Kalachev, Design Principles of the FIAN Cross-Type Wide Range Telescope, in *Radio Telescopes*, D. V. Skobel'tsyn, Ed., *Proc. P. N. Lebedev Phys. Inst. (Acad. Sci. USSR)*, Vol. 28, transl. by Consultants Bureau, New York, 1966.

Welch, W. J., The Berkeley–Illinois–Maryland Association Millimeter Array, in *Astronomy with Millimeter and Submillimeter Wave Interferometry*, M. Ishiguro and W. J. Welch, Eds., Astron. Soc. Pacific. Conf. Ser., **59**, 1–9, 1994.

Whitney, A. R., A. E. E. Rogers, H. F. Hinteregger, C. A. Knight, J. I. Levine, S. Lippincott, T. A. Clark, I. I. Shapiro, and D. S. Robertson, A Very Long Baseline Interferometer System for Geodetic Applications, *Radio Sci.*, **11**, 421–432, 1976.

Whitney, A. R., I. I. Shapiro, A. E. E. Rogers, D. S. Robertson, C. A. Knight, T. A. Clark, R. M. Goldstein, G. E. Marandino, and N. R. Vandenberg, Quasars Revisited: Rapid Time Variations Observed via Very Long Baseline Interferometry, *Science*, **173**, 225–230, 1971.

Yen, J. L., K. I. Kellermann, B. Rayher, N. W. Broten, D. N. Fort, S. H. Knowles, W. B. Waltman, and G. W. Swenson, Jr., Real-Time, Very Long Baseline Interferometry Based on the Use of a Communications Satellite, *Science*, **198**, 289–291, 1977.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

2 Introductory Theory of Interferometry and Synthesis Imaging

In the first chapter we introduced some of the principles of interferometry while reviewing the historical development of the subject. This chapter provides a simplified analysis of interferometry and introduces several Fourier transform relationships and other basic concepts. It is intended to provide a broad introduction to the principles of synthesis imaging to facilitate the understanding of more detailed development in later chapters.

2.1 PLANAR ANALYSIS

The instantaneous response of a radio interferometer to a point source can most simply be analyzed by considering the signal paths in the plane containing the electrical centers of the interferometer antennas and the source under observation. For an extended observation it is necessary to take account of the rotation of the earth and consider the geometric situation in three dimensions, as can be seen from Fig. 1.15. However, the two-dimensional geometry is a good approximation for short-duration observations, and the simplified approach facilitates visualization of the response pattern.

Consider the geometric situation shown in Fig. 2.1, where the antenna spacing is east–west. The two antennas are separated by a distance D, the baseline, and observe the same cosmic source which is in the *far field* of the interferometer; that is, it is sufficiently distant that the incident wavefront can be considered to be a plane over the distance D. The source will be assumed for the moment to have infinitesimal angular dimensions. For this discussion, the receivers will be assumed to have narrow bandpass filters that pass only signal components very close to ν.

As explained for the phase-switching interferometer in Chapter 1, the signal voltages are multiplied and then time-averaged, which has the effect of filtering out high frequencies. The wavefront from the source in direction θ reaches the right-hand antenna at a time $\tau_g = (D/c)\sin\theta$ before it reaches the left-hand one. τ_g is called the *geometric delay* and c is the velocity of light. Thus, in terms of

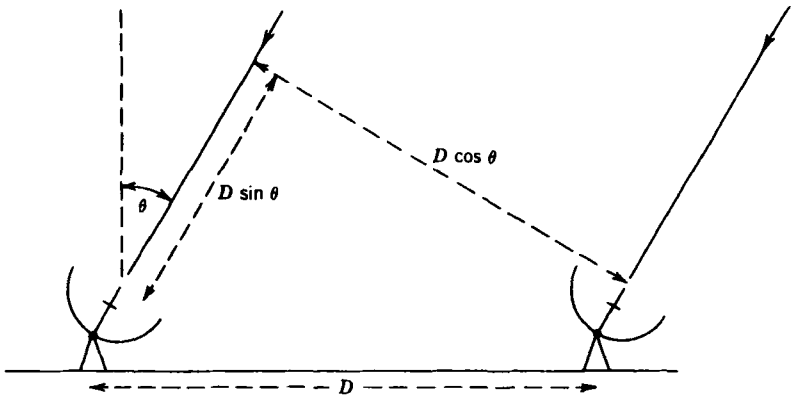


Figure 2.1 Geometry of an elementary interferometer. D is the interferometer baseline.

the frequency ν, the output of the multiplier is proportional to

$$\begin{aligned} F &= 2\sin(2\pi\nu t)\sin 2\pi\nu(t-\tau_g) \\ &= \cos 2\pi\nu\tau_g - \cos(4\pi\nu t)\cos(2\pi\nu\tau_g) - \sin(4\pi\nu t)\sin(2\pi\nu\tau_g). \end{aligned} \tag{2.1}$$

The center frequency of the receivers is generally in the range of tens of megahertz to hundreds of gigahertz. As the earth rotates the most rapid rate of variation of θ is equal to the earth's rotational velocity, which is of the order of 10^{-4} rad s^{-1}. Also, because D cannot be more than, say, 10^7 m for terrestrial baselines, the rate of variation of $\nu\tau_g$ is smaller than that of νt by at least six orders of magnitude. The more rapidly varying terms in Eq. (2.1) are easily filtered out, leaving the fringe function

$$F = \cos 2\pi\nu\tau_g = \cos\left(\frac{2\pi Dl}{\lambda}\right), \tag{2.2}$$

where $l = \sin\theta$; the definition of the variable l is discussed further in Section 2.4. For sidereal sources, the variation of θ with time as the earth rotates generates quasisinusoidal fringes at the correlator, which are the output of the interferometer. Figure 2.2 shows an example of this function, which can be envisaged as the directional power reception pattern of the interferometer for the case where the antennas either track the source or have isotropic responses, and thus do not affect the shape of the pattern.

An alternate and equivalent way of envisaging the formation of the sinusoidal fringes is to note that because of the rotation of the earth, the two antennas have different components of velocity in the direction of the source. The signals reaching the antennas thus suffer different Doppler shifts. When the signals are combined in the multiplying action of the receiving system, the sinusoidal output arises from the beats between the Doppler-shifted signals.

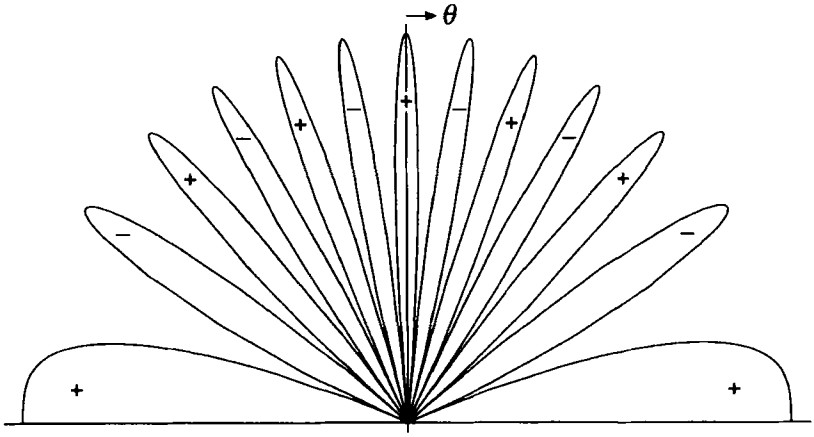


Figure 2.2 Polar plot to illustrate the fringe function $F = \cos[(2\pi D/\lambda)\sin\theta]$. The radial component is equal to $|F|$ and θ is measured with respect to the vertical axis. Alternate lobes correspond to positive and negative half-cycles of the quasisinusoidal fringe pattern, as indicated by the plus and minus signs. To simplify the diagram a very low value of 3 is used for D/λ. The increase in fringe width due to foreshortening of the baseline as $|\theta|$ increases is clearly shown. The maximum in the horizontal direction is a result of the arbitrary choice of an integer value for D/λ.

A development of the simple analysis can be made if we consider two Fourier components of the received signal at frequencies ν_1 and ν_2. These frequency components are statistically independent so that the interferometer output is the linear sum of the responses to each component. Hence the output has components F_1 and F_2, as in Eq. (2.2). For frequency ν_2 the coefficient $2\pi D/\lambda = 2\pi D\,\nu_2/c$ will be different from that for ν_1, so F_2 will have a different period from F_1 at any given angle θ. This difference in period gives rise to interference between F_1 and F_2, so that the fringe maxima have superimposed on them a modulation function that also depends on θ. Similar effects occur in the case of a continuous band of frequencies. For example, if the signals at the correlator are of uniform power spectral density over a band of width $\Delta\nu$ and center frequency ν_0, the output becomes

$$F(l) = \frac{1}{\Delta\nu}\int_{\nu_0-\Delta\nu/2}^{\nu_0+\Delta\nu/2} \cos\left(\frac{2\pi Dl\nu}{c}\right)d\nu$$

$$= \cos\left(\frac{2\pi Dl\nu_0}{c}\right)\frac{\sin(\pi Dl\Delta\nu/c)}{\pi Dl\Delta\nu/c}. \tag{2.3}$$

Thus the fringe pattern has an envelope in the form of a sinc function [sinc $(x) = \sin\pi x/\pi x$]. This is an example of the general result, to be discussed in the fol-

lowing section, that in the case of uniform power spectral density at the antennas the envelope of the fringe pattern is the Fourier transform of the instrumental frequency response.

2.2 EFFECT OF BANDWIDTH

Figure 2.3 shows an interferometer of the same general type as in Fig. 2.1 but with the amplifiers H_1 and H_2, the multiplier, and an integrator (with respect to time) shown explicitly. An instrumental time delay τ_i is inserted into one arm. Assume that for a point source each antenna delivers the same signal voltage $V(t)$ to the correlator, and that one voltage lags the other by a time delay $\tau = \tau_g - \tau_i$, as determined by the baseline D and the source direction θ. The integrator within the correlator has a time constant $2T$; that is, it sums the output from the multiplier for $2T$ seconds and then resets to zero after the sum is suitably recorded. The output of the correlator may be a voltage, a current, or a coded set of logic levels, but in any case it represents a physical quantity with the dimensions of voltage squared.

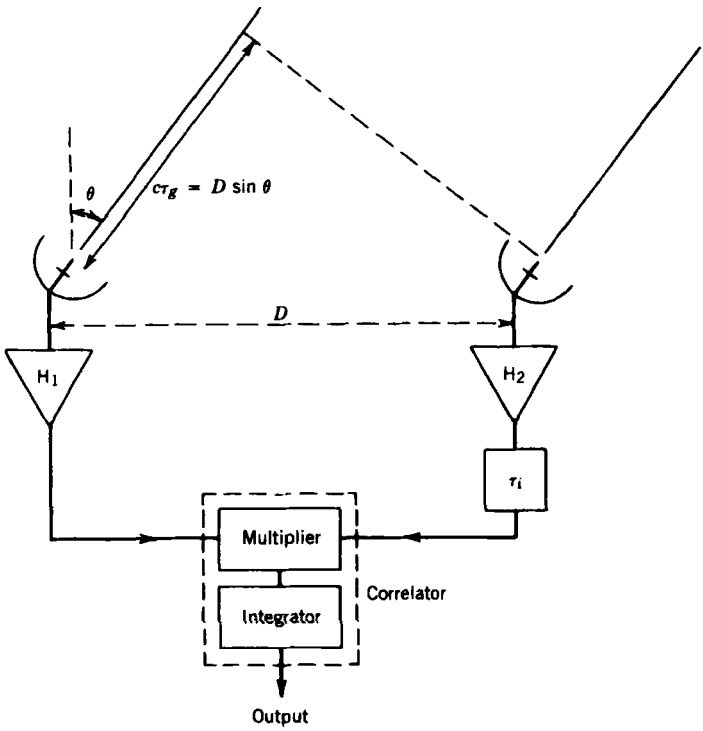


Figure 2.3 Elementary interferometer showing bandpass amplifiers H_1 and H_2, the geometric time delay τ_g, the instrumental time delay τ_i, and the correlator consisting of a multiplier and an integrator.

The output from the correlator resulting from a point source is

$$r = \frac{1}{2T}\int_{-T}^{T} V(t)V(t-\tau)\,dt. \tag{2.4}$$

We have ignored system noise and assumed that the two amplifiers have identical bandpass characteristics, including finite bandwidths $\Delta\nu$ outside which no frequencies are admitted. The integration time $2T$ is typically milliseconds to seconds, that is, very much larger than $\Delta\nu^{-1}$. Thus, Eq. (2.4) can be written as

$$r(\tau) = \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} V(t)V(t-\tau)\,dt, \tag{2.5}$$

which is an (unnormalized) autocorrelation function.* The condition $T \to \infty$ is satisfied if a very large number of variations of the signal amplitude, which have a duration $\sim\Delta\nu^{-1}$, occur in time $2T$. The integration time used in practice must clearly be finite and much less than the fringe period.

As described in Chapter 1, the signal from a natural cosmic source can be considered as a continuous random process that results in a broad spectrum, of which the phases are a random function of frequency. It will be assumed for our immediate purpose that the time-averaged amplitude of the cosmic signal in any finite band is constant with frequency over the passband of the receiver.

The squared amplitude of a frequency spectrum is known as the power density spectrum, or power spectrum. The power spectrum of a signal is the Fourier transform of the autocorrelation function of that signal. This statement is known as the Wiener–Khinchin relation, and is discussed further in Section 3.2. It applies to signals that are either deterministic or statistical in nature, and can be written

$$|H(\nu)|^2 = \int_{-\infty}^{\infty} r(\tau)e^{-j2\pi\nu\tau}\,d\tau, \tag{2.6}$$

and

$$r(\tau) = \int_{-\infty}^{\infty} |H(\nu)|^2 e^{j2\pi\nu\tau}\,d\nu, \tag{2.7}$$

where $H(\nu)$ is the amplitude (voltage) response, and hence $|H(\nu)|^2$ is the power spectrum of the signal input to the correlator. In this case, because the cosmic signal is assumed to have a spectrum of constant amplitude, the spectrum $H(\nu)$ is determined solely by the passband characteristics (frequency response) of the amplifiers. Thus the output of the interferometer as a function of the time delay τ

*For simplicity we consider only the signals from a point source, which are identical except for a time delay. In practical systems the input waveforms at the correlator may contain the partially correlated signals from a partially resolved source as well as instrumental noise. These nonidentical components can be taken into account by considering the *cross-correlation* function.

is the Fourier transform of the power spectrum of the cosmic signal as bandlimited by the amplifiers. Assume, as a simple example, a Gaussian passband centered at ν_0:

$$|H(\nu)|^2 = \frac{1}{2\sigma\sqrt{2\pi}}\left\{\exp\left[-\frac{(\nu-\nu_0)^2}{2\sigma^2}\right] + \exp\left[-\frac{(\nu+\nu_0)^2}{2\sigma^2}\right]\right\}, \tag{2.8}$$

where σ is the bandwidth factor (the full bandwidth at half-maximum level is $\sqrt{8\ln 2}\,\sigma$).

Note that to perform the Fourier transforms in Eqs. (2.6) and (2.7), we include a negative frequency response centered on $-\nu_0$. The spectrum is then symmetrical with respect to zero frequency, which is consistent with the fact that the autocorrelation function is real. The negative frequencies have no physical meaning but arise mathematically from the use of the exponential kernel of the transform. The interferometer response is

$$r(\tau) = e^{-2\pi^2\tau^2\sigma^2}\cos(2\pi\nu_0\tau), \tag{2.9}$$

which is illustrated in Fig. 2.4a. Note that $r(\tau)$ is a cosinusoidal function multiplied by an envelope function, in this case a Gaussian, whose shape and width depend on the amplifier passband. This envelope function is referred to as the *delay pattern*, *bandwidth pattern*, or fringe washing function.

By setting the instrumental delay τ_i to zero and substituting for the geometric delay $\tau_g = (D/c)\sin\theta$ in Eq. (2.9), we obtain the response

$$r(\tau_g) = \exp\left[-2\left(\frac{\pi D\sigma}{c}\sin\theta\right)^2\right]\cos\left(\frac{2\pi\nu_0 D}{c}\sin\theta\right). \tag{2.10}$$

The period of the fringes varies inversely as the quantity $\nu_0 D/c = D/\lambda$ and does not depend on the bandwidth parameter σ. The width of the bandwidth pattern, however, is a function of both σ and D; wide bandwidths and long baselines result in narrow fringe envelopes. This result is quite general. For example, a rectangular amplifier passband of width $\Delta\nu$, as considered in Eq. (2.3), results in an envelope pattern of the form $[\sin(\pi\Delta\nu\tau)]/(\pi\Delta\nu\tau)$, as shown in Fig. 2.4b.

In mapping applications the fringe envelope is often considered a nuisance, although there are some applications, particularly in very-long-baseline interferometry, in which the envelope is useful. In most cases it is desirable to observe the fringes in the vicinity of the maximum of the pattern, where the fringe amplitude is greatest. This condition can be achieved by changing the instrumental delay τ_i continuously or periodically so as to keep $\tau = \tau_g - \tau_i$ suitably small. If τ_i is adjusted in steps of the reciprocal of the center frequency† ν_0, the response

†This adjustment method is useful to consider here, but more commonly used methods are described in Section 7.3 under *Delay-Setting Tolerances*.

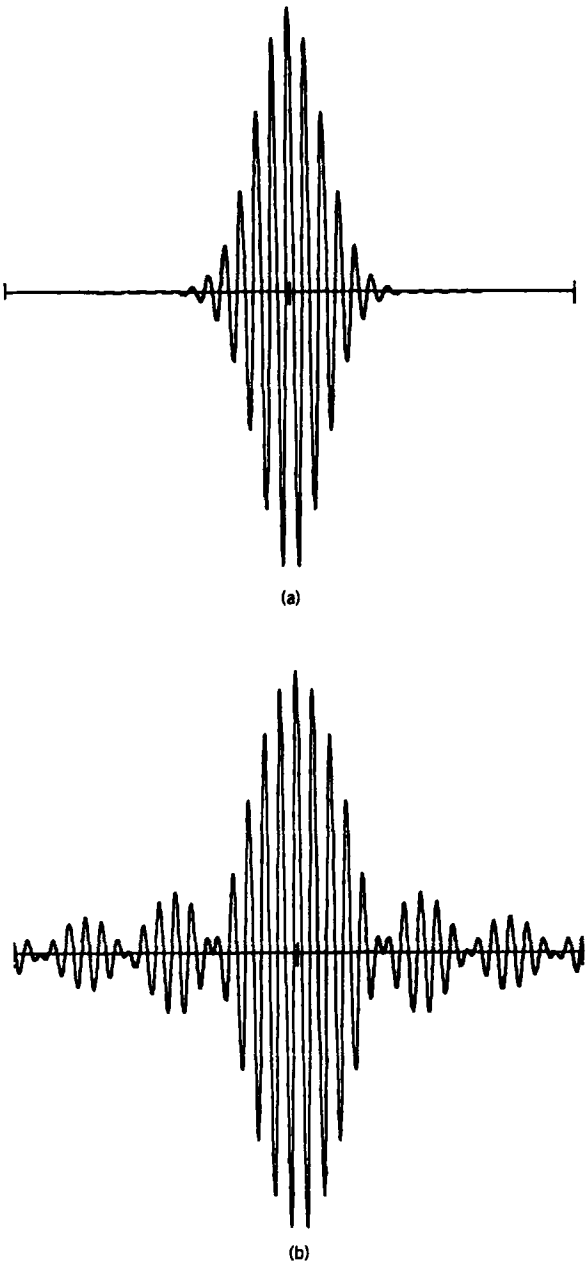


Figure 2.4 Point-source response of an interferometer with (a) Gaussian and (b) rectangular passbands. The abscissa is the geometric delay τ_g. The bandwidth pattern determines the envelope of the fringe term.

remains cosinusoidal with τ_g. Note that as $\Delta\nu$ approaches ν, the width of the envelope function becomes so narrow that only the central fringe remains. This occurs mainly in optics, where a central fringe of this type is called the "white light" fringe.

2.3 ONE-DIMENSIONAL SOURCE SYNTHESIS

Except for a few instruments built for low frequencies, most radio astronomy antennas operate at frequencies above ~1 GHz because broader bandwidths, which enhance sensitivity, and increased angular resolution are more practical at higher frequencies. The ability to track a source across the sky becomes important because the antenna beams become narrower as frequency increases, and also because the rotation of the earth is important in two-dimensional imaging, as illustrated in Fig. 1.15. In the analysis of an interferometer in which the antennas and the instrumental delay track the position of the source, it is convenient to specify angles of the antenna beam and other variables with respect to a reference position on the sky, usually the center or nominal position of the source under observation. This is commonly referred to as the *phase reference position*. Since the range of angles required to specify the source intensity distribution relative to this point is generally no more than a few degrees, small-angle approximations can be used to advantage. The instrumental delay is constantly adjusted to equal the geometric delay for radiation from the reference position. If we designate the reference position as the direction θ_0, then $\tau = \tau_g|_{\theta=\theta_0} - \tau_i = 0$, and $\tau_g|_{\theta=\theta_0} = (D/c)\sin\theta_0$. For radiation from a direction $(\theta_0 - \Delta\theta)$, where $\Delta\theta$ is a small angle, the fringe response term is

$$\begin{aligned}\cos(2\pi\nu_0\tau) &= \cos\left\{2\pi\nu_0\left[\frac{D}{c}\sin(\theta_0 - \Delta\theta) - \tau_i\right]\right\} \\ &\simeq \cos[2\pi\nu_0(D/c)\sin\Delta\theta\cos\theta_0]. \qquad (2.11)\end{aligned}$$

When observing a source at any position in the sky, the angular resolution of the fringes is determined by the length of the baseline projected onto a plane normal to the direction of the source. In Fig. 2.1, for example, this is the distance designated $D\cos\theta$. We therefore introduce a quantity u that is equal to the component of the antenna spacing normal to the direction of the reference position θ_0. u is measured in wavelengths, λ, at the center frequency ν_0, that is,

$$u = \frac{D\cos\theta_0}{\lambda} = \frac{\nu_0 D\cos\theta_0}{c}. \qquad (2.12)$$

Since $\Delta\theta$ is small, we can assume that the bandwidth pattern is near maximum (unity) in the direction $\theta_0 - \Delta\theta$. Then the response to radiation from that direction is, from Eqs. (2.11) and (2.12),

$$F(l) = \cos(2\pi\nu_0\tau) = \cos(2\pi ul), \qquad (2.13)$$

where $l = \sin \Delta\theta$. This is the response to a point source at $\theta = \theta_0 - \Delta\theta$ of an interferometer whose net delay $\tau_g - \tau_i$ is zero at $\theta = \theta_0$. As we shall show, the quantity u is interpreted as *spatial frequency*. It can be measured in cycles per radian, since the spatial variable l, being small, can be expressed in radians.

Interferometer Response as a Convolution

The response of a single antenna or an interferometer to a source can be expressed in terms of a convolution. Consider first the response of a single antenna and a receiver that measures the power received. Figure 2.5 shows the power reception pattern of the antenna $A(\theta)$, which is a polar plot of the effective area of the antenna as a function of angle from the center of the antenna beam. Also shown is the one-dimensional intensity profile of a source $I_1(\theta')$, as defined in Eq. (1.9), in which θ' is measured with respect to the center, or nominal position, of the source. The component of the output power in bandwidth $\Delta\nu$ contributed by each element $d\theta'$ of the source is $\frac{1}{2}\Delta\nu A(\theta'-\theta)I_1(\theta')d\theta'$, where the factor 1/2 takes account of the ability of the antenna to respond to only one component of randomly polarized radiation. The total output power from the antenna, omitting the constant factor $\frac{1}{2}\Delta\nu$, is proportional to

$$\int_{\text{source}} A(\theta' - \theta) I_1(\theta') d\theta'. \tag{2.14}$$

This integral is equal to the cross-correlation of the antenna reception pattern and the intensity distribution of the source. It is convenient to define $\mathcal{A}(\theta) =$

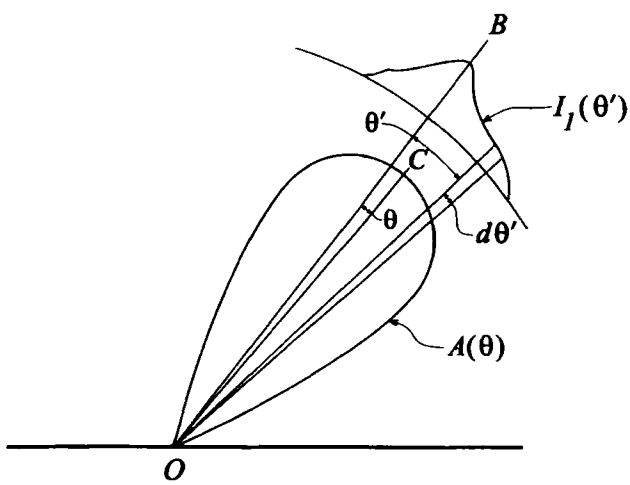


Figure 2.5 The power pattern of an antenna $A(\theta)$ and the intensity profile of a source $I_1(\theta')$ used to illustrate the convolution relationship. The angle θ is measured with respect to the beam center OC and θ' is measured with respect to the direction of the nominal position of the source OB.

$A(-\theta)$, where $\mathcal{A}$ is the mirror image of A with respect to θ. Then expression (2.14) becomes

$$\int_{\text{source}} \mathcal{A}(\theta - \theta')I_1(\theta')d\theta'. \tag{2.15}$$

The integral in expression (2.15) is an example of the *convolution integral*; see, for example, Bracewell (2000) or Champeney (1973). We can say that the output power of the antenna is given by the convolution of the source with the mirror image of the power reception pattern of the antenna. The mirror-image reception pattern can be described as the response of the antenna to a point source.

In the case of an interferometer we can express the response as a convolution by replacing the antenna power pattern in (2.15) by the overall power pattern of the interferometer. From the results presented earlier we find that the response of an interferometer is determined by three functions:

- The reception pattern of the antennas, which we represent as $A(l)$.
- The fringe pattern, $F(l)$, as in the example of Fig. 2.2 and given by Eq. (2.13). Note that the fringe term in the interferometer output, being the product of two voltages, is proportional to power.
- The bandwidth pattern, for example, as given by the sinc-function factor in Eq. (2.3). In the general case we can represent this by $F_B(l)$.

Note that these functions are all ideally symmetrical, and thus we can generally disregard the distinction between the interferometer power pattern and its mirror image in using the convolution relationship.

First consider an interferometer with tracking antennas and an instrumental delay that is adjusted so that the bandwidth pattern also tracks the source across the sky. In effect, the intensity distribution is modified by the antenna and bandwidth patterns. We can therefore envisage the output of the interferometer as the convolution of (the mirror image of) the fringe pattern with the modified intensity. In terms of the convolution integral the response can be written as

$$R(l) = \int_{\text{source}} \cos\left[2\pi u(l - l')\right] A(l')F_B(l')I_1(l')dl'. \tag{2.16}$$

It is often convenient to use the asterisk symbol ($*$) as a concise notation for convolution, with which Eq. (2.16) becomes

$$R(l) = \cos(2\pi ul) * [A(l)F_B(l)I_1(l)]. \tag{2.17}$$

(Note that convolution is commutative, i.e., $f * g = g * f$.) The intensity distribution measured with the interferometer is modified by $A(l)$ and $F_B(l)$, but since these are measurable instrumental characteristics, $I_1(l)$ can generally be recovered. In many cases the angular size of the source is small compared with the antenna beams and the bandwidth pattern, so these two functions introduce only a constant in the expression for the response. To simplify the discussion we

shall consider this case, and omitting constant factors, we can write the essential response of the interferometer as

$$R(l) = \cos(2\pi u l) * I_1(l). \tag{2.18}$$

In the case of the early interferometer shown in Fig. 1.6, in which the antennas are fixed in the meridian and do not track the source, the delays in the signal paths between the antennas and the point at which the signals are multiplied are equal and there is no variable instrumental delay. Thus the three functions that determine the interferometer power pattern are all fixed with respect to the interferometer baseline. The interferometer power pattern is of the form $A(l)\cos(2\pi u l)F_B(l)$. Then the response of the interferometer to the source is $[A(l)\cos(2\pi u l)F_B(l)] * I_1(l)$.

Interferometers with non-tracking antennas, as discussed above, are generally limited to frequencies no greater than a few hundred megahertz. At such long wavelengths it is possible to obtain antennas of large collecting area and still have wide enough beams that some minutes of observing time are obtained as the source passes through in sidereal motion. Generally the bandwidth of such low-frequency instruments is small so $F_B(l)$ is wide and can be omitted. Also, the antenna beams are usually wider than the source and sufficiently wide that several cycles of the fringe pattern can be measured as the source transits the beam. So in the non-tracking case the essential form of the response is also represented by Eq. (2.18). Except for a few instruments built especially for low-frequency observing, non-tracking antennas are mainly a feature of the early years of radio astronomy.

Convolution Theorem and Spatial Frequency

We now examine the interferometer response, as given in Eq. (2.18), using the *convolution theorem* of Fourier transforms. This theorem states that the Fourier transform of the convolution of two functions is the product of their Fourier transforms:

$$f * g \rightleftharpoons \mathcal{F}\mathcal{G}, \tag{2.19}$$

where $f \rightleftharpoons \mathcal{F}$, $g \rightleftharpoons \mathcal{G}$, and $\rightleftharpoons$ indicates Fourier transformation. A proof of the convolution theorem can be found in almost any text on Fourier transforms. Consider the Fourier transforms with respect to l and u of the three functions in Eq. (2.18). For the interferometer response we have $r(u) \rightleftharpoons R(l)$. For a particular value $u = u_0$, the Fourier transform of the fringe term is given by

$$\cos(2\pi u_0 l) \rightleftharpoons \tfrac{1}{2}\left[\delta(u + u_0) + \delta(u - u_0)\right], \tag{2.20}$$

where δ is the delta function. The Fourier transform of $I_1(l)$ is the visibility function $\mathcal{V}(u)$. Thus from Eqs. (2.18), (2.19), and (2.20), we obtain

$$\begin{aligned} r(u) &= \tfrac{1}{2}\left[\delta(u + u_0) + \delta(u - u_0)\right]\mathcal{V}(u) \\ &= \tfrac{1}{2}\left[\mathcal{V}(-u_0)\delta(u + u_0) + \mathcal{V}(u_0)\delta(u - u_0)\right]. \end{aligned} \tag{2.21}$$

This result shows that the instantaneous output of the interferometer as a function of spatial frequency consists of two delta functions situated at plus and minus u_0 on the u axis. Now $\mathcal{V}(u)$, the Fourier transform of $I_1(l)$, represents the amplitude and phase of the sinusoidal component of the intensity profile with spatial frequency u cycles per radian. The interferometer acts as a filter that responds only to spatial frequencies $\pm u_0$. The negative spatial frequency $-u_0$, like the negative frequencies in Eq. (2.8), has no physical meaning. It arises from the use, for mathematical convenience, of the exponential Fourier transform rather than the sine and cosine transforms, which correspond more directly to the physical situation. As a result, the spatial frequency spectra are symmetrical about the origin in the hermitian sense, that is, with even real parts and odd imaginary parts, which is appropriate since the intensity is a real, not complex, quantity.

Fringe visibility, as originally defined by Michelson [$\mathcal{V}_M$, see Eq. (1.8)], is a real quantity and is normalized to unity for an unresolved source. Complex visibility (Bracewell 1958) was defined to take account of the phase of the visibility, measured as the fringe phase, to allow mapping of asymmetric and complicated sources. The normalization is convenient when comparing measurements with simple models, as shown in Fig. 1.5. However, in maps or images it is desirable to display the magnitude of the intensity or brightness temperature, so the general practice is to retain the measured value of visibility, without normalization, since this incorporates the required information. Thus visibility $\mathcal{V}$ as used here is an unnormalized complex quantity with units of flux density (W m^{-2} Hz^{-1}). The quantity u, which was introduced as the projected baseline in wavelengths, is seen also to represent the spatial frequency of the Fourier components of the intensity. The concepts of spatial frequency and spatial frequency spectra are fundamental to the Fourier synthesis of astronomical images, and this general subject is discussed in a seminal paper by Bracewell and Roberts (1954).

Example of One-Dimensional Synthesis

To illustrate the observing process outlined in this chapter, we present a rudimentary simulation of measurements of the complex visibility of a source using arbitrary parameters. The source consists of two components separated by 0.34° of angle, the flux densities of which are in the ratio 2 : 1. The measurements are made with pairs of antennas placed along a line parallel to the direction of separation of the two components. Measurements are made for antenna spacings that are integral multiples of a unit spacing of 30 wavelengths. All spacings from one to 23 times the unit spacing are measured. These results could be obtained using two antennas and a single correlator, observing the source as it transits the meridian on 23 different days, and moving the antennas to provide a new spacing each day. Alternatively, the 23 measurements could be made simultaneously using 23 correlators and a number of antennas that could be as small as eight, if they were set out with minimum redundancy in the spacings, as discussed in Section 5.5. The angular sizes of the two components of the source are too small to be resolved by the interferometer, so they can be regarded as point radiators. The two

components radiate noise, and their two outputs are uncorrelated. The source is at a sufficient distance that incoming wavefronts can be considered to be plane over the measurement baselines.

Figures 2.6a and 2.6b show, respectively, the amplitude[‡] and phase of the visibility function as it would be measured. Since the data are derived from a model, there are no measurement errors, so the points indicate samples of the Fourier transform of the source intensity distribution, which can be represented by two delta functions with strengths in the ratio 1 : 2. Taking the inverse transform of the visibility yields the synthesized image of the source in Fig. 2.6c. The two components of the source are clearly represented. The extraneous oscillations arise from the finite extent of the visibility measurements, which are uniformly weighted out to a cutoff at 23 times the unit spacing. This effect is further illustrated in Fig. 2.6d, which shows the response of the measurement procedure to a single point source; equivalently, it is the synthesized beam. The profile of this response is the sinc function that is the Fourier transform of the rectangular window function, which represents the cutoff of the measurements at the longest spacing. In the image domain the double-source profile can be viewed as the convolution of the source with the point-source response. The point-source nature of the model components maximizes the sidelobe oscillations, which would be partially smoothed out if the width of the components were comparable to that of the sidelobes.

As is clear from the convolution relationship, information on the structure of the source is contained in the whole response pattern in Fig. 2.6c, that is, in the sidelobe oscillations as well as the main-beam peaks. A way to extract the maximum information on the source structure would be to fit scaled versions of the response in Fig. 2.6d to the two peaks in Fig. 2.6c, and then subtract them from the profile. In an actual observation this would leave the noise and any structure that might be present in addition to the point sources, but would remove all or most of the sidelobes. The fitting of the point-source responses could be adjusted to minimize some measure of the residual fluctuations, and further components could be fitted to any remaining peaks and subtracted. This technique would clearly be a good way to estimate the strengths and positions of the two components, and look for evidence of any low-level structure that could be hidden by the sidelobes in Fig 2.6c. The algorithm CLEAN, which is discussed in Chapter 11, uses this principle, but also replaces the components that are removed by model beam responses that are free of sidelobes. Removal of the sidelobes allows any lower-level structure to be investigated, down to the level of the noise. Most synthesis images are processed by nonlinear algorithms of this type, and the range of intensity levels achieved in some two-dimensional images exceeds 10^5 to 1.

[‡]It is arguable that the modulus of the complex visibility should be referred to as *magnitude* rather than *amplitude* since the dimensions of visibility include power rather than voltage. However, the term *visibility amplitude* is widely used in radio astronomy, probably resulting from the early practice of recording the fringe pattern as a quasisinusoidal waveform, and subsequently analyzing the amplitude and phase of the oscillations.

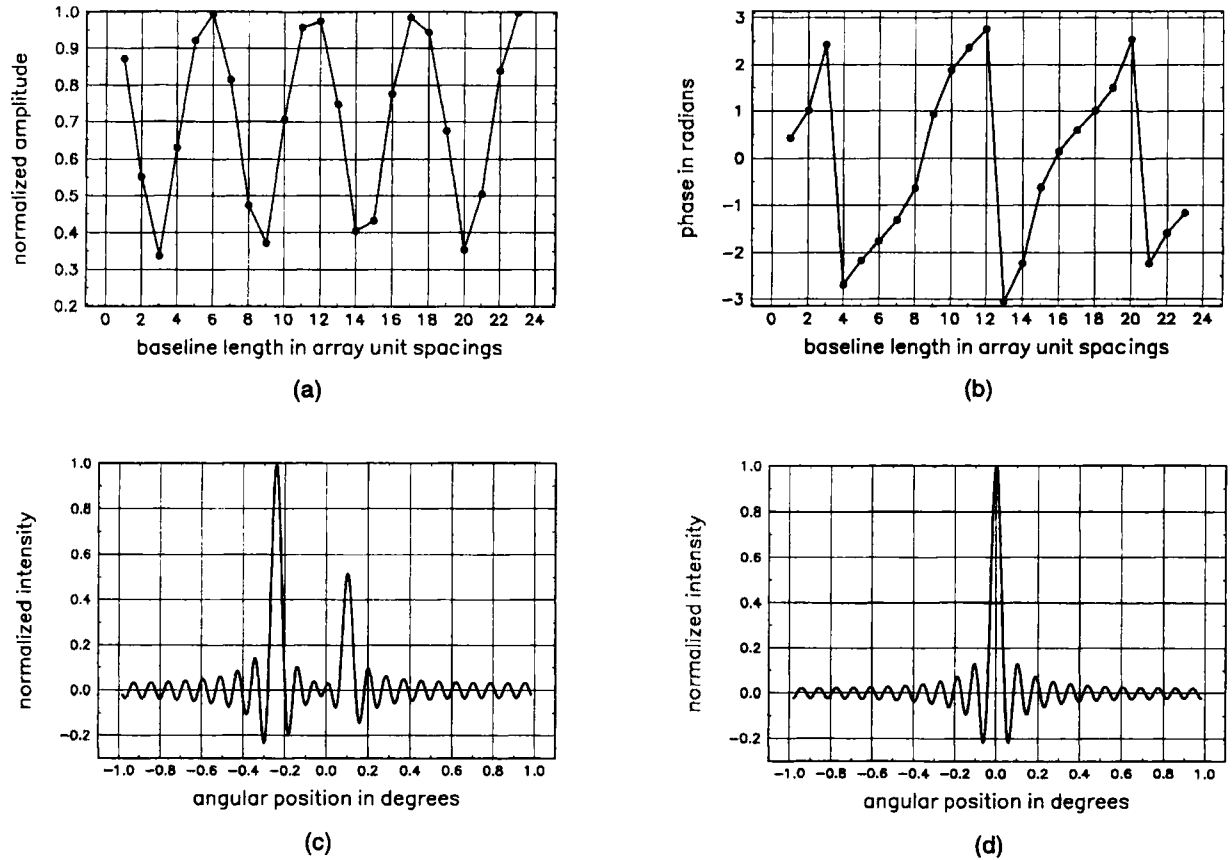


Figure 2.6 Simulated measurements of visibility of a double source: (a) visibility amplitude and (b) visibility phase, each plotted as a function of antenna spacing as a multiple of the unit spacing; (c) the profile derived from the measurements and (d) the response to a point source.

2.4 TWO-DIMENSIONAL SYNTHESIS

To synthesize a map or image of a source in two dimensions on the sky requires measurement of the two-dimensional spatial frequency spectrum in the (u, v) plane, where v is the north–south component as shown in Fig. 2.7a. Similarly, it is necessary to define a two-dimensional coordinate system (l, m) on the sky. The (l, m) origin is the reference position, or phase reference position, introduced in the last section. In considering functions in one dimension in the earlier part of this chapter, it was possible to define l in Eq. (2.2) as the sine of an angle. In two-dimensional analysis l and m are defined as the cosines of the angles between the direction (l, m) and the u and v axes, respectively, as shown in Fig. 2.7c. If the angle between the direction (l, m) and the w axis is small, l and m can be considered as the components of this angle measured in radians in the east–west and north–south directions, respectively.

For a source near the celestial equator, measuring the visibility as a function of u and v requires observing with a two-dimensional array of interferometers, that is, an array in which the baselines between pairs of antennas contain components in the north–south as well as the east–west directions. Although we have considered only east–west baselines, the results derived in terms of angles mea-

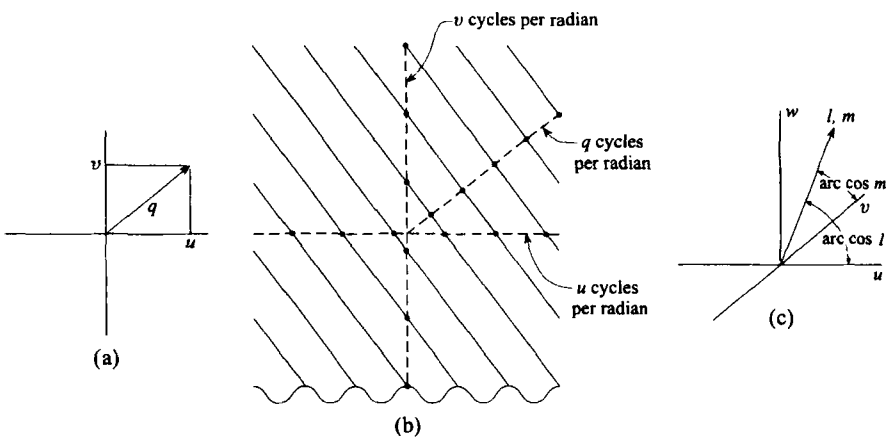


Figure 2.7 (a) The (u, v) plane in which the arrow point indicates the spatial frequency, q cycles per radian, of one Fourier component of an intensity map (or image) of a radio source. The components u and v of the spatial frequency are measured along axes in the east–west and north–south directions, respectively. (b) The (l, m) plane in which a single component of spatial frequency in the intensity domain has the form of sinusoidal corrugations on the sky. The figure shows corrugations that represent one such component. The diagonal lines indicate the ridges of maximum intensity. The dots indicate the positions of these maxima along lines in three directions. In a direction normal to the ridges the frequency of the oscillations is q cycles per radian, and in directions parallel to the u and v axes it is u and v cycles per radian, respectively. (c) The u and v coordinates define a plane and the w coordinate is perpendicular to it. The coordinates (l, m) are used to specify a direction on the sky in two dimensions. l and m are defined as the cosines of the angles made with the u and v axes, respectively.

sured with respect to a plane that is normal to the baseline hold for any baseline direction.

A source at a high declination (near the celestial pole) can be mapped in two dimensions with either one- or two-dimensional arrays, as shown in Fig. 1.15 and further explained in Section 4.1. As the earth rotates, the baseline projection on the celestial sphere rotates and foreshortens. A plot of the variation of the length and direction of the projected baseline as the antennas track the source across the sky is an arc of an ellipse in the (u, v) plane. The parameters of the ellipse depend on the declination of the source, the length and orientation of the baseline, and the latitude of the center of the baseline. In the design of a synthesis array, the relative positions of the antennas are chosen to provide a distribution of measurements in u and v consistent with the angular resolution, field of view, declination range, and sidelobe level required, as discussed in Chapter 5. The two-dimensional intensity distribution is then obtained by taking a two-dimensional Fourier transform of the observed visibility, $\mathcal{V}(u, v)$.

Projection-Slice Theorem

Some important relationships between one-dimensional and two-dimensional functions of intensity and visibility are summarized in Fig. 2.8, which illustrates the "projection-slice" theorem of Fourier transforms (Bracewell 1956, 1995, 2000). At the top left is the two-dimensional intensity distribution of a source $I(l, m)$, and at the bottom right is the corresponding visibility function $\mathcal{V}(u, v)$. These two functions are related by a two-dimensional Fourier transform, as indicated on the arrows shown between them. Note the general property of Fourier transforms that the width in one domain is inversely related to the width in the other domain. At the lower left is the projection of $I(l, m)$ on the l axis, which is equal to the one-dimensional intensity distribution $I_1(l)$. This projection is obtained by line integration along lines parallel to the m axis, as defined in Eq. (1.9). I_1 is related by a one-dimensional Fourier transform to the visibility measured along the u axis at the lower right, that is, the profile of a slice $\mathcal{V}(u, 0)$ through the visibility function $\mathcal{V}(u, v)$, indicated by the shaded area in the diagram. $\mathcal{V}(u, 0)$ could be measured, for example, by observations of a source made at meridian transit with a series of interferometer baselines in an east–west direction. This relationship was encountered in Chapter 1 in the description of the Michelson interferometer, and examples of such pairs of functions are shown in Fig. 1.5. At the upper right is a projection of $\mathcal{V}(u, v)$ on the u axis, $\mathcal{V}_1(u) = \int \mathcal{V}(u, v)\, dv$, and this is related by a one-dimensional Fourier transform to a slice profile of the source intensity $I(l, 0)$ along the l axis at the upper left, indicated by the shaded area. The relationships between the projections and slices are not confined to the u and l axes, but apply to any sets of axes that are parallel in the two domains. For example, integration of $I(l, m)$ along lines parallel to OP results in a curve, the Fourier transform of which is the profile of a slice through $\mathcal{V}(u, v)$ along the line QR.

The relationships in Fig. 2.8 apply to Fourier transforms in general, and their application to radio astronomy was recognized during the early development

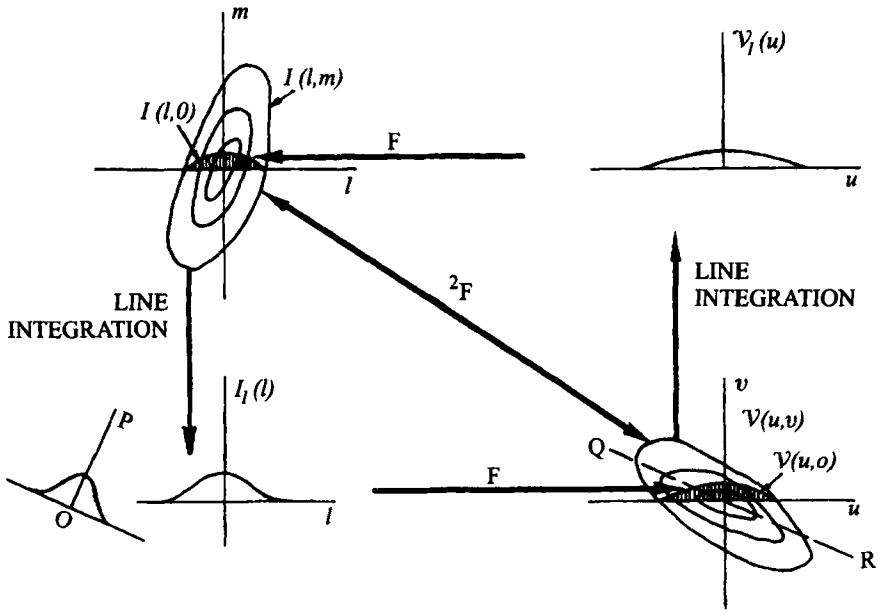


Figure 2.8 Illustration of the "projection-slice" theorem, which explains the relationships between one-dimensional projections and cross sections of intensity and visibility functions. One-dimensional Fourier transforms are organized horizontally and projections vertically. The symbols F and ^{2}F indicate one-dimensional and two-dimensional Fourier transforms, respectively. See the text for further explanation. From Bracewell, Strip Integration in Radio Astronomy, courtesy *Aust. J. Phys.* (Vol. 9, p. 208, 1956).

of the subject. For example, in determining the two-dimensional intensity of a source from a series of fan-beam scans at different angles, one can perform one-dimensional transforms of the scans to obtain values of $\mathcal{V}$ along a series of lines through the origin of the (u, v) plane, thus obtaining the two-dimensional visibility $\mathcal{V}(u, v)$. Then $I(l, m)$ can be obtained by two-dimensional Fourier transformation. In the early years of radio astronomy, before computers were widely available, such computation was a very laborious task and various alternative procedures for image formation from fan-beam scans were devised (Bracewell 1956, Bracewell and Riddle 1967).

REFERENCES

Bracewell, R. N., Radio Interferometry of Discrete Sources, *Proc. IRE*, **46**, 97–105, 1958.

Bracewell, R. N., Strip Integration in Radio Astronomy, *Aust. J. Phys.*, **9**, 198–217, 1956.

Bracewell, R. N., *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 2000 (earlier eds. 1965, 1978).

Bracewell, R. N., *Two-Dimensional Imaging*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

Bracewell, R. N. and A. C. Riddle, Inversion of Fan Beam Scans in Radio Astronomy, *Astrophys. J.*, **150**, 427–434, 1967.

Bracewell, R. N. and J. A. Roberts, Aerial Smoothing in Radio Astronomy, *Aust. J. Phys.*, **7**, 615–640, 1954.

Champeney, D. C., *Fourier Transforms and Their Physical Applications*, Academic Press, London, 1973.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

3 Analysis of the Interferometer Response

In this chapter we introduce the full two-dimensional analysis of the interferometer response, without small-angle assumptions, and then investigate the small-field simplifications commonly used in the transformation from the measured visibility to the intensity distribution. This is followed by a discussion of the relationship between the cross-correlation of the received signals and the cross power spectrum, which results from the Wiener–Khinchin relation and is fundamental to spectral line interferometry. An analysis of the basic response of the receiving system is also given. An appendix considers some approaches to the representation of noise-like signals, including the analytic signal.

3.1 FOURIER TRANSFORM RELATIONSHIP BETWEEN INTENSITY AND VISIBILITY

We begin by deriving the relationship between intensity and visibility in a coordinate-free form and then show how the choice of a coordinate system results in an expression in the familiar form of the Fourier transform. Suppose that the antennas track the source under observation, which is the most common situation, and let the unit vector $\mathbf{s}_0$ in Fig. 3.1 indicate the phase reference position introduced in Section 2.3. This position, sometimes also known as the phase-tracking center, becomes the center of the field to be mapped. An element of the source of solid angle $d\Omega$ at position $\mathbf{s} = \mathbf{s}_0 + \boldsymbol{\sigma}$ contributes a component of power $\frac{1}{2}A(\boldsymbol{\sigma})I(\boldsymbol{\sigma})\Delta\nu d\Omega$ at each of the two antennas, where $A(\boldsymbol{\sigma})$ is the effective collecting area of each antenna, $I(\boldsymbol{\sigma})$ is the source intensity distribution as observed from the distance of the antennas, and $\Delta\nu$ is the bandwidth of the receiving system. It is easily seen that this expression has the dimensions of power since the units of I are W m^{-2} Hz^{-1} sr^{-1}. From the considerations outlined in the derivation of Eqs. (2.1) and (2.2), including the far-field condition for the source, the resulting component of the correlator output is proportional to the received power and to the fringe term $\cos(2\pi\nu\tau_g)$, where τ_g is the geometric delay. If the vector $\mathbf{D}_\lambda$ specifies the baseline measured in wavelengths, then $\nu\tau_g = \mathbf{D}_\lambda \cdot \mathbf{s} = \mathbf{D}_\lambda \cdot (\mathbf{s}_0 + \boldsymbol{\sigma})$. Thus the output from the correlator is represented by

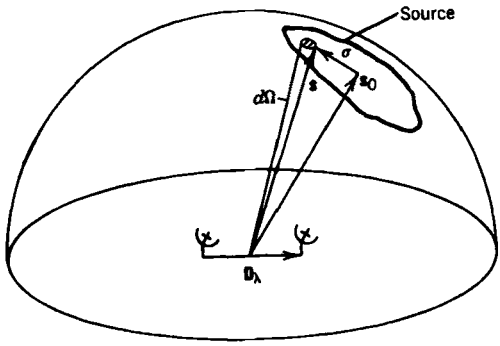


Figure 3.1 Baseline and position vectors that specify the interferometer and the source. The source is represented by the outline on the celestial sphere.

$$
\begin{aligned}
r(\mathbf{D}_\lambda, \mathbf{s}_0) &= \Delta\nu \int_{4\pi} A(\boldsymbol{\sigma}) I(\boldsymbol{\sigma}) \cos[2\pi \mathbf{D}_\lambda \cdot (\mathbf{s}_0 + \boldsymbol{\sigma})]\, d\Omega \\
&= \Delta\nu \cos(2\pi \mathbf{D}_\lambda \cdot \mathbf{s}_0) \int_{4\pi} A(\boldsymbol{\sigma}) I(\boldsymbol{\sigma}) \cos(2\pi \mathbf{D}_\lambda \cdot \boldsymbol{\sigma})\, d\Omega \\
&\quad - \Delta\nu \sin(2\pi \mathbf{D}_\lambda \cdot \mathbf{s}_0) \int_{4\pi} A(\boldsymbol{\sigma}) I(\boldsymbol{\sigma}) \sin(2\pi \mathbf{D}_\lambda \cdot \boldsymbol{\sigma})\, d\Omega. \qquad (3.1)
\end{aligned}
$$

Note that the integration of the response to the element $d\Omega$ over the source in Eq. (3.1) requires the assumption that the source is spatially incoherent, that is, that the radiated waveforms from different elements $d\Omega$ are uncorrelated. This assumption is justified for essentially all cosmic radio sources. Spatial coherence is discussed further in Section 14.2. Let A_0 be the antenna collecting area in direction $\mathbf{s}_0$ in which the beam is pointed. We introduce a normalized reception pattern $A_N(\boldsymbol{\sigma}) = A(\boldsymbol{\sigma})/A_0$ and consider the modified intensity distribution $A_N(\boldsymbol{\sigma}) I(\boldsymbol{\sigma})$. Now we define the complex visibility* as

$$
\mathcal{V} = |\mathcal{V}| e^{j\phi_v} = \int_{4\pi} A_N(\boldsymbol{\sigma}) I(\boldsymbol{\sigma}) e^{-j2\pi \mathbf{D}_\lambda \cdot \boldsymbol{\sigma}}\, d\Omega. \qquad (3.2)
$$

Then by separating the real and imaginary parts we obtain

$$
\int_{4\pi} A_N(\boldsymbol{\sigma}) I(\boldsymbol{\sigma}) \cos(2\pi \mathbf{D}_\lambda \cdot \boldsymbol{\sigma})\, d\Omega = |\mathcal{V}| \cos\phi_v, \qquad (3.3)
$$

*In formulating the fundamental Fourier transform relationship in synthesis mapping, which follows from Eq. (3.2), we use the negative exponent to derive the complex visibility function (or mutual coherence function) from the intensity distribution, and the positive exponent for the inverse operation. From a physical viewpoint the choice is purely arbitrary, and the literature contains examples of both this and the reverse convention. Our choice follows Born and Wolf (1999) and Bracewell (1958).

$$\int_{4\pi} A_N(\boldsymbol{\sigma}) I(\boldsymbol{\sigma}) \sin(2\pi \mathbf{D}_\lambda \cdot \boldsymbol{\sigma}) d\Omega = -|\mathcal{V}| \sin \phi_v, \tag{3.4}$$

and from Eq. (3.1)

$$r(\mathbf{D}_\lambda, \mathbf{s}_0) = A_0 \Delta\nu |\mathcal{V}| \cos(2\pi \mathbf{D}_\lambda \cdot \mathbf{s}_0 - \phi_v). \tag{3.5}$$

Thus the output of the correlator can be expressed in terms of a fringe pattern corresponding to that for a hypothetical point source in the direction $\mathbf{s}_0$, which is the phase reference position. As noted earlier, this is usually the center or nominal position of the source to be mapped. The modulus and phase of $\mathcal{V}$ are equal to the amplitude and phase of the fringes; the phase is measured relative to the fringe

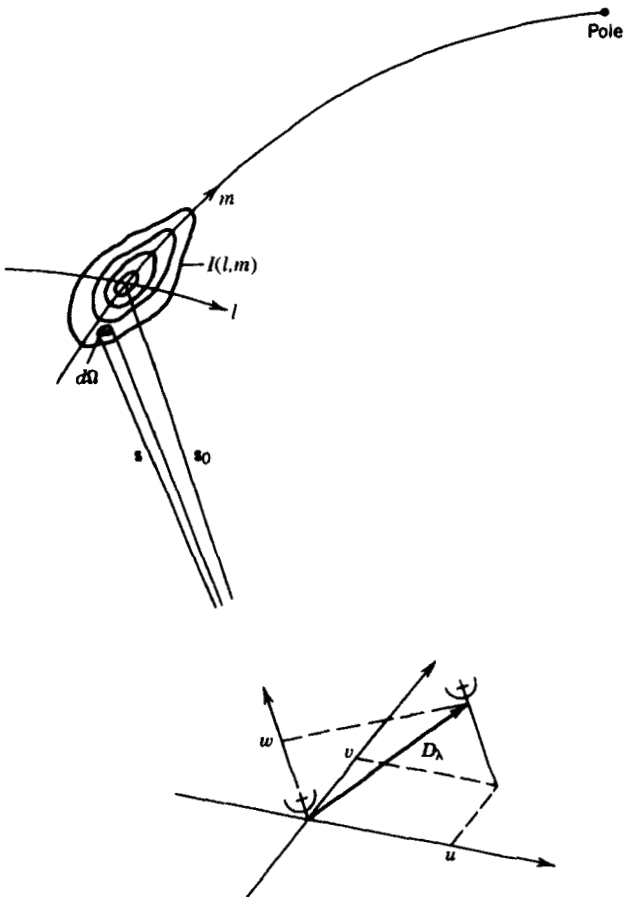


Figure 3.2 Geometric relationship between a source under observation $I(l, m)$ and an interferometer or one antenna pair of an array. The antenna baseline vector, measured in wavelengths, has length D_λ and components (u, v, w).

phase for the hypothetical source. As defined above, $\mathcal{V}$ has the dimensions of flux density (W m^{-2} Hz^{-1}), which is consistent with its Fourier transform relationship with I. Some authors have defined visibility as a normalized, dimensionless quantity, in which case it is necessary to reintroduce the intensity scale in the resulting image. Note that the bandwidth has been assumed to be small compared to the center frequency in deriving Eq. (3.5).

In introducing a coordinate system, the geometry that we now consider is illustrated in Fig. 3.2. The two antennas track the center of the field to be mapped. They are assumed to be identical, but if they differ, $A_N(\boldsymbol{\sigma})$ is the geometric mean of the beam patterns of the two antennas. The magnitude of the baseline vector is measured in wavelengths at the center frequency of the observing band, and the baseline has components (u, v, w) in a right-handed coordinate system, where u and v are measured in a plane normal to the direction of the phase reference position. The spacing component v is measured toward the north as defined by the plane through the origin, the source, and the pole, and u toward the east. The component w is measured in the direction $\mathbf{s}_0$, which is the phase reference position. On Fourier transformation, the phase reference position becomes the origin of the derived intensity distribution $I(l, m)$, where l and m are direction cosines measured with respect to the axes u and v. In terms of these coordinates, we find

$$\begin{aligned} \mathbf{D}_\lambda \cdot \mathbf{s}_0 &= w \\ \mathbf{D}_\lambda \cdot \mathbf{s} &= \left(ul + vm + w\sqrt{1 - l^2 - m^2}\right) \\ d\Omega &= \frac{dl\,dm}{\sqrt{1 - l^2 - m^2}}, \end{aligned} \tag{3.6}$$

where $\sqrt{1 - l^2 - m^2}$ is equal to the third direction cosine n measured with respect to the w axis†. Note also that $\mathbf{D}_\lambda \cdot \boldsymbol{\sigma} = \mathbf{D}_\lambda \cdot \mathbf{s} - \mathbf{D}_\lambda \cdot \mathbf{s}_0$. Thus from Eq. (3.2):

$$\begin{aligned} \mathcal{V}(u, v, w) = &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A_N(l, m) I(l, m) \\ &\times \exp\left\{-j2\pi\left[ul + vm + w\left(\sqrt{1 - l^2 - m^2} - 1\right)\right]\right\} \frac{dl\,dm}{\sqrt{1 - l^2 - m^2}}. \end{aligned} \tag{3.7}$$

†The expression for $d\Omega$ is obtained by considering the unit sphere centered on the (u, v, w) origin. A point P on the sphere with coordinates (u, v, w) is projected onto the (u, v) plane at $u = l$, $v = m$, and the increments dl, dm define a column of square cross section running through $(u, v, 0)$ parallel to the w axis. The column makes an angle $\cos^{-1} n$ with the normal to the spherical surface at P, and $d\Omega$ is equal to the surface area intersected by the column, which is $dl\,dm/n$, or $dl\,dm/\sqrt{1 - l^2 - m^2}$. Alternatively, the solid angle can be expressed in polar coordinates as $d\Omega = \sin\theta\, d\theta\, d\phi$, where θ and ϕ are the polar and azimuthal angles in the (u, v, w) plane, that is, $\theta = \sin^{-1}\sqrt{l^2 + m^2}$ and $\phi = \tan^{-1} m/l$. Calculation of the Jacobian of the transformation from (θ, ϕ) coordinates to (l, m) coordinates gives the result $d\Omega = dl\,dm/\sqrt{1 - l^2 - m^2}$ (Apostol 1962).

A factor $e^{j2\pi w}$ on the right-hand side in Eq. (3.7) results from the measurement of angular position with respect to the w axis. For a source on the w axis, $l = m = 0$, and the argument of the exponential term in Eq. (3.7) is zero. For any other source, the fringe phase is measured relative to that for a source on the w axis, which is the phase reference position, $\mathbf{s}_0$. The function $A_N I$ in Eq. (3.7) is zero for $l^2 + m^2 \geq 1$, and in practice it usually falls to very low values for directions outside the field to be mapped as a result of the antenna beam pattern, the bandwidth pattern, or the finite size of the source. Thus we can extend the limits of integration to $\pm\infty$. Note, however, that Eq. (3.7) requires no small-angle assumptions. The reason why we use direction cosines rather than a linear measure of angle in interferometer theory is that they occur in the exponential term of this relationship.

The coordinate system (l, m) defined above is a convenient one in which to present an intensity distribution. It corresponds to the projection of the celestial sphere onto a plane that is a tangent at the field center, as shown in Fig. 3.3. The distance of any point in the map from the (l, m) origin is proportional to the sine of the corresponding angle on the sky, so for small fields distances on the map are closely proportional to the corresponding angles. The same relationship usually applies to the field of an optical telescope. For a detailed discussion of relationships on the celestial sphere and tangent planes, see König (1962).

If all the measurements could be made with the antennas in a plane normal to the w direction so that $w = 0$, Eq. (3.7) would reduce to an exact two-dimensional Fourier transform. In general this is not possible, and we now consider ways in which the transform relationship can be applied. Recall first that the basis of the synthesis mapping process is the measurement of $\mathcal{V}$ over a wide range of u and v. For a ground-based array this can be achieved by varying the length and direction of the antenna spacing and also by tracking the field-center position as the earth rotates. The rotation causes the projection of $\mathbf{D}_\lambda$ to move across the (u, v) plane. Thus an observation often lasts for 6–12 h. As the earth's rotation carries the

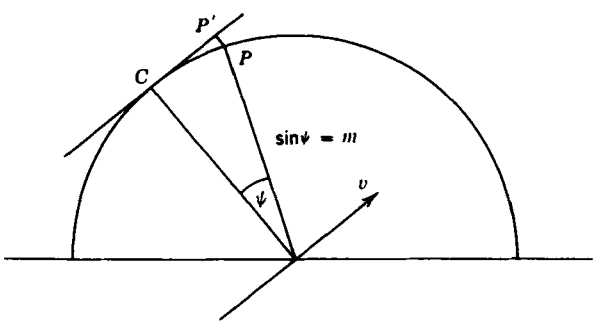


Figure 3.3 Mapping of the celestial sphere onto a plane, shown in one dimension. The position of the point P is measured in terms of the direction cosine m with respect to the v axis. When projected onto a plane surface with a scale linear in m, P appears at P' at a distance from the field center C proportional to $\sin\psi$.

antennas through space, the baseline vector remains in a plane only if $\mathbf{D}_\lambda$ has no component parallel to the rotation axis, that is, the baseline is an east–west line on the earth's surface. In the general case there is a three-dimensional distribution of the measurements of $\mathcal{V}$. The simplest form of the transform relationship that can then be used is based on an approximation that is valid so long as the synthesized field is not too large. If l and m are small enough that the term

$$\left(\sqrt{(1-l^2-m^2)}-1\right)w \simeq -\tfrac{1}{2}(l^2+m^2)w \tag{3.8}$$

can be neglected, Eq. (3.7) becomes

$$\mathcal{V}(u,v,w) \simeq \mathcal{V}(u,v,0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{A_N(l,m)I(l,m)}{\sqrt{1-l^2-m^2}} e^{-j2\pi(ul+vm)}\,dl\,dm. \tag{3.9}$$

Thus for a restricted range of l and m, $\mathcal{V}(u,v,w)$ is approximately independent of w, and for the inverse transform we can write

$$\frac{A_N(l,m)I(l,m)}{\sqrt{1-l^2-m^2}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{V}(u,v)e^{j2\pi(ul+vm)}\,du\,dv. \tag{3.10}$$

With this approximation it is usual to omit the w dependence and write the visibility as the two-dimensional function $\mathcal{V}(u,v)$. Note that the factor $\sqrt{1-l^2-m^2}$ in Eqs. (3.9) and (3.10) can be subsumed into the function $A_N(l,m)$, if desired. Equation (3.10) is a form of the van Cittert–Zernike theorem, which originated in optics and is discussed in Section 14.1 under *Mutual Coherence of an Incoherent Source*.

The approximation in Eq. (3.9) introduces a phase error equal to 2π times the neglected term, that is, $\pi(l^2+m^2)w$. Limitation of this error to some tolerable value places a restriction on the size of the synthesized field, which can be estimated approximately as follows. If the antennas track the source under observation down to low elevation angles, the values of w can approach the maximum spacings $(D_\lambda)_{\max}$ in the array, as shown in Fig. 3.4. Also, if the spatial frequen-

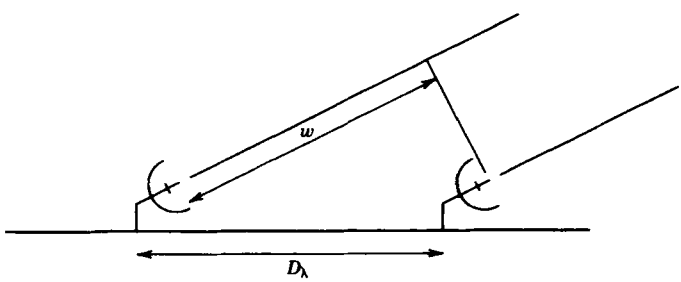


Figure 3.4 When observations are made at a low angle of elevation, and at an azimuth close to that of the baseline, the spacing component w becomes comparable to the baseline length D_λ, which is measured in wavelengths.

cies measured are evenly distributed out to the maximum spacing, the synthesized beamwidth θ_b is approximately equal to $(D_\lambda)^{-1}_{\max}$. Thus the maximum phase error is approximately

$$\pi \left(\frac{\theta_f}{2}\right)^2 \theta_b^{-1}, \tag{3.11}$$

where θ_f is the width of the synthesized field. The condition that no phase errors can exceed, say, 0.1 rad then requires that

$$\theta_f < \tfrac{1}{3}\sqrt{\theta_b}, \tag{3.12}$$

where the angles are measured in radians. For example, if $\theta_b = 1$ arcsec, $\theta_f < 2.5$ arcmin. Much synthesis mapping in astronomy is performed within this restriction, but ways of mapping larger fields will be discussed later.

We now return to the case of arrays with east–west spacings only, and discuss further the conditions for which we can put $w = 0$, and the resulting effects. Let us first rotate the (u, v, w) coordinate system about the u axis until the w axis points toward the pole as shown in Fig. 3.5. We indicate by primes the quantities measured in the rotated system. The (u', v') axes lie in a plane parallel to the earth's equator. The east–west antenna spacings contain components in this plane only (i.e., $w' = 0$), and as the earth rotates, the spacing vectors sweep out circles

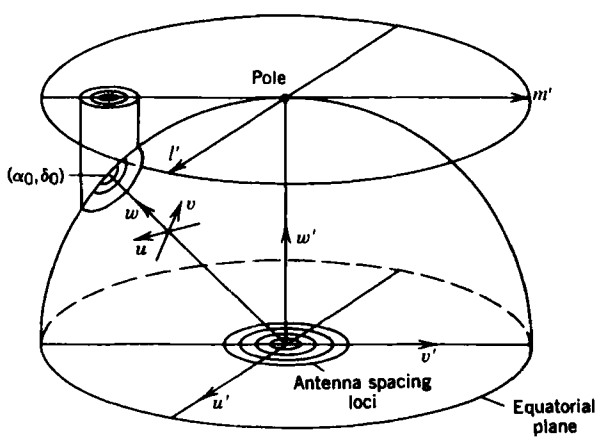


Figure 3.5 The (u', v', w') coordinate system for an east–west array. The (u', v') plane is the equatorial plane and the antenna spacing vectors trace out arcs of concentric circles as the earth rotates. Note that the directions of the u' and v' axes are chosen so that the v' axis lies in the plane containing the pole, the observer, and the point under observation (α_0, δ_0). In Fourier transformation from the (u', v') to the (l', m') planes the celestial hemisphere is mapped as a projection onto the tangent plane at the pole. The (u, v, w) coordinates for observation in the direction (α_0, δ_0) are also shown.

concentric with the (u', v') origin. From Eq. (3.7) we can write

$$\mathcal{V}(u', v') = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A_N(l', m')I(l', m')e^{-j2\pi(u'l'+v'm')}\frac{dl'\,dm'}{\sqrt{1-l'^2-m'^2}}, \tag{3.13}$$

where (l', m') are direction cosines measured with respect to (u', v'). Equation (3.13) holds for the whole hemisphere above the equatorial plane. The inverse transformation yields

$$\frac{A_N(l', m')I(l', m')}{\sqrt{1-l'^2-m'^2}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{V}(u', v')e^{j2\pi(u'l'+v'm')}\,du'\,dv'. \tag{3.14}$$

In this mapping the hemisphere is projected onto the tangent plane at the pole, as shown in Fig. 3.5. In practice, however, a map is usually confined to a small area within the antenna beams. In the vicinity of such an area, centered at right ascension and declination (α_0, δ_0), angular distances in the map are compressed by a factor $\sin\delta_0$ in the m' dimension. Also, in mapping the (α_0, δ_0) vicinity it is convenient if the origin of the angular position variables is shifted to (α_0, δ_0). Expansion of the scale and shift of the origin can be accomplished by the coordinate transformation

$$l = l', \qquad m'' = (m' - \cos\delta_0)\operatorname{cosec}\delta_0. \tag{3.15}$$

If we write $F(l', m')$ for the left-hand side of Eq. (3.14), then

$$F(l', m') \rightleftharpoons \mathcal{V}(u', v'), \tag{3.16}$$

and

$$F\left[l', (m' - \cos\delta_0)\operatorname{cosec}\delta_0\right] \rightleftharpoons \left|\sin\delta_0\right|\mathcal{V}(u', v'\sin\delta_0)e^{-j2\pi v'\cos\delta_0}, \tag{3.17}$$

where $\rightleftharpoons$ indicates Fourier transformation. Equation (3.17) follows from the behavior of Fourier pairs with change of variable and involves the application of the similarity and shift theorems [see, e.g., Bracewell (2000)]. The coordinates $(u', v'\sin\delta_0)$ on the right-hand side of Eq. (3.17) represent the projection of the equatorial plane onto the (u, v) plane, which is normal to the direction (α_0, δ_0). In the (u, v, w) system $u = u'$ and $v = v'\sin\delta_0$. The coordinate w shown in Fig. 3.5 is equal to $-v'\cos\delta_0$. Thus $e^{-j2\pi v'\cos\delta_0}$ in Eq. (3.17) is the same factor that occurs in Eq. (3.7) as a result of the measurement of visibility phase relative to that for a point source in the w direction. Equation (3.14) now becomes

$$\frac{A_N(l, m'')I(l, m'')}{\sqrt{1-l^2-m''^2}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{V}(u', v'\sin\delta_0)\left|\sin\delta_0\right|e^{-j2\pi v'\cos\delta_0} \times e^{j2\pi(u'l'+v'm')}\,du'\,dv'$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{V}(u, v) e^{j2\pi(ul+vm'')} \, du \, dv. \tag{3.18}$$

A similar analysis is given by Brouw (1971).

The derivation of Eq. (3.18) from Eq. (3.14) involves a redefinition of the m coordinate, but no approximations. Equation (3.18) is of the same form as Eq. (3.10) in which the term in Eq. (3.8) was neglected. Thus if we apply the mapping scheme of Eq. (3.10), which is based on omitting this term, to observations made with an east–west array, the phase errors introduced distort the map in a way that corresponds exactly to the change of definition of the m variable to m''. Since m'' is derived from a direction cosine measured from the v' axis in the equatorial plane, there is a progressive change in the north–south angular scale over the map. The factor cosec δ_0 in Eq. (3.15) establishes the correct angular scale at the center of the map, but this simple correction is acceptable only for small fields. The crucial point to note here is that when visibility data measured in a plane are projected into (u, v, w) coordinates, w is a linear function of u and v (and a linear function of v alone for east–west baselines). Hence the phase error $\pi(l^2 + m^2)w$ is linear in u and v. Phase errors of this kind have the effect of introducing position shifts in the resulting map, but there remains a one-to-one correspondence between points in the map and on the sky. The effect is simply to produce a predictable, and hence correctable, distortion of the coordinates.

It is clear from Fig. 3.5 that if all the measurements lie in the (u', v') plane, then the values of v in the (u, v) plane become seriously foreshortened for directions close to the celestial equator. To obtain two-dimensional resolution in such directions requires components of antenna spacing parallel to the earth's axis. The design of such arrays is discussed in Chapter 5. The effect of the earth's rotation is then to distribute the measurements in (u, v, w) space so that they no longer lie in a plane, unless the observation is of short time duration. In many cases the restriction of the synthesized field in Eq. (3.12) is acceptable. However, at low frequencies (~100 MHz and lower) antennas have wide primary beams and it is often necessary to map the entire beam to avoid source confusion. In such circumstances several techniques are possible, based on the following approaches:

1. Equation (3.7) can be written in the form of a three-dimensional Fourier transform. The resulting intensity distribution is then taken from the surface of a unit sphere in (l, m, n) space.
2. Large maps can be constructed as mosaics of smaller ones that individually comply with the field restriction for two-dimensional transformation. The centers of the individual maps must be taken at tangent points on the same unit sphere referred to in 1.
3. Since in most terrestrial arrays the antennas are mounted on an approximately plane area of ground, measurements taken over a short time interval lie close to a plane in (u, v, w) space. It is therefore possible to analyze an observation lasting several hours as a series of short duration maps, which are subsequently combined after adjustment of the coordinate scales.

Practical implementation of the three approaches outlined above requires the nonlinear deconvolution techniques described in Chapter 11. A more detailed discussion of the resulting methods is given in Section 11.8.

3.2 CROSS-CORRELATION AND THE WIENER–KHINCHIN RELATION

The Fourier transform relationship between the power spectrum of a waveform and its autocorrelation function, expressed in Eqs. (2.6) and (2.7), is known as the Wiener–Khinchin relation. It is also useful to examine the corresponding relation for the cross-correlation function of two different waveforms. The response of a correlator, as used in a radio interferometer, can be written as

$$r(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} V_1(t) V_2^*(t-\tau)\, dt. \tag{3.19}$$

In practice the correlation is measured for a finite time period $2T$, which is usually a few seconds or minutes, but is long compared with both the period and the reciprocal bandwidth of the waveforms. The factor $1/2T$ is sometimes omitted, but for the waveforms considered here it is required to obtain convergence. Cross-correlation is represented by the pentagram symbol ($\star$):

$$V_1(t) \star V_2(t) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} V_1(t) V_2^*(t-\tau)\, dt. \tag{3.20}$$

This integral can be expressed as a convolution in the following way:

$$V_1(t) \star V_2(t) = \lim_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty} V_1(t) V_{2-}^*(\tau - t)\, dt = V_1(t) * V_{2-}^*(t), \tag{3.21}$$

where $V_{2-}(t) = V_2(-t)$. Now the ν, t Fourier transforms, denoted by $\rightleftharpoons$, are as follows[‡]: $V_1(t) \rightleftharpoons \widehat{V}_1(\nu)$, $V_2(t) \rightleftharpoons \widehat{V}_2(\nu)$, and $V_{2-}^*(t) \rightleftharpoons \widehat{V}_2^*(\nu)$. Then from the convolution theorem

$$V_1(t) \star V_2(t) \rightleftharpoons \widehat{V}_1(\nu) \widehat{V}_2^*(\nu). \tag{3.22}$$

The right-hand side of Eq. (3.22) is known as the cross power spectrum of $V_1(t)$ and $V_2(t)$. The cross power spectrum is a function of frequency, and we see that it is the Fourier transform of the cross-correlation, which is a function of τ. This is a useful result, and in the case where $V_1 = V_2$ it becomes the Wiener–Khinchin relation. The relationship expressed in Eq. (3.22) is the basis of cross-correlation spectrometry, described in Section 8.7 under *Principles of Digital Spectral Measurements*.

[‡]In cases where the same letter is used for functions of both time and frequency, the circumflex (hat) accent is used to distinguish functions of frequency.

3.3 BASIC RESPONSE OF THE RECEIVING SYSTEM

From a mathematical viewpoint, the basic components of the interferometer receiving system are the antennas that transform the incident electric fields into voltage waveforms, the filters that select the frequency components to be processed, and the correlator that forms the averaged product of the signals. These components are shown in Fig. 3.6. Most other effects can be represented by multiplicative gain constants, which we shall ignore here, or as variations of the frequency response that can be subsumed into the expressions for the filters. Thus we assume that the frequency response of the antennas and the strength of the received signal are effectively constant over the filter passband, which is realistic for most continuum observations.

Antennas

In order to consider the responses of the two antennas independently, we should introduce their voltage reception patterns, since the correlator responds to the product of the signal voltages. The voltage reception pattern of an antenna $V_A(l, m)$ has the dimension *length*, and responds to the electric field specified in volts per meter. $V_A(l, m)$ is the Fourier transform of the field distribution in the aperture $\overline{\mathcal{E}}(X, Y)$, as shown in Section 14.1 under *Diffraction at an Aperture and the Response of an Antenna*. X and Y are coordinates of position within the antenna aperture. Omitting constant factors, we can write

$$V_A(l, m) \propto \int\int_{-\infty}^{\infty} \overline{\mathcal{E}}(X, Y) e^{j2\pi[(X/\lambda)l + (Y/\lambda)m]}\, dX\, dY, \tag{3.23}$$

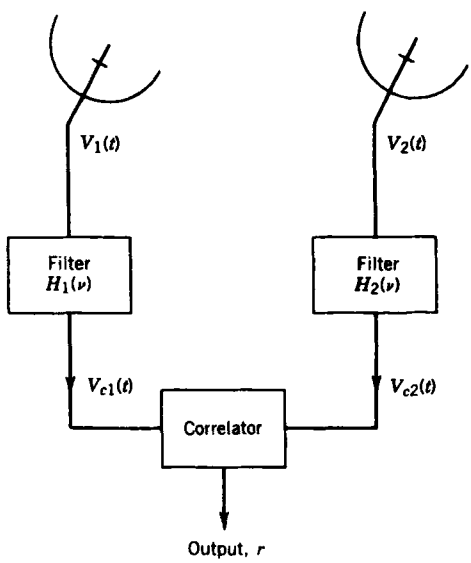


Figure 3.6 Basic components of the receiving system of a two-element interferometer.

where λ is the wavelength. In applying Eq. (3.23), X and Y are measured from the center of each antenna aperture. The power reception pattern is proportional to the squared modulus of the voltage reception pattern. $V_A(l, m)$ is a complex quantity, and it represents the phase of the radio frequency voltage at the antenna terminals as well as the amplitude. For an interferometer (with antennas denoted by subscripts 1 and 2) the response is proportional to $V_{A1} V_{A2}^*$, which is purely real if the antennas are identical. For each antenna the collecting area $A(l, m)$ is a real quantity. In practice, it is usual to specify the antenna response in terms of $A(l, m)$, and to replace $V_A(l, m)$ by $\sqrt{A(l, m)}$, which is proportional to the modulus of $V_A(l, m)$. Any phase introduced by differences between the antennas is ignored in the analysis, but in effect is combined with the phase responses of the amplifiers, filters, transmission lines, and other elements that make up the signal path to the correlator input. The overall instrumental response of the interferometer in both phase and amplitude is calibrated by observing an unresolved source of known position and flux density.

For the case where the antennas track the source, both the antenna beam center and the center of the source are at the (l, m) origin. If $E(l, m)$ is the incident field, the output voltage of an antenna can be written (omitting constant factors) as

$$\widehat{V} = \int \int_{-\infty}^{\infty} E(l, m) \sqrt{A(l, m)}\, dl\, dm. \tag{3.24}$$

If the antennas do not track the source, a convolution relationship of the form shown in Eq. (2.15) applies.

Filters

The filters in Fig. 3.6 will be regarded as a representation of the overall effect of components that determine the frequency response of the receiving channels, including amplifiers, cables, and other components as well as filters. The frequency response of a filter will be represented by $H(\nu)$. The output of the filter $\widehat{V}_c(\nu)$ is related to the input $\widehat{V}(\nu)$ by

$$\widehat{V}_c(\nu) = H(\nu) \widehat{V}(\nu). \tag{3.25}$$

The Fourier transform of $H(\nu)$ with respect to time and frequency is the impulse response of the filter $h(t)$, which is the response to a voltage impulse $\delta(t)$ at the input. Thus in the time domain the corresponding expression to Eq. (3.25) is

$$V_c(t) = \int_{-\infty}^{\infty} h(t') V(t - t')\, dt' = h(t) * V(t). \tag{3.26}$$

In specifying filters it is usual to use the frequency response rather than the impulse response because the former is more directly related to the properties of interest in a receiving system, and is usually easier to measure.

Correlator

The correlator[§] produces the cross-correlation of the two voltages fed to it. If $V_1(t)$ and $V_2(t)$ are the input voltages, the correlator output is

$$r(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} V_1(t) V_2^*(t-\tau)\, dt. \tag{3.27}$$

τ is the time by which voltage V_2 is delayed with respect to voltage V_1 and for continuum observations is maintained small or zero. The functions V_1 and V_2 that represent the signals in Eq. (3.27) may be complex. The output of a single multiplying device is a real voltage or number. To obtain the complex cross-correlation, which represents both the amplitude and the phase of the visibility, one can record the fringe oscillations and measure their phase, or use a *complex correlator* which contains two multiplying circuits, as described in Section 6.1 under *Simple and Complex Correlators*. As follows from Eqs. (3.20) and (3.22), the Fourier transform of $r(\tau)$ is the cross power spectrum, which is required in observations of spectral lines. This is usually obtained by inserting a series of instrumental delays in the signal to determine the cross-correlation as a function of τ, as described in Section 8.7 under *Lag (XF) Correlator*.

Response to the Incident Radiation

We use subscripts 1 and 2 to indicate the two antennas and receiving channels as in Fig. 3.6. The response of antenna 1 to the signal field $E(l, m)$ given by Eq. (3.24) is the voltage spectrum $\widehat{V}(\nu)$. We multiply this by $H(\nu)$ to obtain the signal at the output of the filter, and then take the Fourier transform to go from the frequency to the time domain. Thus

$$V_{c1}(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E(l, m)\sqrt{A_1(l, m)}H_1(\nu)e^{j2\pi\nu t}\, dl\, dm\, d\nu. \tag{3.28}$$

A similar expression can be written for the signal $V_{c2}(t)$ from antenna 2, and the output of the correlator is obtained from Eq. (3.27). Note also that if the radiation were to have some degree of spatial coherence, we should integrate over (l, m) independently for each antenna (Swenson and Mathur 1968), but here we make the usual assumption of incoherence. Thus the correlator output is

$$r(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E(l, m)E^*(l, m)\sqrt{A_1(l, m)A_2(l, m)}$$
$$\times\, H_1(\nu)H_2^*(\nu)e^{j2\pi\nu t}e^{-j2\pi\nu(t-\tau)}\, dl\, dm\, dt\, d\nu$$

[§]The term *correlator* basically refers to a device that measures the complex cross-correlation function $r(\tau)$ as given in Eq. (3.27). It is also used to denote simpler systems where the time delay τ is zero, or where both signals are represented by real functions. Large systems that cross-correlate the signal pairs of multielement arrays may contain 10^7 or more correlator circuits to accommodate many antennas and many spectral channels. Complete systems of this type are also commonly referred to as *correlators*.

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(l,m)\sqrt{A_1(l,m)A_2(l,m)}H_1(\nu)H_2^*(\nu)e^{j2\pi\nu\tau}\,dl\,dm\,d\nu. \tag{3.29}$$

Here we have replaced the squared field amplitude by the intensity I. The result is a very general one since the use of separate response functions A_1 and A_2 for the two antennas can accommodate different antenna designs, or different pointing offset errors, or both. Also different frequency responses H_1 and H_2 are used. In the case where the antennas and filters are identical Eq. (3.29) becomes

$$r(\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(l,m)A(l,m)|H(\nu)|^2 e^{j2\pi\nu\tau}\,dl\,dm\,d\nu. \tag{3.30}$$

The result is a function of the delay τ of the signal $V_{c2}(t)$ with respect to $V_{c1}(t)$. The geometric component of the delay is generally compensated by an adjustable instrumental delay (discussed in Chapters 6 and 7), so that $\tau = 0$ for radiation from the direction of the (l, m) origin. For spectral line observations the correlator system may incorporate additional delay elements so that the correlation is measured as a function of τ. For a wavefront incident from the direction (l, m), the difference in propagation times through the two antennas to the correlator results from a difference in path lengths of $(ul + vm)$ wavelengths, for the conditions indicated in Eqs. (3.8) and (3.9). The corresponding time difference is $(ul + vm)/\nu$. If we take as V_1 the signal from the antenna for which the path length is the greater (for positive l and m), then from Eq. (3.30), the correlator output becomes

$$r = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(l,m)A(l,m)|H(\nu)|^2 e^{-j2\pi(lu+mv)}\,dl\,dm\,d\nu. \tag{3.31}$$

Equation (3.31) indicates that the correlator output measures the Fourier transform of the intensity distribution modified by the antenna pattern. Let us assume that, as is often the case, the intensity and the antenna pattern are constant over the bandpass range of the filters, and the width of the source is small compared with the antenna beam. The correlator output then becomes

$$\begin{aligned} r &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(l,m)A(l,m)e^{-j2\pi(lu+mv)}dl\,dm \int_{-\infty}^{\infty}|H(\nu)|^2 d\nu \\ &= A_0\mathcal{V}(u,v)\int_{-\infty}^{\infty}|H(\nu)|^2\,d\nu, \end{aligned} \tag{3.32}$$

where A_0 is the collecting area of the antennas in the direction of the maximum beam response and $\mathcal{V}$ is the visibility. The filter response $H(\nu)$ is a dimensionless (gain) quantity. If the filter response is essentially constant over a bandwidth $\Delta\nu$, Eq. (3.32) becomes

$$r = A_0\mathcal{V}(u,v)\,\Delta\nu. \tag{3.33}$$

$\mathcal{V}(u, v)$ has units of W m^{-2} Hz^{-1}, A_0 has units of m^2, and $\Delta\nu$ has units of Hz. This is consistent with r, the output of the correlator, which is proportional to the correlated component of the received power.

APPENDIX 3.1 MATHEMATICAL REPRESENTATION OF NOISE-LIKE SIGNALS

Electromagnetic fields and voltage waveforms that result from the emissions of astronomical objects are generally characterized by variations of a random nature. The received waveforms are usually described as ergodic (time averages and ensemble averages converge to equal values), which implies strict stationarity. For a detailed discussion see, for example, Goodman (1985). Although such fields and voltages are entirely real, it is often convenient to represent them mathematically as complex functions. These complex functions can be manipulated in exponential form, and it is then necessary to take the real part as a final step in a calculation.

Analytic Signal

A formulation that is often used in optical and radio signal analysis to represent a function of time is known as the *analytic signal*, which was introduced by Gabor (1946); see, for example, Born and Wolf (1999), Bracewell (2000), or Goodman (1985). Let $V_R(t)$ represent a real function of which the Fourier (voltage) spectrum is

$$\widehat{V}(\nu) = \int_{-\infty}^{\infty} V_R(t) e^{-j2\pi\nu t}\, dt, \tag{A3.1}$$

where we use the circumflex accent to designate a function of frequency. The inverse transform is

$$V_R(t) = \int_{-\infty}^{\infty} \widehat{V}(\nu) e^{j2\pi\nu t}\, d\nu. \tag{A3.2}$$

To form the analytic signal, the imaginary part that is added to produce a complex function is the Hilbert transform [see, e.g., Bracewell (2000)] of $V_R(t)$. One way of forming the Hilbert transform is to multiply the Fourier spectrum of the original function by $j\,\mathrm{sgn}(\nu)$[¶]. In forming the Hilbert transform of a function the amplitudes of the Fourier spectral components are unchanged, but the phases are shifted by $\pi/2$, with the sign of the shift reversed for negative and positive frequencies. The Hilbert transform of $V_R(t)$, which becomes the imaginary part $V_I(t)$, is obtained as the inverse Fourier transform of the modified spectrum, as follows:

[¶]The function $\mathrm{sgn}(\nu)$ is equal to 1 for $\nu \geq 0$ and -1 for $\nu < 0$. The Fourier transform of $\mathrm{sgn}(\nu)$ is $-j/\pi t$.

$$V_I(t) = -j \int_{-\infty}^{\infty} \operatorname{sgn}(\nu)\widehat{V}(\nu)e^{j2\pi\nu t}d\nu$$
$$= j\int_{-\infty}^{0} \widehat{V}(\nu)e^{j2\pi\nu t}\,d\nu - j\int_{0}^{\infty} \widehat{V}(\nu)e^{j2\pi\nu t}\,d\nu. \qquad \text{(A3.3)}$$

The analytic signal is the complex function that represents $V_R(t)$, and is

$$V(t) = V_R(t) + jV_I(t)$$
$$= \int_{-\infty}^{0} (1+j^2)\widehat{V}(\nu)e^{j2\pi\nu t}\,d\nu + \int_{0}^{\infty} (1-j^2)\widehat{V}(\nu)e^{j2\pi\nu t}\,d\nu \qquad \text{(A3.4)}$$
$$= 2\int_{0}^{\infty} \widehat{V}(\nu)e^{j2\pi\nu t}\,d\nu.$$

It can be seen that the analytic signal contains no negative-frequency components. From Eq. (A3.4), another way of obtaining the analytic signal for a real function $V_R(t)$ is to suppress the negative-frequency components of the spectrum and double the amplitudes of the positive ones. It can also be shown [see, e.g., Born and Wolf (1999)] that

$$\langle [V_R(t)]^2\rangle = \langle [V_I(t)]^2\rangle = \tfrac{1}{2}\langle V(t)V^*(t)\rangle, \qquad \text{(A3.5)}$$

where angle brackets $\langle\ \rangle$ indicate the expectation. The analytic signal is so called because, considered as a function of a complex variable, it is analytic in the lower half of the complex plane.

From Eqs. (A3.2) and (A3.4), we obtain

$$\int_{-\infty}^{\infty} \widehat{V}(\nu)e^{j2\pi\nu t}\,dt = 2\,\mathcal{R}e\left[\int_{0}^{\infty} \widehat{V}(\nu)e^{j2\pi\nu t}\,dt\right]. \qquad \text{(A3.6)}$$

This is a useful equality that can be used with any *hermitian* function and its conjugate variable.

In most cases of interest in radio astronomy and optics the bandwidth of a signal is small compared with the mean frequency ν_0, which in many instrumental situations is the center frequency of a filter. Such a waveform resembles a sinusoid with amplitude and phase that vary with time on a scale that is slow compared with the period $1/\nu_0$. The analytic signal can then be written as

$$V(t) = \mathcal{C}(t)\,e^{j[2\pi\nu_0 t - \Phi(t)]}, \qquad \text{(A3.7)}$$

where $\mathcal{C}$ and Φ are real. The spectral components of the function under consideration are appreciable only for small values of $|\nu - \nu_0|$. Thus $\mathcal{C}(t)$ and $\Phi(t)$ consist of low-frequency components, and the period of the time variation of $\mathcal{C}$ and Φ is characteristically the reciprocal of the bandwidth. The real and imaginary parts

of the analytic signal can be written as

$$V_R(t) = \mathcal{C}(t)\cos[2\pi\nu_0 t - \Phi(t)] \tag{A3.8}$$

$$V_I(t) = \mathcal{C}(t)\sin[2\pi\nu_0 t - \Phi(t)] \tag{A3.9}$$

The modulus $\mathcal{C}(t)$ of the complex analytic signal can be regarded as a modulation envelope and $\Phi(t)$ represents the phase. In cases where the width of the signal band and the effect of the modulation are not important, it is clearly possible to consider $\mathcal{C}$ and Φ as constants, that is, to represent the signals as monochromatic waveforms of frequency ν_0, as in the introductory discussion. The case where the bandwidth is small compared with the center frequency, as represented by Eq. (A3.7), is referred to as the quasimonochromatic case.

As a simple example, $e^{j2\pi\nu t}$ is the analytic signal corresponding to the real function of time $\cos(2\pi\nu t)$. The Fourier spectrum of $e^{j2\pi\nu t}$ has a component at frequency ν only, but the Fourier spectrum of $\cos(2\pi\nu t)$ has components at the two frequencies $\pm\nu$. In general it is necessary to consider the negative-frequency components in the analysis of waveforms, unless they are represented by the analytic signal formulation, for which negative-frequency components are zero. For example, in Eq. (2.8) we included negative-frequency components. If we had omitted the negative frequencies and doubled the amplitude of the positive ones, the cosine term in Eq. (2.9) would have been replaced by $e^{j2\pi\nu_0\tau}$. We would then have taken the real part to arrive at the correct result. In the approach used in Chapter 2 it is necessary to include the negative frequencies since the autocorrelation function is real, and thus its Fourier transform is hermitian; that is, the real and imaginary parts have even and odd symmetry, respectively, along the frequency axis. In this book we have generally included the negative frequencies rather than using the analytic signal, and have made use of the relationship in Eq. (A3.6) when it was advantageous to do so.

It is interesting to note another property of functions of which the real and imaginary parts are a Hilbert transform pair. If the real and imaginary parts of a waveform (i.e., a function of time) are a Hilbert transform pair, then its spectral components are zero for negative frequencies. If we consider the inverse Fourier transforms, it is seen that if the waveform amplitude is zero for $t < 0$, the real and imaginary parts of the spectrum are a Hilbert transform pair. The response of any electrical system to an impulse function applied at time $t = 0$ is zero for $t < 0$, since an effect cannot precede its cause. A function representing such a response is referred to as a *causal function*, and the Hilbert transform relationship applies to its spectrum.

Truncated Function

Another consideration in the representation of waveforms concerns the existence of the Fourier transform. A condition of the existence of the transform is that the Fourier integral over the range $\pm\infty$ be finite. Although this is not always the case, it is possible to form a function for which the Fourier transform exists and that approaches the original function as the value of some parameter tends toward a

limit. For example, the original function can be multiplied by a Gaussian so that the product falls to zero at large values, and the Fourier integral exists. The Fourier transform of the product approaches that of the original function as the width of the Gaussian tends to infinity. Such transforms in the limit are applicable to periodic functions such as $\cos(2\pi\nu t)$, as shown by Bracewell (2000). In the case of noise-like waveforms the frequency spectrum of a time function can always be determined with satisfactory accuracy by analyzing a sufficiently long (but finite) time interval. In practice the time interval needs to be long compared with the physically significant timescales that are associated with the waveform, such as the reciprocals of the mean frequency and of the bandwidth. Thus if the function $V(t)$ is truncated at $\pm T$, the Fourier transform with respect to frequency becomes

$$\widehat{V}(\nu) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} V(t)e^{-j2\pi\nu t}\,dt. \tag{A3.10}$$

It is sometimes useful to define the truncated function as $V_T(t)$, where

$$\begin{aligned} V_T(t) &= V(t) \quad |t| \le T \\ V_T(t) &= 0 \qquad\;\; |t| > T, \end{aligned} \tag{A3.11}$$

and to write the Fourier transform as

$$\widehat{V}(\nu) = \lim_{T\to\infty} \frac{1}{2T}\int_{-\infty}^{\infty} V_T(t)e^{-j2\pi\nu t}\,dt. \tag{A3.12}$$

In the case of the analytic signal, truncation of the real part does not necessarily result in truncation of its Hilbert transform. It may therefore be necessary that the limits of the integral over time should be $\pm\infty$ as in Eq. (A3.12), rather than $\pm T$.

REFERENCES

Apostol, T. M., *Calculus, Vol. II*, Blaisdel, Waltham, MA, 1962, p. 82.

Born, M. and E. Wolf, *Principles of Optics*, 7th ed., Cambridge Univ. Press, Cambridge, UK, 1999.

Bracewell, R. N., Radio Interferometry of Discrete Sources, *Proc. IRE*, **46**, 97–105, 1958.

Bracewell, R. N., *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 2000 (earlier eds. 1965, 1978).

Brouw, W. N., *Data Processing for the Westbrook Synthesis Radio Telescope*, Univ. Leiden, 1971.

Gabor, D., Theory of Communication, *J. Inst. Elect. Eng.*, **93**, Part III, 429–457, 1946.

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

König, A., Astrometry with Astrographs, in *Astronomical Techniques, Stars and Stellar Systems*, Vol. 2, W. A. Hiltner, Ed., Univ. Chicago Press, Chicago, 1962, pp. 461–486.

Swenson, G. W., Jr. and N. C. Mathur, The Interferometer in Radio Astronomy, *Proc. IEEE*, **56**, 2114–2130, 1968.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

4 Geometric Relationships and Polarimetry

In this chapter we start to examine some of the practical aspects of interferometry. These include baselines, antenna mounts and beamshapes, and the response to polarized radiation, all of which involve geometric considerations and coordinate systems. The discussion is concentrated on earth-based arrays with tracking antennas, which illustrate the principles involved, although the same principles apply to other systems such as those that include one or more antennas in earth orbit.

4.1 ANTENNA SPACING COORDINATES AND (u, v) LOCI

Various coordinate systems are used to specify the relative positions of the antennas in an array, and of these one of the more convenient for terrestrial arrays is shown in Fig. 4.1. A right-handed Cartesian coordinate system is used where X and Y are measured in a plane parallel to the earth's equator, X in the meridian plane (defined as the plane through the poles of the earth and the reference point in the array), Y is measured toward the east, and Z toward the north pole. In terms of hour angle H and declination δ, the coordinates (X, Y, Z) are measured toward $(H = 0, \delta = 0)$, $(H = -6^{\mathrm{h}}, \delta = 0)$, and $(\delta = 90°)$, respectively. If $(X_\lambda, Y_\lambda, Z_\lambda)$ are the components of $\mathbf{D}_\lambda$ in the (X, Y, Z) system, the components (u, v, w) are given by

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} \sin H & \cos H & 0 \\ -\sin\delta\cos H & \sin\delta\sin H & \cos\delta \\ \cos\delta\cos H & -\cos\delta\sin H & \sin\delta \end{bmatrix} \begin{bmatrix} X_\lambda \\ Y_\lambda \\ Z_\lambda \end{bmatrix}. \tag{4.1}$$

Here (H, δ) are usually the hour angle and declination of the phase reference position. (In VLBI observations it is customary to set the X axis in the Greenwich meridian, in which case H is measured with respect to that meridian rather than a local one.) The elements of the transformation matrix given above are the direction cosines of the (u, v, w) axes with respect to the (X, Y, Z) axes and can easily be derived from the relationships in Fig. 4.2. Another method of specifying the baseline vector is in terms of its length, D, and the hour angle and declination, (h, d), of the intersection of the baseline direction with the northern celestial

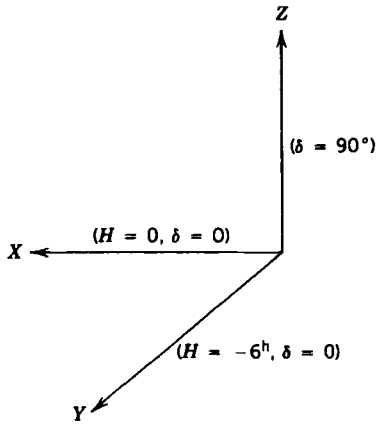


Figure 4.1 The (X, Y, Z) coordinate system for specification of relative positions of antennas. Directions of the axes specified are in terms of hour angle H and declination δ.

hemisphere. The coordinates in the (X, Y, Z) system are then given by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = D \begin{bmatrix} \cos d \cos h \\ -\cos d \sin h \\ \sin d \end{bmatrix}. \tag{4.2}$$

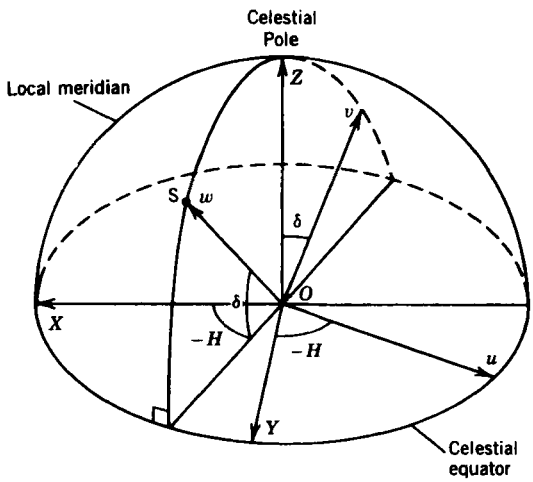


Figure 4.2 Relationships between the (X, Y, Z) and (u, v, w) coordinate systems. The (u, v, w) system is defined for observation in the direction of the point S, which has hour angle and declination H and δ. As shown, S is in the eastern half of the hemisphere and H is therefore negative. The direction cosines in the transformation matrix in Eq. (4.1) follow from the relationships in this diagram. The relationship in Eq. (4.2) can also be derived if we let S represent the direction of the baseline and put the baseline coordinates (h, d) for (H, δ).

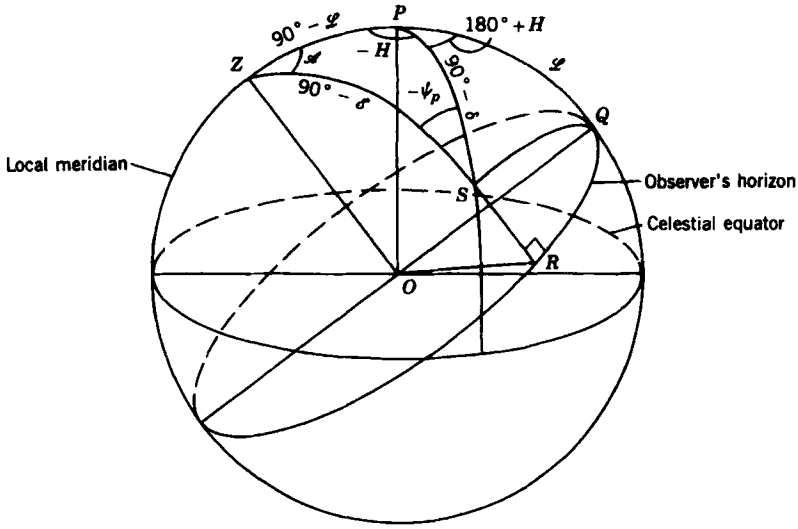


Figure 4.3 Relationship between the celestial coordinates (H, δ) and the elevation and azimuth $(\mathcal{E}, \mathcal{A})$ of a point S as seen by an observer at latitude $\mathcal{L}$. P is the celestial pole and Z the observer's zenith. The parallactic angle ψ_p is the position angle of the observer's vertical on the sky measured from north toward east. The lengths of the arcs measured in terms of angles subtended at the center of the sphere O are as follows:

$$ZP = 90° - \mathcal{L} \qquad PQ = \mathcal{L} \qquad SR = \mathcal{E} \quad RQ = \mathcal{A}$$

$$SZ = 90° - \mathcal{E} \qquad SP = 90° - \delta \qquad SQ = \cos^{-1}(\cos\mathcal{E}\cos\mathcal{A})$$

The required relationships can be obtained by application of the sine and cosine rules for spherical triangles to ZPS and PQS, and are given in Appendix 4.1. Note that with S in the eastern half of the observer's sky, as shown, H and ψ_p are negative.

The coordinates in the (u, v, w) system are, from Eqs. (4.1) and (4.2),

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = D_\lambda \begin{bmatrix} \cos d \sin(H - h) \\ \sin d \cos\delta - \cos d \sin\delta \cos(H - h) \\ \sin d \sin\delta + \cos d \cos\delta \cos(H - h) \end{bmatrix}. \tag{4.3}$$

The (D, h, d) system was used more widely in the earlier literature, particularly for instruments involving only two antennas; see, for example, Rowson (1963).

When the (X, Y, Z) components of a new baseline are first established, the usual practice is to determine the elevation $\mathcal{E}$, azimuth $\mathcal{A}$, and length of the baseline by field surveying techniques. Figure 4.3 shows the relationship between $(\mathcal{E}, \mathcal{A})$ and other coordinate systems; see also Appendix 4.1. For latitude $\mathcal{L}$, using Eqs. (4.2) and (A4.2), we obtain

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = D \begin{bmatrix} \cos \mathcal{L} \sin \mathcal{E} - \sin \mathcal{L} \cos \mathcal{E} \cos \mathcal{A} \\ \cos \mathcal{E} \sin \mathcal{A} \\ \sin \mathcal{L} \sin \mathcal{E} + \cos \mathcal{L} \cos \mathcal{E} \cos \mathcal{A} \end{bmatrix}. \tag{4.4}$$

Examination of Eq. (4.1) or (4.3) shows that the locus of the projected antenna spacing components u and v defines an ellipse with hour angle as the variable. Let (H_0, δ_0) be the phase reference position. Then from Eq. (4.1), we have

$$u^2 + \left(\frac{v - Z_\lambda \cos \delta_0}{\sin \delta_0} \right)^2 = X_\lambda^2 + Y_\lambda^2. \tag{4.5}$$

In the (u, v) plane Eq. (4.5) defines an ellipse with the semimajor axis equal to $\sqrt{X_\lambda^2 + Y_\lambda^2}$, and the semiminor axis equal to $\sin \delta_0 \sqrt{X_\lambda^2 + Y_\lambda^2}$, as in Fig. 4.4a. The ellipse is centered on the v axis at $(u, v) = (0, Z_\lambda \cos \delta_0)$. The arc of the ellipse that is traced out during any observation depends on the azimuth, elevation, and latitude of the baseline; the declination of the source; and the range of hour angle covered, as illustrated in Fig. 4.5. Since $\mathcal{V}(-u, -v) = \mathcal{V}^*(u, v)$, any observation supplies simultaneous measurements on two arcs, which are part of the same ellipse only if $Z_\lambda = 0$.

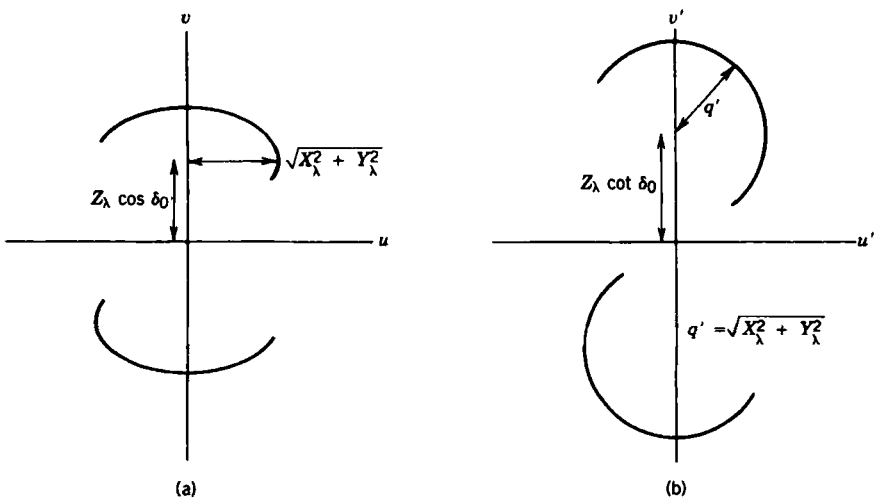


Figure 4.4 (a) Spacing vector locus in the (u, v) plane from Eq. (4.5). (b) Spacing vector locus in the (u', v') plane from Eq. (4.8). The lower arc in each diagram represents the locus of conjugate values of visibility. Unless the source is circumpolar the cutoff at the horizon limits the lengths of the arcs.

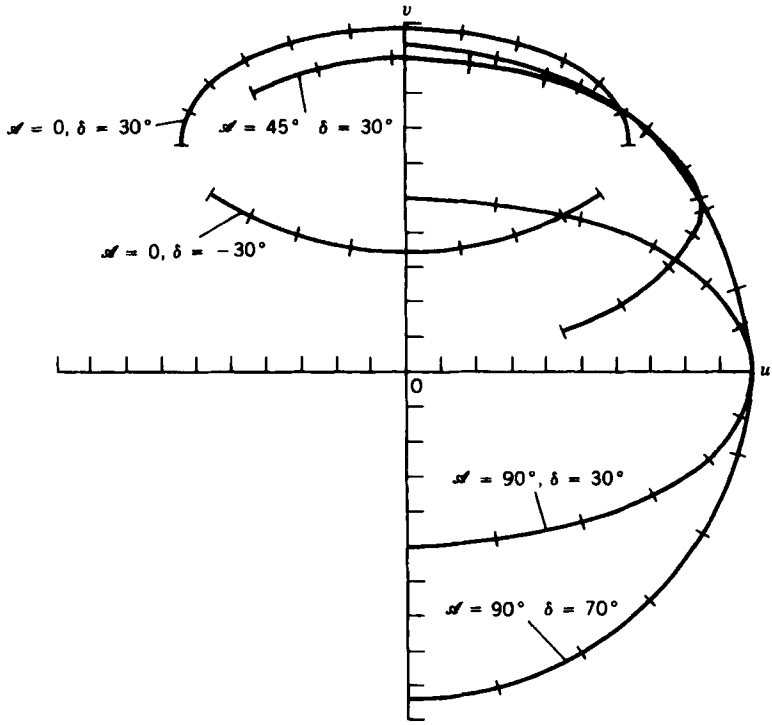


Figure 4.5 Examples of (u, v) loci to show the variation with baseline azimuth $\mathcal{A}$ and observing declination δ (the baseline elevation $\mathcal{E}$ is zero). The baseline length in all cases is equal to the length of the axes measured from the origin. The tracking range is -4 to $+4$ h for $\delta = -30°$, and -6 to $+6$ h in all other cases. Marks along the loci indicate 1-h intervals in tracking. Note the change in ellipticity for east–west baselines ($\mathcal{A} = 90°$) with $\delta = 30°$ and with $\delta = 70°$. The loci are calculated for latitude 40°.

4.2 (u', v') PLANE

The (u', v') plane, which was introduced in Section 3.1 with regard to east–west baselines, is also useful in discussing certain aspects of the behavior of arrays in general. This plane is normal to the direction of the pole and can be envisaged as the equatorial plane of the earth. For non-east–west baselines we can also consider the projection of the spacing vectors onto the (u', v') plane. All such projected vectors sweep out circular loci as the earth rotates. The spacing components in the (u', v') plane are derived from those in the (u, v) plane by the transformation $u' = u$, $v' = v \operatorname{cosec} \delta_0$. In terms of the components of the baseline $(X_\lambda, Y_\lambda, Z_\lambda)$ for two antennas, we obtain from Eq. (4.1)

$$u' = X_\lambda \sin H_0 + Y_\lambda \cos H_0 \tag{4.6}$$

$$v' = -X_\lambda \cos H_0 + Y_\lambda \sin H_0 + Z_\lambda \cot \delta_0. \tag{4.7}$$

The loci are circles centered on $(0, Z_\lambda \cot \delta_0)$, with radii q' given by

$$q'^2 = u'^2 + (v' - Z_\lambda \cot \delta_0)^2 = X_\lambda^2 + Y_\lambda^2, \tag{4.8}$$

as shown in Fig. 4.4b. The projected spacing vectors that generate the loci rotate with constant angular velocity ω_e, the rotation velocity of the earth, which is easier to visualize than the elliptic motion in the (u, v) plane. In particular, problems involving the effect of time, such as the averaging of visibility data, are conveniently dealt with in the (u', v') plane. Examples of its use will be found in Sections 4.4, 6.4, and 15.2. In Fourier transformation the conjugate variables of (u', v') are (l', m'), where $l' = l$ and $m' = m \sin \delta_0$, that is, the map plane is compressed by a factor $\sin \delta_0$ in the m direction.

4.3 FRINGE FREQUENCY

The component w of the baseline represents the path difference to the two antennas for a plane wave incident from the phase reference position. The corresponding time delay is w/ν_0, where ν_0 is the center frequency of the observing band. The relative phase of the signals at the two antennas changes by 2π radians when w changes by unity. Thus the frequency of the oscillations at the output of the correlator that combines the signals is

$$\frac{dw}{dt} = \frac{dw}{dH}\frac{dH}{dt} = -\omega_e[X_\lambda \cos\delta \sin H + Y_\lambda \cos\delta \cos H] = -\omega_e u \cos\delta, \tag{4.9}$$

where $\omega_e = dH/dt = 7.29115 \times 10^{-5}$ rad s^{-1} is the rotation velocity of the earth: for greater accuracy, see Seidelmann (1992). The sign of dw/dt indicates whether the phase is increasing or decreasing with time. The result shown above applies to the case where the signals suffer no time-varying instrumental phase changes between the antennas and the correlator inputs. In an array in which the antennas track a source, time delays to compensate for the space path differences w are usually applied under computer control to maintain correlation of the signals. If an exact compensating delay were introduced in the radio frequency section of the receivers, the relative phases of the signals at the correlator input would remain constant, and the correlator output would show no fringes. However, the compensating delays are usually introduced at an intermediate frequency, of which the band center ν_d is usually much less than the radio frequency ν_0. The adjustment of the compensating delay introduces a rate of phase change $2\pi\nu_d(dw/dt)/\nu_0 = -\omega_e u(\cos\delta)\nu_d/\nu_0$. The resulting fringe frequency at the correlator output is

$$\nu_f = \frac{dw}{dt}\left(1 \mp \frac{\nu_d}{\nu_0}\right) = -\omega_e u \cos\delta \left(1 \mp \frac{\nu_d}{\nu_0}\right), \tag{4.10}$$

where the negative sign refers to upper-sideband reception and the positive sign to lower-sideband reception; these distinctions and the double-sideband case are explained in Section 6.1. From Eq. (4.3) the right-hand side of Eq. (4.10) is equal to $-\omega_e D \cos d \cos\delta \sin(H-h)(\nu_0 \mp \nu_d)/c$. Note that $(\nu_0 \mp \nu_d)$ is usually determined by one or more local oscillator frequencies.

4.4 VISIBILITY FREQUENCIES

As explained in Section 3.1, the phase of the complex visibility is measured with respect to that of a hypothetical point source at the phase reference position. The fringe-frequency variations do not appear in the visibility function, but slower variations occur that depend on the position of the radiating sources within the field. We now examine the maximum temporal frequency of the visibility variations. Consider a point source represented by the delta function $\delta(l_1, m_1)$. The visibility function is the Fourier transform of $\delta(l_1, m_1)$, which is

$$e^{-j2\pi(ul_1+vm_1)} = \cos 2\pi(ul_1 + vm_1) - j \sin 2\pi(ul_1 + vm_1). \tag{4.11}$$

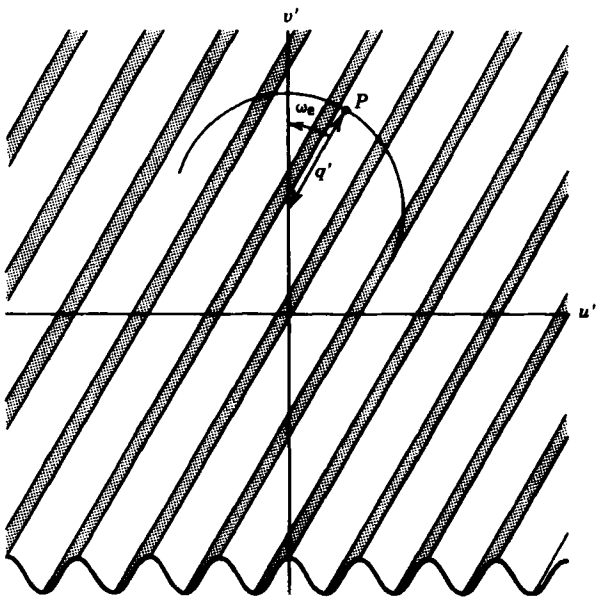


Figure 4.6 The (u', v') plane showing sinusoidal corrugations that represent the visibility of a point source. For simplicity only the real part of the visibility is included. The most rapid variation in the visibility is encountered at the point P where the direction of the spacing locus is normal to the ridges in the visibility. ω_e is the rotation velocity of the earth.

This expression represents two sets of sinusoidal corrugations, one real and one imaginary. The corrugations represented by the real part of Eq. (4.11) are shown in (u', v') coordinates in Fig. 4.6, where the arguments of the trigonometric functions in Eq. (4.11) become $2\pi(u'l_1 + v'm_1 \sin \delta_0)$. The frequency of the corrugations in terms of cycles per unit distance in the (u', v') plane is l_1 in the u' direction, $m_1 \sin \delta_0$ in the v' direction, and

$$r_1' = \sqrt{l_1^2 + m_1^2 \sin^2 \delta_0} \tag{4.12}$$

in the direction of most rapid variations. Expression (4.12) is maximized at the pole and then becomes equal to r_1, which is the angular distance of the source from the (l, m) origin. For any antenna pair the spatial frequency locus in the (u', v') plane is a circle of radius q' generated by a vector rotating with angular velocity ω_e, where q' is as defined in Eq. (4.8). From Fig. 4.6 it is clear that the temporal variation of the measured visibility is greatest at the point P and is equal to $\omega_e r_1' q'$. This is a useful result, since if r_1 represents a position at the edge of the field to be mapped, it indicates that to follow the most rapid variations the visibility must be sampled at time intervals sufficiently small compared with $(\omega_e r_1' q')^{-1}$. Also, we may wish to alternate between two frequencies or polarizations during an observation, and these changes must be made on a similarly short timescale. Note that this requirement is also covered by the sampling theorem in Section 5.2.

4.5 CALIBRATION OF THE BASELINE

The position parameters (X, Y, Z) for each antenna relative to a common reference point can usually be established to a few centimeters or millimeters by a conventional engineering survey. Except at long wavelengths, the accuracy required is greater than this. We must be able to compute the phase at any hour angle for a point source at the phase reference position to an accuracy of, say, 1° and subtract it from the observed phase. This reference phase is represented by the factor $e^{j2\pi w}$ in Eq. (3.7), and it is therefore necessary to calculate w to 1/360 of the observing wavelength. The baseline parameters can be obtained to the required accuracy from observations of calibration sources for which the positions are accurately known. The phase of such a calibrator observed at the phase reference position (H_0, δ_0) should ideally be zero. However, if practical uncertainties are taken into account, the measured phase is, from Eq. (4.1),

$$2\pi \Delta w + \phi_{\text{in}} = 2\pi(\cos \delta_0 \cos H_0 \Delta X_\lambda - \cos \delta_0 \sin H_0 \Delta Y_\lambda + \sin \delta_0 \Delta Z_\lambda) + \phi_{\text{in}}, \tag{4.13}$$

where the prefix Δ indicates the uncertainty in the associated quantity, and ϕ_{in} is an instrumental phase term for the two antennas involved. If a calibrator is observed over a wide range of hour angle, ΔX_λ and ΔY_λ can be obtained from

the even and odd components, respectively, of the phase variation with H_0. To measure ΔZ_λ calibrators at more than one declination must be included. A possible procedure is to observe several calibrators at different declinations, repeating a cycle of observations for several hours. For the kth observation we can write, from Eq. (4.13),

$$a_k \Delta X_\lambda + b_k \Delta Y_\lambda + c_k \Delta Z_\lambda + \phi_{\text{in}} = \phi_k, \tag{4.14}$$

where a_k, b_k, and c_k are known source parameters and ϕ_k is the measured phase. The calibrator source position need not be accurately known since the phase measurements can be used to estimate both the source positions and the baselines. Techniques for this analysis are discussed in Section 12.2. In practice, the instrumental phase ϕ_{in} will vary slowly with time: instrumental stability is discussed in Chapter 7. Also there will be atmospheric phase variations, which are discussed in Chapter 13. These effects set the final limit on the attainable accuracy in observing both calibrators and sources under investigation.

Measurement of baseline parameters to an accuracy of order 1 part in 10^7 (e.g., 3 mm in 30 km) implies timing accuracy of order $10^{-7}\omega_e^{-1} \simeq 1$ ms. Timekeeping is discussed in Sections 9.5 under *Time Synchronization* and 12.3 under *Universal Time*.

4.6 ANTENNA MOUNTS

In discussing the dependence of the measured phase on the baseline components, we have ignored any effects introduced by the antennas, which is tantamount to assuming that the antennas are identical and their effects on the signals cancel out. This, however, is only approximately true. In most synthesis arrays the antennas must have collecting areas of tens or hundreds of square meters for reasons of sensitivity. These large structures must be capable of accurately tracking a radio source across the sky. Tracking antennas are almost always constructed either on equatorial mounts (also called polar mounts) or on altazimuth mounts, as illustrated in Fig. 4.7. In an equatorial mount the polar axis is parallel to the earth's axis of rotation, and to track a source requires only that the antenna be turned about the polar axis at the sidereal rate. Equatorial mounts are mechanically more difficult to construct than altazimuth ones and are found mainly on antennas built prior to the introduction of computers for control and coordinate conversion.

In most tracking arrays used in radio astronomy the antennas are circularly symmetrical reflectors. A desirable feature is that the axis of symmetry of the reflecting surface intersect both the rotation axes of the mount. If this is not the case, pointing motions will cause the antenna to have a component of motion along the direction of the beam. It is then necessary to take account of phase changes associated with small pointing corrections, which may differ from one antenna to another. In most antenna mounts, however, whether of equatorial or altazimuth type, the reflector axis intersects the rotation axes with sufficient precision that phase errors of this type are negligible.

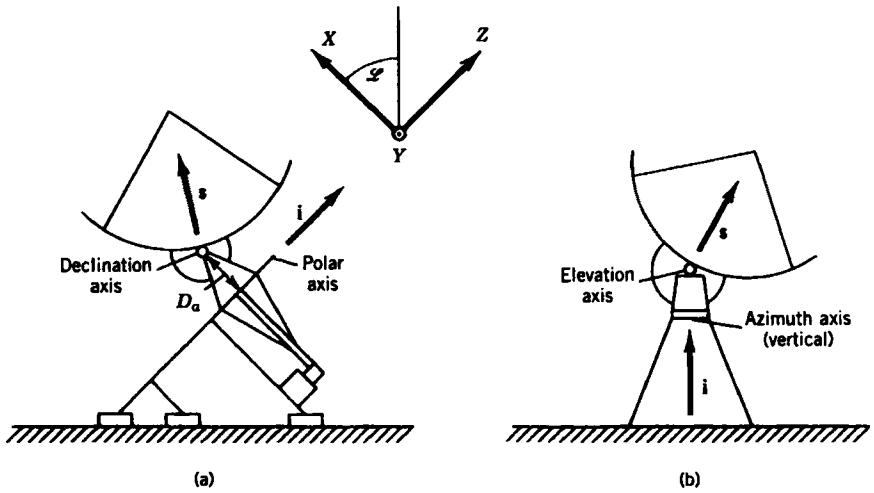


Figure 4.7 Schematic diagrams of antennas on (a) equatorial (polar) and (b) altazimuth mounts. In the positions shown the declination and elevation axes are normal to the plane of the page. In the equatorial mount there is a distance D_a between the two rotational axes, but in the altazimuth mount the axes often intersect, as shown.

It is convenient but not essential that the two rotation axes of the mount intersect. The intersection point then provides an appropriate reference point for defining the baseline between antennas, since whatever direction in which the antenna points, its aperture plane is always the same distance from that point as measured along the axis of the beam. In most large, equatorially mounted antennas the polar and declination axes do not intersect. In many cases there is an offset of several meters between the polar and declination axes. Wade (1970) has considered the implication of this offset for high-accuracy phase measurements and shown that it is necessary to take account of variations in the offset distance and in the accuracy of alignment of the polar axis. These results can be obtained as follows. Let $\mathbf{i}$ and $\mathbf{s}$ be unit vectors in the direction of the polar axis and the direction of the source under observation, respectively, and let $\mathbf{D}_a$ be the spacing vector between the two axes measured perpendicular to $\mathbf{i}$ (see Fig. 4.7a). The quantity that we need to compute is the projection of $\mathbf{D}_a$ in the direction of observation, $\mathbf{D}_a \cdot \mathbf{s}$. Since $\mathbf{D}_a$ is perpendicular to $\mathbf{i}$, the cosine of the angle between $\mathbf{D}_a$ and $\mathbf{s}$ is $\sqrt{1 - (\mathbf{i} \cdot \mathbf{s})^2}$. Thus

$$\mathbf{D}_a \cdot \mathbf{s} = D_a \sqrt{1 - (\mathbf{i} \cdot \mathbf{s})^2}, \tag{4.15}$$

where D_a is the magnitude of $\mathbf{D}_a$. In the (X, Y, Z) coordinate system in which the baseline components are measured, $\mathbf{i}$ has direction cosines (i_X, i_Y, i_Z) and $\mathbf{s}$ has direction cosines given by the transformation matrix on the right-hand side of Eq. (4.2), but with h and d replaced by H and δ, which refer to the direction of observation. If the polar axis is correctly aligned to within about 1 arcmin, i_X and

i_Y are of order 10^{-3} and $i_Z \simeq 1$. Thus we can use the direction cosines to evaluate Eq. (4.15), and ignoring second-order terms in i_X and i_Y we obtain

$$\mathbf{D}_a \cdot \mathbf{s} = D_a(\cos\delta - i_X \sin\delta \cos H + i_Y \sin\delta \sin H). \tag{4.16}$$

If the magnitude of $\mathbf{D}_a$ is expressed in wavelengths, the difference in the values of $\mathbf{D}_a \cdot \mathbf{s}$ for the two antennas must be added to the w component of the baseline given by Eq. (4.1) when calculating the reference phase at the field center. To do this it is first necessary to determine the unknown constants in Eq. (4.16), which can be done by adding a term of the form $2\pi(\alpha \cos \delta_0 + \beta \sin \delta_0 \cos H_0 + \gamma \sin \delta_0 \sin H_0)$ to the right-hand side of Eq. (4.13) and extending the solution to include α, β, and γ. The result then represents the differences in the corresponding mechanical dimensions of the two antennas. Note that the terms in i_X and i_Y in Eq. (4.16) are important only when D_a is large. If D_a is no more than one wavelength, it should be possible to ignore them.

The preceding analysis can be extended to the case of an altazimuth mount by letting $\mathbf{i}$ represent the direction of the azimuth axis as in Fig. 4.7b. Then $i_X = \cos(\mathcal{L} + \varepsilon)$, $i_Y = \sin\varepsilon'$, and $i_Z = \sin(\mathcal{L} + \varepsilon)$, where $\mathcal{L}$ is the latitude and ε and ε' are, respectively, the tilt errors in the XZ plane and in the plane containing the Y axis and the local vertical. The errors again should be quantities of order 10^{-3}. In many altazimuth mounts the axes are designed to intersect, and D_a represents only a structural tolerance. Thus we assume that D_a is small enough to allow terms in $i_Y D_a$ and εD_a to be ignored, and evaluation of Eq. (4.15) gives

$$\mathbf{D}_a \cdot \mathbf{s} = D_a\left[1 - (\sin\mathcal{L}\sin\delta + \cos\mathcal{L}\cos\delta\cos H)^2\right] = D_a \cos\mathcal{E}, \tag{4.17}$$

where $\mathcal{E}$ is the elevation of direction $\mathbf{s}$: see Eq. (A4.1) of Appendix 4.1. Correction terms of this form can be added to the expressions for the baseline calibration and for w.

4.7 BEAMWIDTH AND BEAM-SHAPE EFFECTS

The interpretation of data taken with arrays containing antennas with nonidentical beamwidths is not always a straightforward matter. Each antenna pair responds to an effective intensity distribution that is the product of the actual intensity of the sky and the geometric mean of the normalized beam profiles. If different pairs of antennas respond to different effective distributions, then, in principle, the Fourier transform relationship between $I(l, m)$ and $\mathcal{V}(u, v)$ cannot be applied to the ensemble of observations. Mixed arrays are frequently used in VLBI when it is necessary to make use of antennas that have different designs. However, in VLBI studies the source structure under investigation is very small compared with the widths of the antenna beams, so the differences in the beams can usually be ignored. If cases arise where different beams are used and the source is not small compared with beamwidths, it is possible to restrict the measurements to

the field defined by the narrowest beam by convolution of the visibility data with an appropriate function in the (u, v) plane.

A problem similar to that of unmatched beams occurs if the antennas have altazimuth mounts and the beam contours are not circularly symmetrical about the nominal beam axis. As a point in the sky is tracked using an altazimuth mount, the beam rotates with respect to the sky about this nominal axis. This rotation does not occur for equatorial mounts. The angle between the vertical at the antenna and the direction of north at the point being observed (defined by the great circle through the point and the north pole) is the parallactic angle ψ_p in Fig. 4.3. Application of the sine rule to the spherical triangle ZPS gives

$$\frac{-\sin\psi_p}{\cos\mathcal{L}} = \frac{-\sin H}{\cos\mathcal{E}} = \frac{\sin\mathcal{A}}{\cos\delta}, \tag{4.18}$$

which can be combined with Eq. (A4.1) or (A4.2) to express ψ_p as a function of $(\mathcal{A}, \mathcal{E})$ or (H, δ). If the beam has elongated contours and width comparable to the source under observation, rotation of the beam causes the effective intensity distribution to vary with hour angle. However, in most tracking arrays the antenna beams are sufficiently circularly symmetrical that this problem is rarely serious.

4.8 POLARIMETRY

Parameters Defining Polarization

Polarization measurements are very important in radio astronomy. For example, most synchrotron radiation shows a small degree of polarization which indicates the distribution of the magnetic fields within the source. As noted in Chapter 1, this polarization is generally linear (plane) and can vary in magnitude and position angle over the source. As frequency is increased, the percentage polarization often increases because the depolarizing action of Faraday rotation is reduced. Polarization of radio emission also results from the Zeeman effect in atoms and molecules, cyclotron radiation and plasma oscillations in the solar atmosphere, and Brewster angle effects at planetary surfaces. The measure of polarization that is almost universally used in astronomy is the set of four parameters introduced by Sir George Stokes in 1852. We assume here that readers have some familiarity with the concept of Stokes parameters or can refer to one of numerous texts that describe them [e.g., Born and Wolf (1999), Kraus and Carver (1973), Rohlfs and Wilson (1996)].

Stokes parameters are related to the amplitudes of the components of the electric field, E_x and E_y, resolved in two perpendicular directions normal to the direction of propagation. Thus if E_x and E_y are represented by $\mathcal{E}_x(t)\cos[2\pi\nu t + \delta_x(t)]$ and $\mathcal{E}_y(t)\cos[2\pi\nu t + \delta_y(t)]$, respectively, Stokes parameters are defined as follows:

$$I = \langle\mathcal{E}_x^2(t)\rangle + \langle\mathcal{E}_y^2(t)\rangle$$

$$Q = \langle\mathcal{E}_x^2(t)\rangle - \langle\mathcal{E}_y^2(t)\rangle$$

$$U = 2\langle \mathcal{E}_x(t)\mathcal{E}_y(t)\cos[\delta_x(t) - \delta_y(t)]\rangle$$
$$V = 2\langle \mathcal{E}_x(t)\mathcal{E}_y(t)\sin[\delta_x(t) - \delta_y(t)]\rangle, \tag{4.19}$$

where the angular brackets denote the expectation or time average. This averaging is necessary because in radio astronomy we are dealing with fields that vary with time in random manner. Of the four parameters, I is a measure of the total intensity of the wave, Q and U represent the linearly polarized component, and V represents the circularly polarized component. Stokes parameters can be converted to a measure of polarization with a more direct physical interpretation as follows:

$$m_\ell = \frac{\sqrt{Q^2 + U^2}}{I} \tag{4.20}$$

$$m_c = \frac{V}{I} \tag{4.21}$$

$$m_t = \frac{\sqrt{Q^2 + U^2 + V^2}}{I} \tag{4.22}$$

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{U}{Q}\right), \qquad 0 \le \theta \le \pi, \tag{4.23}$$

where m_ℓ, m_c, and m_t are the degrees of linear, circular, and total polarization, respectively, and θ is the position angle of the plane of linear polarization. For monochromatic signals, $m_t = 1$ and the polarization can be fully specified by just three parameters. For random signals such as those of cosmic origin, $m_t \le 1$ and all four parameters are required. The Stokes parameters all have the dimensions of flux density or intensity, and they propagate in the same manner as the electromagnetic field. Thus they can be determined by measurement or calculation at any point along a wave path, and their relative magnitudes define the state of polarization at that point. Stokes parameters combine additively for independent waves. When they are used to specify the total radiation from any point on a source, I, which measures the total intensity, is always positive, but Q, U, and V can take both positive and negative values depending on the position angle or sense of rotation of the polarization. The corresponding visibility values measured with an interferometer are complex quantities, as will be discussed later.

In considering the response of interferometers and arrays up to this point we have ignored the question of polarization. This simplification can be justified by the assumption that we have been dealing with completely unpolarized radiation for which only the parameter I is nonzero. In that case the response of an interferometer with identically polarized antennas is proportional to the total flux density of the radiation. As will be shown below, in the more general case the response is proportional to a linear combination of two or more Stokes parameters, where the combination is determined by the polarizations of the antennas. By observing with different states of polarization of the antennas it is possible to separate the responses to the four parameters and determine the corresponding compo-

nents of the visibility. The variation of each parameter over the source can thus be mapped individually, and the polarization of the radiation emitted at any point can be determined. There are alternative methods of describing the polarization state of a wave, of which the coherency matrix is perhaps the most important (Ko 1967a,b). However, the classical treatment in terms of Stokes parameters has remained essentially universal in its usage by astronomers, and we therefore follow it here.

Antenna Polarization Ellipse

The polarization of an antenna in either transmission or reception can be described in general by stating that the electric vector of a transmitted signal traces out an elliptical locus in the wavefront plane. Most antennas are designed so that the ellipse approximates a line or circle, corresponding to linear or circular polarization, in the central part of the main beam. However, exact linear or circular responses are never achieved in practice. As shown in Fig. 4.8a, the essential characteristics of the polarization ellipse are given by the position angle ψ of the

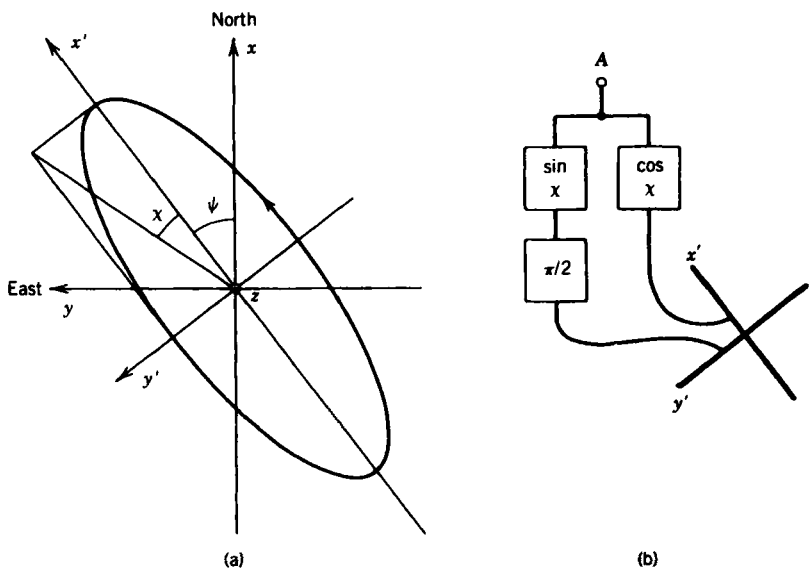


Figure 4.8 (a) Description of the general state of polarization of an antenna in terms of the characteristics of the ellipse generated by the electric vector in the transmission of a sinusoidal signal. The position angle ψ of the major axis is measured with respect to the x axis, which points toward the direction of north on the sky. A wave approaching from the sky is traveling toward the reader, in the direction of the positive z axis. For such a wave the arrow on the ellipse indicates the direction of right-handed polarization. (b) Model antenna that radiates the electric field represented by the ellipse in (a) when a signal is applied to the terminal A. Cos χ and sin χ indicate the voltage responses of the units shown, and $\pi/2$ indicates a phase lag.

major axis, and by the axial ratio, which it is convenient to express as the tangent of an angle χ, where $-\pi/4 \leq \chi \leq \pi/4$.

An antenna of arbitrary polarization can be modeled in terms of two idealized dipoles as shown in Fig. 4.8b. Consider *transmitting* with this antenna by applying a signal waveform to the terminal A. The signals to the dipoles pass through networks with voltage responses proportional to $\cos\chi$ and $\sin\chi$, and the signal to the y' dipole also passes through a network that introduces a $\pi/2$ phase lag. Thus the antenna produces field components of amplitude $\mathcal{E}_{x'}$ and $\mathcal{E}_{y'}$ in phase quadrature along the directions of the major and minor axes of the ellipse. If the antenna input is a radio frequency sine wave $V_0 \cos 2\pi\nu t$, then the field components are

$$\begin{aligned} \mathcal{E}_{x'} \cos 2\pi\nu t &\propto V_0 \cos\chi \cos(2\pi\nu t) \\ \mathcal{E}_{y'} \sin 2\pi\nu t &\propto V_0 \sin\chi \sin(2\pi\nu t). \end{aligned} \tag{4.24}$$

In these equations the y' component lags the x' component by $\pi/2$. If $\chi = \pi/4$, the radiated electric vector traces a circular locus with the sense of rotation from the x' axis to the y' axis (i.e., *counterclockwise* in Fig. 4.8a). This is consistent with the quarter-cycle delay in the signal to the y' dipole. Then a wave propagating in the positive z' direction of a right-handed coordinate system (i.e., toward the reader in Fig. 4.8a) is right circularly polarized in the IEEE (1977) definition. (This definition is now widely adopted, but in some of the older literature such a wave would be defined as left circularly polarized.) The International Astronomical Union (IAU 1973) has adopted the IEEE definition and states that the position angle of the electric vector on the sky should be measured from north through east with reference to the system of right ascension and declination. The IAU also states that "the polarization of incoming radiation, for which the position angle, θ, of the electric vector, measured at a fixed point in space, increases with time, is described as right-handed and positive." Note that Stokes parameters in Eqs. (4.19) specify only the field in the (x, y) plane, and to determine whether a circularly polarized wave is left- or right-handed, the direction of propagation must be given. From Eqs. (4.19), and the definitions of E_x and E_y that precede them, a wave traveling in the positive z direction in right-handed coordinates is right circularly polarized for positive V.

In reception an electric vector that rotates in a *clockwise* direction in Fig. 4.8 produces a voltage in the y' dipole that leads the voltage in the x' dipole by $\pi/2$ in phase, and the two signals therefore combine in phase at A. For counterclockwise rotation the signals at A are in antiphase and cancel one another. Thus the antenna in Fig. 4.8 receives right-handed waves incident from the positive z direction (that is, traveling toward negative z), and it transmits right-handed polarization in the direction toward positive z. To receive a right-handed wave propagating down from the sky (in the positive z direction), the polarity of one of the dipoles must be reversed, which requires that $\chi = -\pi/4$.

To determine the interferometer response, we begin by considering the output of the antenna modeled in Fig. 4.8b. We define the field components in complex form:

$$E_x(t) = \mathcal{E}_x(t) e^{j[2\pi\nu t + \delta_x(t)]}, \qquad E_y(t) = \mathcal{E}_y(t) e^{j[2\pi\nu t + \delta_y(t)]}. \tag{4.25}$$

The signal voltage received at A in Fig. 4.8b, expressed in complex form, is

$$V' = E_{x'} \cos\chi - jE_{y'} \sin\chi, \tag{4.26}$$

where the factor $-j$ represents the $\pi/2$ phase lag applied to the y' signal, for the fields represented by Eqs. (4.25). Now we need to specify the polarization of the incident wave in terms of Stokes parameters. In accordance with IAU (1973) the axes used are in the directions of north and east on the sky, which are represented by x and y in Fig. 4.8a. In terms of the field in the x and y directions the components of the field in the x' and y' directions are

$$\begin{aligned} E_{x'}(t) &= \left[\mathcal{E}_x(t)e^{j\delta_x(t)} \cos\psi + \mathcal{E}_y(t)e^{j\delta_y(t)} \sin\psi\right] e^{j2\pi\nu t} \\ E_{y'}(t) &= \left[-\mathcal{E}_x(t)e^{j\delta_x(t)} \sin\psi + \mathcal{E}_y(t)e^{j\delta_y(t)} \cos\psi\right] e^{j2\pi\nu t}. \end{aligned} \tag{4.27}$$

Derivation of the response at the output of the correlator for antennas m and n of an array involves straightforward manipulation of some rather lengthy expressions that are not reproduced here. The steps are as follows:

1. Substitute $E_{x'}$ and $E_{y'}$ from Eqs. (4.27) into Eq. (4.26) to obtain the output of each antenna.
2. Indicate values of ψ, χ, and V' for the two antennas by subscripts m and n and calculate the correlator output, $R_{mn} = G_{mn}\langle V'_m V'^*_n\rangle$, where G_{mn} is an instrumental gain factor.
3. Substitute Stokes parameters for $\mathcal{E}_x$, $\mathcal{E}_y$, δ_x, δ_y using Eqs. (4.19) as follows:

$$\begin{aligned} \langle(\mathcal{E}_x e^{j\delta_x})(\mathcal{E}_x e^{j\delta_x})^*\rangle &= \langle\mathcal{E}_x^2\rangle = \tfrac{1}{2}(I + Q) \\ \langle(\mathcal{E}_y e^{j\delta_y})(\mathcal{E}_y e^{j\delta_y})^*\rangle &= \langle\mathcal{E}_y^2\rangle = \tfrac{1}{2}(I - Q) \\ \langle(\mathcal{E}_x e^{j\delta_x})(\mathcal{E}_y e^{j\delta_y})^*\rangle &= \langle\mathcal{E}_x\mathcal{E}_y e^{j(\delta_x-\delta_y)}\rangle = \tfrac{1}{2}(U + jV) \\ \langle(\mathcal{E}_x e^{j\delta_x})^*(\mathcal{E}_y e^{j\delta_y})\rangle &= \langle\mathcal{E}_x\mathcal{E}_y e^{-j(\delta_x-\delta_y)}\rangle = \tfrac{1}{2}(U - jV). \end{aligned} \tag{4.28}$$

The result is

$$\begin{aligned} R_{mn} = \tfrac{1}{2}G_{mn}\{ & I_v\,[\cos(\psi_m - \psi_n)\cos(\chi_m - \chi_n) + j\sin(\psi_m - \psi_n)\sin(\chi_m + \chi_n)] \\ & + Q_v\,[\cos(\psi_m + \psi_n)\cos(\chi_m + \chi_n) + j\sin(\psi_m + \psi_n)\sin(\chi_m - \chi_n)] \\ & + U_v\,[\sin(\psi_m + \psi_n)\cos(\chi_m + \chi_n) - j\cos(\psi_m + \psi_n)\sin(\chi_m - \chi_n)] \\ & - V_v\,[\cos(\psi_m - \psi_n)\sin(\chi_m + \chi_n) + j\sin(\psi_m - \psi_n)\cos(\chi_m - \chi_n)]\}. \end{aligned} \tag{4.29}$$

In this equation a subscript v has been added to Stokes parameter symbols to indicate that they represent the complex visibility for the distribution of the cor-

responding parameter over the source, not simply the intensity or brightness of the radiation. Equation (4.29) is a general and very useful formula that applies to all cases. It was originally derived by Morris, Radhakrishnan, and Seielstad (1964) and later by Weiler (1973). In the derivation by Morris et al. the sign of V_υ is opposite to that given by Weiler and in Eq. (4.29). This difference results from the convention for the sense of rotation for circular polarization. In the convention we have followed in Fig. 4.8, two identical antennas both adjusted to receive right circularly polarized radiation would have parameters $\psi_m = \psi_n$ and $\chi_m = \chi_n = -\pi/4$. In Eq. (4.29) these values correspond to a positive sign for V_υ. Thus in Eq. (4.29) positive V_υ represents right circular polarization incident from the sky, which is in agreement with the IAU definition. The derivation by Morris et al. predates the IAU definition and follows an earlier convention.

Note that in what follows the factor $1/2$ in Eq. (4.29) is omitted and considered to be subsumed within the overall gain factor.

Stokes Visibilities

As noted above, the symbols I_υ, Q_υ, U_υ, and V_υ in Eq. (4.29) refer to the corresponding visibility values as measured by the spaced antennas. We shall therefore refer to these quantities as *Stokes visibilities*, following the nomenclature of Hamaker, Bregman, and Sault (1996). Stokes visibilities are the quantities required in mapping polarized emission, and they can be derived from the correlator output values by using Eq. (4.29). This equation is considerably simplified when the nominal polarization characteristics of practical antennas are inserted. First consider the case where both antennas are identically polarized. Then $\chi_m = \chi_n$, $\psi_m = \psi_n$, and Eq. (4.29) becomes

$$R_{mn} = G_{mn}[I_\upsilon + Q_\upsilon \cos 2\psi_m \cos 2\chi_m + U_\upsilon \sin 2\psi_m \cos 2\chi_m - V_\upsilon \sin 2\chi_m]. \tag{4.30}$$

In considering linearly polarized antennas it is convenient to use subscripts x and y to indicate two orthogonal planes of polarization. For example, R_{xy} represents the correlator output for antenna m with polarization x and antenna n with polarization y. For linearly polarized antennas $\chi_m = \chi_n = 0$. Consider two antennas, each with separate outputs for linear polarizations x and y. Then for parallel polarizations, omitting gain constants, we obtain from Eq. (4.30)

$$R_{xx} = I_\upsilon + Q_\upsilon \cos 2\psi_m + U_\upsilon \sin 2\psi_m. \tag{4.31}$$

Here ψ_m is the position angle of the antenna polarization measured from celestial north in the direction of east. The y polarization angle is equal to the x polarization angle plus $\pi/2$. For ψ_m equal to 0°, 45°, 90°, and 135° the output R_{xx} is proportional to $(I_\upsilon + Q_\upsilon)$, $(I_\upsilon + U_\upsilon)$, $(I_\upsilon - Q_\upsilon)$, and $(I_\upsilon - U_\upsilon)$, respectively. By using antennas with these polarization angles, I_υ, Q_υ, and U_υ, but not V_υ, can

be measured. In many cases, circular polarization is negligibly small and the inability to measure V_v is not a serious problem. However, Q_v and U_v are often only a few percent of I_v, and in attempting to measure them with identical feeds one faces the usual problems of measuring a small difference in two much larger quantities. The same is true if one attempts to measure V_v using identical circular feeds for which $\chi = \pm\pi/4$ and the response is proportional to $(I_v \mp V_v)$. These problems are reduced by using oppositely polarized feeds to measure Q_v, U_v, or V_v. For an example of measurement of V_v, see Weiler and Raimond (1976).

With oppositely polarized feeds we insert in Eq. (4.29) $\psi_n = \psi_m + \pi/2$, and $\chi_m = -\chi_n$. For linear polarization the χ terms are zero and the planes of polarization orthogonal. The antennas are then described as cross-polarized, as typified by crossed dipoles. Omitting constant gain factors and using the x and y subscripts defined above, we obtain for the correlator output

$$\begin{aligned} R_{xy} &= -Q_v \sin 2\psi_m + U_v \cos 2\psi_m + jV_v \\ R_{yx} &= -Q_v \sin 2\psi_m + U_v \cos 2\psi_m - jV_v, \end{aligned} \tag{4.32}$$

where ψ_m refers to the angle of the plane of polarization in the direction (x or y) indicated by the first subscript of the R term in the same equation. Then for ψ_m equal to 0° and 45° the R_{xy} response is proportional to $(U_v + jV_v)$ and $(-Q_v + jV_v)$. If V_v is assumed to be zero, this suffices to measure the polarized component. If both antennas provide outputs for cross-polarized signals, the outputs of which go to two separate receiving channels at each antenna, four correlators can be used for each antenna pair. These provide responses for both crossed and parallel pairs as follows:

Position Angles			
m	n	Stokes Visibilities Measured	
0°	0°	$I_v + Q_v$	Position angle I
0°	90°	$U_v + jV_v$	"
90°	0°	$U_v - jV_v$	"
90°	90°	$I_v - Q_v$	"
45°	45°	$I_v + U_v$	Position angle II
45°	135°	$-Q_v + jV_v$	"
135°	45°	$-Q_v - jV_v$	"
135°	135°	$I_v - U_v$	"

(4.33)

Thus if the planes of polarization can be periodically rotated through 45° as indicated by position angles I and II above, for example, by rotating antenna feeds, then Q_v, U_v, and V_v can be measured without taking differences between re-

sponses involving I_υ. The use of rotating feeds has, however, proved to be of limited practicality. Rotating the feed relative to the main reflector is likely to have a small but significant effect on the beam shape and polarization properties. This is because the rotation will cause deviations from circular symmetry in the radiation pattern of the feeds to interact differently with the shadowing effects of the focal support structure and any departures from circular symmetry in the main reflector. Furthermore, in radio astronomy systems designed for the greatest sensitivity, the feed together with the low-noise amplifiers and a cryogenically refrigerated Dewar are often built as one monolithic unit that cannot easily be rotated. However, for antennas on altazimuth mounts, the variation of the parallactic angle with hour angle causes the antenna response pattern to rotate on the sky as a source is tracked in hour angle. Conway and Kronberg (1969) pointed out this advantage of altazimuth mounts, which enables instrumental effects to be distinguished from the true polarization of the source if observations continue for a period of several hours.

In the case of the Westerbork Synthesis Radio Telescope the antennas are equatorially mounted and the parallactic angle of the polarization remains fixed as a source is tracked. Crossed, linearly polarized feeds are used, as described by Weiler (1973). The outputs of a series of antennas that are movable on rail track are correlated with those from a series of antennas in fixed locations. The position angles of the planes of polarization for the movable antennas are 45° and 135° and those of the fixed antennas 0° and 90°. These angles result in the following responses:

Position Angle		
m	n	Stokes Visibilities Measured
0°	45°	$(I_\upsilon + Q_\upsilon + U_\upsilon + jV_\upsilon)/\sqrt{2}$
0°	135°	$(-I_\upsilon - Q_\upsilon + U_\upsilon + jV_\upsilon)/\sqrt{2}$
90°	45°	$(I_\upsilon - Q_\upsilon + U_\upsilon - jV_\upsilon)/\sqrt{2}$
90°	135°	$(I_\upsilon - Q_\upsilon - U_\upsilon + jV_\upsilon)/\sqrt{2}$

(4.34)

Although the responses are reduced by a factor of $\sqrt{2}$ relative to those in (4.33), there is no loss in sensitivity since each Stokes visibility appears at all four correlator outputs. Note, however, that since only signals from antennas with different polarization configurations are correlated, this scheme does not make use of all possible antenna pairs.

Opposite circularly polarized feeds offer certain advantages for measurements of linear polarization. In determining the responses we retain an arbitrary position angle ψ_m for antenna m to show the effect of rotation caused, for example, by an altazimuth antenna mount. If the antennas provide simultaneous outputs for opposite senses of rotation (denoted by r and ℓ) and four correlators are used for each antenna pair, the outputs are proportional to

Sense of Rotation		
m	n	Stokes Visibilities Measured
r	r	$I_v + V_v$
r	ℓ	$(-jQ_v + U_v)e^{-j2\psi_m}$
ℓ	r	$(-jQ_v - U_v)e^{j2\psi_m}$
ℓ	ℓ	$I_v - V_v$

(4.35)

Here we have made $\psi_\ell = \psi_r + \pi/2$, and $\chi = -\pi/4$ for right circular polarization and $\chi = \pi/4$ for left circular. The feeds need not be rotated during an observation, and the responses to Q_v and U_v are separated from those to I_v. The expressions in (4.35) can be simplified by choosing values of ψ_r such as $\pi/2$, $\pi/4$, or 0. For example, if $\psi_r = 0$, the sum of the $r\ell$ and ℓr responses is a measure of Stokes visibility U_v. Again, the effects of the rotation of the position angle with altazimuth mounts must be taken into account. Conway and Kronberg (1969) appear to have been the first to use an interferometer with circularly polarized antennas to measure linear polarization in weakly polarized sources. Circularly polarized antennas have since become widely used in radio astronomy.

Instrumental Polarization

The responses with the various combinations of linearly and circularly polarized antennas discussed above are derived on the assumption that the polarization is exactly linear or circular and that the position angles of the linear feeds are exactly determined. This is not the case in practice, and the polarization ellipse can never be maintained as a perfect circle or straight line. The nonideal characteristics of the antennas cause an unpolarized source to appear polarized and are therefore referred to as *instrumental polarization*. The effect of these deviations from ideal behavior can be calculated from Eq. (4.29) if the deviations are known. In the expressions in (4.33), (4.34), and (4.35) the responses given are only the major terms, and if the instrumental terms are included, all four Stokes visibilities are, in general, involved. For example, consider the case of crossed linear feeds with nominal position angles 0° and 90°. Let the actual values of ψ and χ be such that $(\psi_x + \psi_y) = \pi/2 + \Delta\psi^+$, $(\psi_x - \psi_y) = -\pi/2 + \Delta\psi^-$, $\chi_x + \chi_y = \Delta\chi^+$, $\chi_x - \chi_y = \Delta\chi^-$. Then from Eq. (4.29),

$$R_{xy} \simeq I_v(\Delta\psi^- - j\Delta\chi^+) - Q_v(\Delta\psi^+ - j\Delta\chi^-) + U_v + jV_v. \tag{4.36}$$

Generally, antennas can be adjusted so that the Δ terms are no more than $\sim 1°$, and here we have assumed that they are small enough that their cosines can be approximated by unity, their sines by the angles, and products of two sines by zero. Instrumental polarization is often different for each antenna even if they are structurally similar, and corrections must be made to the visibility data before they are combined into a map.

Although we have derived expressions for deviations of the antenna polarizations from the ideal in terms of the ellipticity and orientation of the polarization ellipse in Eq. (4.29), it is not necessary to know these parameters for the antennas so long as it is possible to remove the instrumental effects from the measurements, so that they do not appear in the final map or image. In calibrating the antenna responses, an approach that is widely preferred is to specify the instrumental polarization in terms of the response of the antenna to a wave of polarization that is orthogonal or opposite-handed with respect to the nominal antenna response. Thus, for linearly polarized antennas, following the analysis of Sault, Killeen, and Kesteven (1991), we can write

$$v'_x = v_x + D_x v_y, \quad \text{and} \quad v'_y = v_y + D_y v_x, \tag{4.37}$$

where subscripts x and y indicate two orthogonal planes of polarization, v' indicates the signal received, v indicates the signal that would be received with an ideally polarized antenna, and the D terms indicate the response of the real antenna to the polarization orthogonal to the nominal polarization. The D terms are often described as the *leakage* of the orthogonal polarization into the antenna (Bignell 1982) and represent the instrumental polarization. For each polarization state the leakage is specified by one complex number, that is, the same number of terms as the two real numbers required to specify the ellipticity and orientation of the polarization ellipse. In Appendix 4.2 expressions for D_x and D_y are derived in terms of the parameters of the polarization ellipse:

$$D_x \simeq \psi_x - j\chi_x, \quad \text{and} \quad D_y \simeq -\psi_y + j\chi_y, \tag{4.38}$$

where the approximations are valid for small values of the χ and ψ parameters. Note that in Eq. (4.38) ψ_y is measured with respect to the y direction. For an ideal linearly polarized antenna, χ_x and χ_y are both zero, and the polarization in the x and y planes is precisely aligned with, and orthogonal to, the x direction with respect to the antenna. Thus for an ideal antenna, ψ_x and ψ_y are also zero. For a practical antenna, the terms in Eqs. (4.38) represent limits of accuracy in the hardware, and we see that the real and imaginary parts of the leakage terms can be related to the misalignment and ellipticity, respectively.

For a pair of antennas, m and n, the leakage terms allow us to express the measured correlator outputs R'_{xx}, R'_{yy}, R'_{xy}, and R'_{yx} in terms of the unprimed quantities that represent the corresponding correlations as they would be measured with ideally polarized antennas:

$$\begin{aligned}
R'_{xx}/(g_{xm}g^*_{xn}) &= R_{xx} + D_{xm}R_{yx} + D^*_{xn}R_{xy} + D_{xm}D^*_{xn}R_{yy} \\
R'_{xy}/(g_{xm}g^*_{yn}) &= R_{xy} + D_{xm}R_{yy} + D^*_{yn}R_{xx} + D_{xm}D^*_{yn}R_{yx} \\
R'_{yx}/(g_{ym}g^*_{xn}) &= R_{yx} + D_{ym}R_{xx} + D^*_{xn}R_{yy} + D_{ym}D^*_{xn}R_{xy} \\
R'_{yy}/(g_{ym}g^*_{yn}) &= R_{yy} + D_{ym}R_{xy} + D^*_{yn}R_{yx} + D_{ym}D^*_{yn}R_{xx}.
\end{aligned} \tag{4.39}$$

The g terms represent the voltage gains of the corresponding signal channels. They are complex quantities representing amplitude and phase, and the equations can be normalized so that the values of the individual g terms do not differ greatly from unity. Note that Eqs. (4.39) contain no small-term approximations. However, the leakage terms are typically no more than a few percent, and products of two such terms will be omitted at this point. Then from Eqs. (4.31) and (4.32) the responses can be written in terms of the Stokes visibilities as follows:

$$\begin{aligned}
R'_{xx}/(g_{xm}g^*_{xn}) &= I_v + Q_v[\cos 2\psi_m - (D_{xm} + D^*_{xn})\sin 2\psi_m] \\
&\quad + U_v[\sin 2\psi_m + (D_{xm} + D^*_{xn})\cos 2\psi_m] - jV_v(D_{xm} - D^*_{xn}) \\
R'_{xy}/(g_{xm}g^*_{yn}) &= I_v(D_{xm} + D^*_{yn}) - Q_v[\sin 2\psi_m + (D_{xm} - D^*_{yn})\cos 2\psi_m] \\
&\quad + U_v[\cos 2\psi_m - (D_{xm} - D^*_{yn})\sin 2\psi_m] + jV_v \\
R'_{yx}/(g_{ym}g^*_{xn}) &= I_v(D_{ym} + D^*_{xn}) - Q_v[\sin 2\psi_m - (D_{ym} - D^*_{xn})\cos 2\psi_m] \\
&\quad + U_v[\cos 2\psi_m + (D_{ym} - D^*_{xn})\sin 2\psi_m] - jV_v \\
R'_{yy}/(g_{ym}g^*_{yn}) &= I_v - Q_v[\cos 2\psi_m + (D_{ym} + D^*_{yn})\sin 2\psi_m] \\
&\quad - U_v[\sin 2\psi_m - (D_{ym} + D^*_{yn})\cos 2\psi_m] + jV_v(D_{ym} - D^*_{yn}).
\end{aligned} \tag{4.40}$$

Note that ψ_m refers to the polarization (x or y) indicated by the first of the two subscripts of the R' term in the same equation. Sault, Killeen, and Kesteven (1991) describe Eqs. (4.40) as representing the strongly polarized case. In deriving them no restriction was placed on the magnitudes of the Stokes visibility terms, but the leakage terms of the antennas are assumed to be small. In the case where the source is only weakly polarized, the products of Q_v, U_v, and V_v with leakage terms can be omitted. Equations (4.40) then become

$$\begin{aligned}
R'_{xx}/(g_{xm}g^*_{xn}) &= I_v + Q_v\cos 2\psi_m + U_v\sin 2\psi_m \\
R'_{xy}/(g_{xm}g^*_{yn}) &= I_v(D_{xm} + D^*_{yn}) - Q_v\sin 2\psi_m + U_v\cos 2\psi_m + jV_v \\
R'_{yx}/(g_{ym}g^*_{xn}) &= I_v(D_{ym} + D^*_{xn}) - Q_v\sin 2\psi_m + U_v\cos 2\psi_m - jV_v \\
R'_{yy}/(g_{ym}g^*_{yn}) &= I_v - Q_v\cos 2\psi_m - U_v\sin 2\psi_m.
\end{aligned} \tag{4.41}$$

If the antennas are operating well within the upper frequency limit of their performance, the polarization terms can be expected to remain largely constant with time since gravitational deflections that vary with pointing should be small. The instrumental gain terms can contain components due to the atmosphere, which may vary on timescales of seconds or minutes, and they also include the effect of the electronics.

In the case of circularly polarized antennas, leakage terms can also be defined and similar expressions for the instrumental response derived. The leakage terms are given by the following equations:

$$v'_r = v_r + D_r v_\ell, \qquad v'_\ell = v_\ell + D_\ell v_r, \tag{4.42}$$

where, as before, the v' terms are the measured signal voltages, the unprimed v terms are the signals that would be observed with an ideally polarized antenna, and the D terms are the leakages. The subscripts r and ℓ indicate the right and left senses of rotation. Again, the relationship between the leakage terms and the orientation and ellipticity of the antenna responses is derived in Appendix 4.2. The results, which in this case require no small-angle approximations, are

$$D_r = e^{j2\psi_r} \tan \Delta\chi_r, \qquad D_\ell = e^{-j2\psi_\ell} \tan \Delta\chi_\ell, \tag{4.43}$$

where the Δ terms are defined by $\chi_r = -45° + \Delta\chi_r$ and $\chi_\ell = 45° + \Delta\chi_\ell$. To derive expressions for the outputs of an interferometer in terms of the leakage terms and Stokes visibilities, the four measured correlator outputs are represented by R'_{rr}, $R'_{\ell\ell}$, $R'_{r\ell}$, and $R'_{\ell r}$. These are related to the corresponding (unprimed) quantities that would be observed with ideally polarized antennas as follows:

$$\begin{aligned}
R'_{rr}/(g_{rm}g^*_{rn}) &= R_{rr} + D_{rm}R_{\ell r} + D^*_{rn}R_{r\ell} + D_{rm}D^*_{rn}R_{\ell\ell} \\
R'_{r\ell}/(g_{rm}g^*_{\ell n}) &= R_{r\ell} + D_{rm}R_{\ell\ell} + D^*_{\ell n}R_{rr} + D_{rm}D^*_{\ell n}R_{\ell r} \\
R'_{\ell r}/(g_{\ell m}g^*_{rn}) &= R_{\ell r} + D_{\ell m}R_{rr} + D^*_{rn}R_{\ell\ell} + D_{\ell m}D^*_{rn}R_{r\ell} \\
R'_{\ell\ell}/(g_{\ell m}g^*_{\ell n}) &= R_{\ell\ell} + D_{\ell m}R_{r\ell} + D^*_{\ell n}R_{\ell r} + D_{\ell m}D^*_{\ell n}R_{rr}.
\end{aligned} \tag{4.44}$$

Now from the expressions in (4.35) the outputs in terms of the Stokes visibilities are

$$\begin{aligned}
R'_{rr}/(g_{rm}g^*_{rn}) &= I_v(1 + D_{rm}D^*_{rn}) - jQ_v(D_{rm}e^{j2\psi_m} + D^*_{rn}e^{-j2\psi_m}) \\
&\quad - U_v(D_{rm}e^{j2\psi_m} - D^*_{rn}e^{-j2\psi_m}) + V_v(1 - D_{rm}D^*_{rn}) \\
R'_{r\ell}/(g_{rm}g^*_{\ell n}) &= I_v(D_{rm} + D^*_{\ell n}) - jQ_v(e^{-j2\psi_m} + D_{rm}D^*_{\ell n}e^{j2\psi_m}) \\
&\quad + U_v(e^{-j2\psi_m} - D_{rm}D^*_{\ell n}e^{j2\psi_m}) - V_v(D_{rm} - D^*_{\ell n}) \\
R'_{\ell r}/(g_{\ell m}g^*_{rn}) &= I_v(D_{\ell m} + D^*_{rn}) - jQ_v(e^{j2\psi_m} + D_{\ell m}D^*_{rn}e^{-j2\psi_m}) \\
&\quad - U_v(e^{j2\psi_m} - D_{\ell m}D^*_{rn}e^{-j2\psi_m}) + V_v(D_{\ell m} - D^*_{rn}) \\
R'_{\ell\ell}/(g_{\ell m}g^*_{\ell n}) &= I_v(1 + D_{\ell m}D^*_{\ell n}) - jQ_v(D_{\ell m}e^{-j2\psi_m} + D^*_{\ell n}e^{j2\psi_m}) \\
&\quad + U_v(D_{\ell m}e^{-j2\psi_m} - D^*_{\ell n}e^{j2\psi_m}) - V_v(1 - D_{\ell m}D^*_{\ell n}).
\end{aligned} \tag{4.45}$$

Here again, ψ_m refers to the polarization (r or ℓ) indicated by the first of the two subscripts of the R' term in the same equation. The angle ψ_m represents the parallactic angle plus any instrumental offset. We have made no approximations in deriving Eqs. (4.45) [in the similar Eqs. (4.40), products of two D terms were omitted]. If the leakage terms are small, then any product of two of them can be omitted, as in the strongly polarized case for linearly polarized antennas in Eqs. (4.40). The weakly polarized case is derived from the strongly polarized case by further omitting products of Q_v, U_v, and V_v with the leakage terms, and

is as follows:

$$
\begin{aligned}
R'_{rr}/(g_{rm}g^*_{rn}) &= I_v + V_v \\
R'_{r\ell}/(g_{rm}g^*_{\ell n}) &= I_v(D_{rm} + D^*_{\ell n}) - (jQ_v - U_v)e^{-j2\psi_m} \\
R'_{\ell r}/(g_{\ell m}g^*_{rn}) &= I_v(D_{\ell m} + D^*_{rn}) - (jQ_v + U_v)e^{j2\psi_m} \\
R'_{\ell\ell}/(g_{\ell m}g^*_{\ell n}) &= I_v - V_v.
\end{aligned}
\tag{4.46}
$$

Similar expressions* are given by Fomalont and Perley (1989). To make use of the expressions that have been derived for the response in terms of the leakage and gain factors, we need to consider how such quantities can be calibrated, and this is discussed later.

Matrix Formulation

The description of polarimetry given above, using the ellipticity and orientation of the antenna response, is based on a physical model of the antenna and the electromagnetic wave. Historically, studies of optical polarization have developed over a much longer period. A description of radio polarimetry following an approach originally developed in optics is given by Hamaker, Bregman, and Sault in four papers (Hamaker, Bregman, and Sault 1996; Sault, Hamaker, and Bregman 1996; Hamaker and Bregman 1996; Hamaker 2000), and also described in Hamaker (1996). The mathematical analysis is largely in terms of matrix algebra, and in particular it allows the responses of different elements of the signal path such as the atmosphere, the antennas, and the electronic system, to be represented independently and then combined in the final solution.

In the matrix formulation the electric fields of the polarized wave are represented by a two-component column vector. The effect of any linear system on the wave, or on the voltage waveforms of the signal after reception, can be represented by a 2×2 matrix of the form shown below:

$$
\begin{bmatrix} E'_p \\ E'_q \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} E_p \\ E_q \end{bmatrix}, \tag{4.47}
$$

where E_p and E_q represent the input polarization state (orthogonal linear or opposite circular) and E'_p and E'_q represent the outputs. The 2×2 matrix in Eq. (4.47) is referred to as a Jones matrix (Jones 1941), and any simple linear operation on the wave can be represented by such a matrix. Jones matrices can represent a rotation of the wave relative to the antenna; the response of the antenna, including polarization leakage effects; or the amplification of the signals in the receiving system up to the correlator input. The combined effect of these operations is rep-

*Note that in comparing expressions for polarimetry by different authors, differences of signs or of the factor j can result from differences in the way the parallactic angle is defined with respect to the antenna, and similar arbitrary factors.

resented by the product of the corresponding Jones matrices, just as the effect on a scalar voltage can be represented by the product of gains and response factors for different stages of the receiving system. For a wave specified in terms of opposite circularly polarized components, Jones matrices for these operations can take the following forms:

$$\mathbf{J}_{\text{rotation}} = \begin{bmatrix} \exp(j\theta) & 0 \\ 0 & \exp(-j\theta) \end{bmatrix} \tag{4.48}$$

$$\mathbf{J}_{\text{leakage}} = \begin{bmatrix} 1 & D_r \\ D_\ell & 1 \end{bmatrix} \tag{4.49}$$

$$\mathbf{J}_{\text{gain}} = \begin{bmatrix} G_r & 0 \\ 0 & G_\ell \end{bmatrix}. \tag{4.50}$$

Here θ represents a rotation, and the cross polarization in the antenna is represented by the off-diagonal leakage terms D_r and D_ℓ. For a nonideal antenna the diagonal terms will be slightly different from unity, but in this case the difference is subsumed into the gain matrix of the two channels. The gain of both the antenna and the electronics can be represented by a single matrix, and since any cross-coupling of the signals in the amplifiers can be made negligibly small, only the diagonal terms are nonzero in the gain matrix.

Let $\mathbf{J}_m$ represent the product of the Jones matrices required to represent the linear operations on the signal of antenna m up to the point where it reaches the correlator input. Let $\mathbf{J}_n$ be the same matrix for antenna n. The signals at the inputs to the correlator are $\mathbf{J}_m\mathbf{E}_m$ and $\mathbf{J}_n\mathbf{E}_n$, where $\mathbf{E}_m$ and $\mathbf{E}_n$ are the vectors representing the signals at the antenna. The correlator output is the *outer product* (also known as the Krönecker, or tensor, product) of the signals at the input:

$$\mathbf{E}'_m \otimes \mathbf{E}'^*_n = (\mathbf{J}_m\mathbf{E}_n) \otimes (\mathbf{J}^*_n\mathbf{E}^*_n), \tag{4.51}$$

where $\otimes$ represents the outer product. The outer product $\mathbf{A} \otimes \mathbf{B}$ is formed by replacing each element a_{ik} of $\mathbf{A}$ by $a_{ik}\mathbf{B}$. Thus the outer product of two $n \times n$ matrices is a matrix of order $n^2 \times n^2$. It is also a property of the outer product that

$$(\mathbf{A}_i\mathbf{B}_i) \otimes (\mathbf{A}_k\mathbf{B}_k) = (\mathbf{A}_i \otimes \mathbf{A}_k)(\mathbf{B}_i \otimes \mathbf{B}_k). \tag{4.52}$$

Thus we can write Eq. (4.51) as

$$\mathbf{E}'_m \otimes \mathbf{E}'^*_n = (\mathbf{J}_m \otimes \mathbf{J}^*_n)(\mathbf{E}_m \otimes \mathbf{E}^*_n). \tag{4.53}$$

The time average of Eq. (4.53) represents the correlator output, which is

$$\mathbf{R}_{mn} = \langle \mathbf{E}'_m \otimes \mathbf{E}'^*_n \rangle = \begin{bmatrix} R^{pp}_{mn} \\ R^{pq}_{mn} \\ R^{qp}_{mn} \\ R^{qq}_{mn} \end{bmatrix}, \tag{4.54}$$

where p and q indicate opposite polarization states. The column vector in Eq. (4.54) is known as the coherency vector and represents the four cross-products from the correlator outputs for antennas m and n. From Eq. (4.53) it is evident that the measured coherency vector $\mathbf{R}'_{mn}$, which includes the effects of instrumental responses, and the true coherency vector $\mathbf{R}_{mn}$, which is free from such effects, are related by the outer product of the Jones matrices that represent the instrumental effects:

$$\mathbf{R}'_{mn} = (\mathbf{J}_m \otimes \mathbf{J}_n^*)(\mathbf{E}_m \otimes \mathbf{E}_n^*)\mathbf{R}_{mn}. \tag{4.55}$$

To determine the response of an interferometer in term of the Stokes visibilities of the input radiation, which are complex quantities, we introduce the Stokes visibility vector

$$\mathbf{V}_{Smn} = \begin{bmatrix} I_v \\ Q_v \\ U_v \\ V_v \end{bmatrix}. \tag{4.56}$$

The Stokes visibilities can be regarded as an alternate coordinate system for the coherency vector. Let $\mathbf{S}$ be a 4×4 transformation matrix from Stokes parameters to the polarization coordinates of the antennas. Then we have

$$\mathbf{R}'_{mn} = (\mathbf{J}_m \otimes \mathbf{J}_n^*)\mathbf{S}\mathbf{V}_{Smn}. \tag{4.57}$$

For ideal antennas with crossed (orthogonal) linear polarization, the response in terms of Stokes visibilities is given by the expressions in (4.33). We can write this result in matrix form as

$$\begin{bmatrix} R_{xx} \\ R_{xy} \\ R_{yx} \\ R_{yy} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & j \\ 0 & 0 & 1 & -j \\ 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} I_v \\ Q_v \\ U_v \\ V_v \end{bmatrix}, \tag{4.58}$$

where the subscripts x and y here refer to polarization position angles 0° and 90° respectively. Similarly for opposite-hand circular polarization, we can write the expressions in (4.35) as

$$\begin{bmatrix} R_{rr} \\ R_{r\ell} \\ R_{\ell r} \\ R_{\ell\ell} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & -je^{-j2\psi_m} & e^{-j2\psi_m} & 0 \\ 0 & -je^{j2\psi_m} & -e^{j2\psi_m} & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} I_v \\ Q_v \\ U_v \\ V_v \end{bmatrix}. \tag{4.59}$$

The 4×4 matrices in Eqs. (4.58) and (4.59) provide transformation matrices from Stokes visibilities to the coherency vector for crossed linear and opposite circular polarizations, respectively. Note that these matrices depend on the particular formulation we have used to specify the angles ψ and χ, and other factors in Fig. 4.8, which may not be identical to corresponding parameters used by other authors.

The expression $\mathbf{S}^{-1}(\mathbf{J}_m \otimes \mathbf{J}_n^*)\mathbf{S}$ is a matrix that relates the input and output coherency vectors of a system where these quantities are in Stokes coordinate form. In optics this type of matrix is known as a Mueller matrix (Mueller 1948). Further explanations of Jones and Mueller matrices can be found in some textbooks on optics [e.g., O'Neill (1963)].

As an example of the matrix usage, let us derive the effect of the leakage and gain factors in the case of opposite circular polarizations. For antenna m, the Jones matrix $\mathbf{J}_m$ is the product of the Jones matrices for leakage and gain as follows:

$$\mathbf{J}_m = \begin{bmatrix} g_{rm} & 0 \\ 0 & g_{\ell m} \end{bmatrix} \begin{bmatrix} 1 & D_{rm} \\ D_{\ell m} & 1 \end{bmatrix} = \begin{bmatrix} g_{rm} & g_{rm} D_{rm} \\ g_{\ell m} D_{\ell m} & g_{\ell m} \end{bmatrix}. \tag{4.60}$$

Here the g terms represent voltage gain, the D terms represent leakage, and the subscripts r and ℓ indicate polarization. A corresponding matrix $\mathbf{J}_n$ is required for antenna n. Then if we use primes to indicate the components of the coherency vector (i.e., the correlator outputs) for antennas m and n, we can write

$$\begin{bmatrix} R'_{rr} \\ R'_{r\ell} \\ R'_{\ell r} \\ R'_{\ell\ell} \end{bmatrix} = \mathbf{J}_m \otimes \mathbf{J}_n^* \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & -je^{-j2\psi_m} & e^{-j2\psi_m} & 0 \\ 0 & -je^{j2\psi_m} & -e^{j2\psi_m} & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} I_v \\ Q_v \\ U_v \\ V_v \end{bmatrix}, \tag{4.61}$$

where the 4×4 matrix is the one relating Stokes visibilities to the coherency vector in Eq. (4.59). Also we have

$$\mathbf{J}_m \otimes \mathbf{J}_n^* = \begin{bmatrix} g_{rm} g_{rn}^* & g_{rm} g_{rn}^* D_{rn}^* & g_{rm} g_{rn}^* D_{rm} & g_{rm} g_{rn}^* D_{rm} D_{rn}^* \\ g_{rm} g_{\ell n}^* D_{\ell n}^* & g_{rm} g_{\ell n}^* & g_{rm} g_{\ell n}^* D_{rm} D_{\ell n}^* & g_{rm} g_{\ell n}^* D_{rm} \\ g_{\ell m} g_{rn}^* D_{\ell m} & g_{\ell m} g_{rn}^* D_{\ell m} D_{rn}^* & g_{\ell m} g_{rn}^* & g_{\ell m} g_{rn}^* D_{rn}^* \\ g_{\ell m} g_{\ell n}^* D_{\ell m} D_{\ell n}^* & g_{\ell m} g_{\ell n}^* D_{\ell m} & g_{\ell m} g_{\ell n}^* D_{\ell n}^* & g_{\ell m} g_{\ell n}^* \end{bmatrix}. \tag{4.62}$$

Insertion of Eq. (4.62) into Eq. (4.61) and reduction of the matrix products results in Eqs. (4.45) for the response with circularly polarized feeds.

Calibration of Instrumental Polarization†

The fractional polarization of almost all astronomical sources is of magnitude comparable to that of the leakage and gain terms that are used above to define the

† A general description of calibration is given in Chapter 10, in particular Section 10.1, which it may be helpful to read before embarking on the more complicated considerations of polarization calibration. The discussion of polarization calibration is placed in this chapter for ease of reference to the development of polarimetry given here.

instrumental polarization. Thus to obtain an accurate measure of the polarization of a source the leakage and gain terms must be accurately calibrated. It is usually necessary to determine the calibration independently for each set of observations since the gain terms are likely to be functions of the temperature and state of adjustment of the electronics, and cannot be assumed to remain constant from one observing session to another. Making observations (i.e., measuring the coherency vector) of sources for which the polarization parameters are already known is clearly a way of determining the leakage and gain terms. The number of unknown parameters to be calibrated is proportional to the number of antennas, n_a, but the number of measurements is proportional to the number of baselines, $n_a(n_a - 1)/2$. The unknown parameters are therefore usually overdetermined, and a least-squares solution is usually the preferred procedure.

For any antenna with orthogonally polarized receiving channels there are seven degrees of freedom, that is, seven unknown quantities, that must be calibrated to allow full interpretation of the measured Stokes visibilities. This applies to the general case, and the number can be reduced if approximations are made for weak polarization or small instrumental polarization. In terms of the polarization ellipses, these unknowns can be regarded as the orientations and ellipticities of the two orthogonal feeds and the complex gains (amplitudes and phases) of the two receiving channels. When the outputs of two antennas are combined, only the differences in the instrumental phases are required, leaving seven degrees of freedom per antenna. Sault, Hamaker, and Bregman (1996) make the same point from the consideration of the Jones matrix of an antenna, which contains four complex quantities. They also give a general result that illustrates the seven degrees of freedom or unknown terms. This expresses the relationship between the uncorrected (measured) Stokes visibilities (indicated by primes) and the true values of the Stokes visibilities, in terms of seven γ and δ terms:

$$\begin{bmatrix} I'_v - I_v \\ Q'_v - Q_v \\ U'_v - U_v \\ V'_v - V_v \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} \gamma_{++} & \gamma_{+-} & \delta_{+-} & -j\delta_{-+} \\ \gamma_{+-} & \gamma_{++} & \delta_{++} & -j\delta_{--} \\ \delta_{+-} & -\delta_{++} & \gamma_{++} & j\gamma_{--} \\ -j\delta_{-+} & -j\delta_{--} & j\gamma_{--} & j\gamma_{++} \end{bmatrix} \begin{bmatrix} I_v \\ Q_v \\ U_v \\ V_v \end{bmatrix}. \quad (4.63)$$

The seven γ and δ terms are defined as follows:

$$\begin{aligned}
\gamma_{++} &= (\Delta g_{xm} + \Delta g_{ym}) + (\Delta g^*_{xn} + \Delta g^*_{yn}) \\
\gamma_{+-} &= (\Delta g_{xm} - \Delta g_{ym}) + (\Delta g^*_{xn} - \Delta g^*_{yn}) \\
\gamma_{--} &= (\Delta g_{xm} - \Delta g_{ym}) - (\Delta g^*_{xn} - \Delta g^*_{yn}) \\
\delta_{++} &= (D_{xm} + D_{ym}) + (D^*_{xn} + D^*_{yn}) \\
\delta_{+-} &= (D_{xm} - D_{ym}) + (D^*_{xn} - D^*_{yn}) \\
\delta_{-+} &= (D_{xm} + D_{ym}) - (D^*_{xn} + D^*_{yn}) \\
\delta_{--} &= (D_{xm} - D_{ym}) - (D^*_{xn} - D^*_{yn}).
\end{aligned} \quad (4.64)$$

Here it is assumed that Eqs. (4.39) are normalized so that the gain terms are close to unity, and the Δg terms are defined by $g_{ik} = 1 + \Delta g_{ik}$. The D (leakage) terms and the Δg terms are small enough that products of two such terms can be neglected. The results, as shown in Eqs. (4.63) and (4.64), apply to antennas that are linearly polarized in directions x and y. The same results apply to circularly polarized antennas if the subscripts x and y are replaced by r and ℓ, respectively, and, in the column matrix on the right-hand side of Eq. (4.63), Q_v, U_v, and V_v are replaced by V_v, Q_v, and U_v respectively. A similar result is given by Sault, Killeen, and Kesteven (1991). The seven γ and δ terms defined above are subject to errors in the calibration process, so there are seven degrees of freedom in the error mechanisms.

An observation of a single calibration source for which the four Stokes parameters are known enables four of the degrees of freedom to be determined. However, because of the relationships of the quantities involved, it takes at least three calibration observations to solve for all seven unknown parameters (Sault, Hamaker, and Bregman 1996). In the calibration observations it is useful to observe one unpolarized source, but observing a second unpolarized one would add no further solutions. At least one observation of a linearly polarized source is required to determine the relative phases of the two oppositely polarized channels, that is, the relative phases of the complex gain terms $g_{xm}g^*_{yn}$ and $g_{ym}g^*_{xn}$, or $g_{rm}g^*_{\ell n}$ and $g_{\ell m}g^*_{rn}$. Note that with antennas on altazimuth mounts, observations of a calibrator with linear polarization, taken at intervals between which large rotations of the parallactic angle occur, can essentially be regarded as observations of independent calibrators. Under these circumstances three observations of the same calibrator will suffice for the full solution. Furthermore, the polarization of the calibrator need not be known in advance but can be determined from the observations.

In cases where only an unpolarized calibrator can be observed, it is possible to determine two more degrees of freedom by introducing the constraint that the sum of the leakage factors over all antennas should be zero. As shown by the expressions for the leakage terms in Appendix 4.2, this is a reasonable assumption for a homogeneous array, that is, one in which the antennas are of nominally identical design. However, the phase difference between the signal paths from the feeds to the correlator for the two orthogonal polarizations of each antenna remains unknown. This requires an observation of a calibrator with a component of linear polarization, or a scheme to measure the instrumental component of the phase. For example, on the compact array of the Australia Telescope (Frater and Brooks 1992), noise sources are provided at each antenna to inject a common signal into the two polarization channels (Sault, Hamaker, and Bregman 1996). With such a system it is necessary to provide an additional correlator for each antenna, or to be able to rearrange correlator inputs, to measure the relative phase of the injected signals in the two polarizations.

In the case of the approximations for weak polarization, Eqs. (4.41) and (4.46) show that if the gain terms are known, the leakage terms can be calibrated by observing an unpolarized source. For opposite circular polarizations, Eq. (4.46) shows that if V_v is small, it is possible to obtain solutions for the gain terms from the outputs for the $\ell\ell$ and rr combinations only, provided also that the

number of baselines is several times larger than the number of antennas. The leakage terms can then be solved for separately. For crossed linear polarizations, Eq. (4.41) shows that this is only possible if the linear polarization (Q_ν and U_ν parameters) for the calibrator have been determined independently.

Optimum strategies for calibration of polarization observations is a subject that leads to highly detailed discussions involving the characteristics of particular synthesis arrays, the hour angle range of the observations, the availability of calibration sources (which can depend on the observing frequency), and other factors, especially if the solutions for strong polarization are used. Such discussions can be found, for example, in Conway and Kronberg (1969), Weiler (1973), Bignell (1982), Sault, Killeen, and Kesteven (1991), Sault, Hamaker, and Bregman (1996), and Smegal et al. (1997). Polarization measurements with VLBI involve some special considerations: see, for example, Roberts, Brown, and Wardle (1991), Cotton (1993), Roberts, Wardle, and Brown (1994), Kemball, Diamond, and Cotton (1995).

For most large synthesis arrays, effective calibration techniques have been devised and the software to implement them has been developed. Thus a prospective observer need not be discouraged if the necessary calibration procedures appear complicated. Some general considerations relevant to observations of polarization are given below:

- Since the polarization of many sources varies on a timescale of months, it is usually advisable to regard the polarization of the calibration source as one of the variables to be solved for.
- Two sources with relatively strong linear polarization at position angles that do not appear to vary are 3C286 and 3C138. These are useful for checking the phase difference for oppositely polarized channels.
- For most sources the circular polarization parameter V_ν is very small, ~0.2% or less, and can be neglected. Measurements with circularly polarized antennas of the same sense therefore generally give an accurate measure of I_ν. However, circular polarization is important in measurement of magnetic fields by Zeeman splitting. As an example of positive detection at a very low level, Fiebig and Güsten (1989) describe measurements for which $V/I \simeq 5 \times 10^{-5}$. Zeeman splitting of several components of the OH line at 22.235 GHz were observed using a single antenna, the 100-m paraboloid of the Max Planck Institute for Radio Astronomy, with a receiving system that switched between opposite circular polarizations at 10 Hz. Rotation of the feed and receiver unit was used to identify spurious instrumental responses to linearly polarized radiation, and calibration of the relative pointing of the two beams to one arcsecond accuracy was required.
- Although the polarized emission from most sources is small compared with the total emission, it is possible for Stokes visibilities Q_ν and U_ν to be comparable to I_ν in cases where there is a broad unpolarized component that is highly resolved and a narrower polarized component that is not resolved. In such cases errors may occur if the approximations for weak polarization [Eqs. (4.41) and (4.46)] are used in the data analysis.

- For most antennas the instrumental polarization varies over the main beam and increases toward the beam edges. Sidelobes that are cross polarized relative to the main beam tend to peak near the beam edges. Thus polarization measurements are usually made for cases where the source is small compared with the width of the main beam, and for such measurements the beam should be centered on the source.
- Faraday rotation of the plane of polarization of incoming radiation occurs in the ionosphere, and becomes important for frequencies below a few gigahertz; see Table 13.6 in Chapter 13. During polarization measurements periodic observations of a strongly polarized source are useful for monitoring changes in the rotation, which varies with the total column density of electrons in the ionosphere. If not accounted for, Faraday rotation can cause errors in calibration; see, for example, Sakurai and Spangler (1994).
- In some antennas the feed is displaced from the axis of the main reflector, for example, when the Cassegrain focus is used and the feeds for different bands are located in a circle around the vertex. For circularly polarized feeds, this departure from circular symmetry results in pointing offsets of the beams for the two opposite hands. The pointing directions of the two beams are typically separated by ~0.1 beamwidths, which makes measurements of circular polarization difficult or impractical because V_ν is proportional to $(R_{rr} - R_{\ell\ell})$. For linearly polarized feeds the corresponding effect is an increase in the cross polarized sidelobes near the beam edges. See also Section 5.1.
- In VLBI the large distances between antennas result in different parallactic angles at different sites, which must be taken into account.
- The quantities m_ℓ and m_t, of Eqs. (4.20) and (4.22), have Rice distributions of the form of Eq. (6.63a), and the position angle has a distribution of the form of Eq. (6.63b). The percentage polarization can be overestimated, and a correction should be applied (Wardle and Kronberg 1974). The discussion in Section 9.3, at the end of *Noise in VLBI Observations*, is relevant to this problem.

The following points concern choices in designing an array for polarization measurements:

- The rotation of an antenna on an altazimuth mount relative to the sky is in most cases a distinct advantage in polarimetry. The rotation could be a disadvantage in cases where polarization mapping over a large part of the antenna beam is being attempted. Correction for the variation of instrumental polarization over the beam could be more difficult if the beam rotates on the sky.
- With linearly polarized antennas, errors in calibration are likely to cause I_ν to corrupt the linear parameters Q_ν and U_ν, so for measurement of linear polarization, circularly polarized antennas offer an advantage. Similarly, with circularly polarized antennas, calibration errors are likely to cause I_ν to corrupt V_ν, so for measurements of circular polarization, linearly polarized antennas are to be preferred.

- Linearly polarized feeds for reflector antennas can be made with relative bandwidths of at least 2 : 1, whereas for circularly polarized feeds the maximum relative bandwidth is commonly about 1.4 : 1. In many designs of circularly polarized feeds, orthogonal linear components of the field are combined with $\pm 90°$ relative phase shifts, and the phase shifting element limits the bandwidth. For this reason linear polarization is the choice for some synthesis arrays such as the Australia Telescope (James 1992), and with careful calibration good polarization performance is obtainable.
- The stability of the instrumental polarization, which greatly facilitates accurate calibration over a wide range of hour angle, is perhaps the most important feature to be desired. Caution should therefore be used if feeds are rotated relative to the main reflector, or if antennas are used near the high end of their frequency range.

APPENDIX 4.1 CONVERSION BETWEEN HOUR ANGLE–DECLINATION AND AZIMUTH–ELEVATION COORDINATES

Although the positions of cosmic sources are almost always specified in celestial coordinates, for purposes of observation it is generally necessary to convert to elevation and azimuth. The conversion formulas between hour angle and declination (H, δ) and elevation and azimuth $(\mathcal{E}, \mathcal{A})$ can be derived by applying the sine and cosine rules for spherical triangles to the system in Fig. 4.3. For an observer at latitude $\mathcal{L}$ they are, for (H, δ) to $(\mathcal{A}, \mathcal{E})$,

$$\begin{aligned} \sin \mathcal{E} &= \sin \mathcal{L} \sin \delta + \cos \mathcal{L} \cos \delta \cos H \\ \cos \mathcal{E} \cos \mathcal{A} &= \cos \mathcal{L} \sin \delta - \sin \mathcal{L} \cos \delta \cos H \\ \cos \mathcal{E} \sin \mathcal{A} &= -\cos \delta \sin H. \end{aligned} \tag{A4.1}$$

Similarly for $(\mathcal{A}, \mathcal{E})$ to (H, δ) we obtain

$$\begin{aligned} \sin \delta &= \sin \mathcal{L} \sin \mathcal{E} + \cos \mathcal{L} \cos \mathcal{E} \cos \mathcal{A} \\ \cos \delta \cos H &= \cos \mathcal{L} \sin \mathcal{E} - \sin \mathcal{L} \cos \mathcal{E} \cos \mathcal{A} \\ \cos \delta \sin H &= -\cos \mathcal{E} \sin \mathcal{A}. \end{aligned} \tag{A4.2}$$

Here azimuth is measured from north through east.

APPENDIX 4.2 LEAKAGE PARAMETERS IN TERMS OF THE POLARIZATION ELLIPSE

The polarization leakage terms used to express the instrumental polarization are related to the ellipticity and orientation of the polarization ellipses of each antenna, as shown below.

Linear Polarization

Consider the antenna in Fig. 4.8, and suppose that it is nominally linearly polarized in the x direction, in which case χ and ψ are small angles that represent engineering tolerances. A field E aligned with the x axis in Fig. 4.8a produces components $E_{x'}$ and $E_{y'}$ along the (x', y') axes with which the dipoles in Fig. 4.8b are aligned. Then from Eq. (4.26) we obtain the voltage at the output of the antenna (point A in Fig. 4.8b), which is

$$V_x' = E(\cos\psi\cos\chi + j\sin\psi\sin\chi). \tag{A4.3}$$

The response to the same field, but aligned with the y axis, is

$$V_y' = E(\sin\psi\cos\chi - j\cos\psi\sin\chi). \tag{A4.4}$$

V_x' represents the wanted response to the field along the x axis and V_y' represents the unwanted response to a cross-polarized field. The leakage term is equal to the cross-polarized response expressed as a fraction of the wanted x-polarization response, that is,

$$D_x = \frac{V_y'}{V_x'} = \frac{(\sin\psi_x\cos\chi_x - j\cos\psi_x\sin\chi_x)}{(\cos\psi_x\cos\chi_x + j\sin\psi_x\sin\chi_x)} \simeq \psi_x - j\chi_x, \tag{A4.5}$$

where the subscript x indicates the x-polarization case. The corresponding term D_y for the condition in which Fig. 4.8 represents the nominal y polarization of the antenna is obtained as V_x'/V_y' by inverting Eq. (A4.5), replacing ψ_x by $\psi_y + \pi/2$, and replacing χ_x by χ_y. Then ψ_y is measured from the y axis in the same sense as ψ_x is measured from the x axis, that is, increasing in a counterclockwise direction in Fig. 4.8. Thus we obtain

$$\begin{aligned} D_y = \frac{V_x'}{V_y'} &= \frac{[\cos(\psi_y + \pi/2)\cos\chi_y + j\sin(\psi_y + \pi/2)\sin\chi_y]}{[\sin(\psi_y + \pi/2)\cos\chi_y - j\cos(\psi_y + \pi/2)\sin\chi_y]} \\ &= \frac{(-\sin\psi_y\cos\chi_y + j\cos\psi_y\sin\chi_y)}{(\cos\psi_y\cos\chi_y + j\sin\psi_y\sin\chi_y)} \simeq -\psi_y + j\chi_y. \end{aligned} \tag{A4.6}$$

Similar expressions for D_x and D_y have also been derived by Sault, Killeen, and Kesteven (1991). Note that D_x and D_y are of comparable magnitude and opposite sign, so one would expect the average of all the D terms for an array of antennas to be very small. As used earlier in this chapter, subscripts m and n are added to the D terms to indicate individual antennas.

Circular Polarization

To receive right circular polarization from the sky, the antenna in Fig. 4.8b must respond to a field with counterclockwise rotation in the plane of the diagram, as explained earlier. This requires $\chi = -45°$. In terms of fields in the x and y directions, counterclockwise rotation requires that E_x leads E_y in phase by $\pi/2$; that is, $E_x = jE_y$ for the fields as defined in Eq. (4.25). For fields E_x and E_y, we determine the components in the x' and y' directions, and then obtain expressions for the output of the antenna for both counterclockwise and clockwise rotation of the incident field. For counterclockwise rotation

$$E'_x = E_x \cos\psi + E_y \sin\psi = E_x(\cos\psi - j\sin\psi), \tag{A4.7}$$

$$E'_y = -E_x \sin\psi + E_y \cos\psi = -E_x(\sin\psi + j\cos\psi). \tag{A4.8}$$

For nominal right-circular polarization, $\chi_r = -\pi/4 + \Delta\chi_r$, where $\Delta\chi_r$ is a measure of the departure of the polarization from circularity. Then from Eq. (4.26), we obtain

$$V'_r = E_x e^{-j\psi_r}(\cos\chi_r - \sin\chi_r) = \sqrt{2}E_x e^{-j\psi_r}\cos\Delta\chi_r. \tag{A4.9}$$

The next step is to repeat the procedure for left circular polarization from the sky, for which we have clockwise rotation of the electric vector and $E_y = jE_x$. The result is

$$V'_\ell = E_x e^{j\psi_r}(\cos\chi_r + \sin\chi_r) = \sqrt{2}E_x e^{j\psi_r}\sin\Delta\chi_r. \tag{A4.10}$$

The relative magnitude of the opposite-hand response of the nominally right-handed polarization state, that is, the leakage term, is

$$D_r = \frac{V'_\ell}{V'_r} = e^{j2\psi_r}\tan\Delta\chi_r \simeq e^{j2\psi_r}\Delta\chi_r. \tag{A4.11}$$

For the nominal left-hand polarization the relative magnitude of the opposite-hand response is obtained by inverting the right-hand side of Eq. (A4.11) and also substituting $\Delta\chi_\ell + \pi/2$ for $\Delta\chi_r$ and $\psi_\ell - \pi/2$ for ψ_r. For the corresponding leakage term D_ℓ, which represents the right circular leakage of the nominally left circularly polarized antenna, we then obtain

$$D_\ell = e^{-j2\psi_\ell}\tan\Delta\chi_\ell \simeq e^{-j2\psi_\ell}\Delta\chi_\ell. \tag{A4.12}$$

Since $-\pi/4 \leq \chi \leq \pi/4$, $\Delta\chi_r$ and $\Delta\chi_\ell$ take opposite signs. Thus, as in the case of the leakage terms for linear polarization, D_r and D_ℓ are of comparable magnitude and opposite sign.

REFERENCES

Bignell, R. C., Polarization, in *Synthesis Mapping, Proceedings of NRAO Workshop No. 5*, Socorro, NM, June 21–25, 1982, A. R. Thompson and L. R. D'Addario, Eds., National Radio Astronomy Observatory, Green Bank, WV, 1982.

Born, M. and E. Wolf, *Principles of Optics*, 7th ed, Cambridge Univ. Press, Cambridge, UK, 1999.

Conway, R. G. and P. P. Kronberg, Interferometer Measurement of Polarization Distribution in Radio Sources, *Mon. Not. R. Astron. Soc.*, **142**, 11–32, 1969.

Cotton, W. D., Calibration and Imaging of Polarization Sensitive Very Long Baseline Interferometer Observations, *Astron. J.*, **106**, 1241–1248, 1993.

Fiebig, D. and R. Güsten, Strong Magnetic Fields in Interstellar Maser Clumps, *Astron. Astrophys.*, **214**, 333–338, 1989.

Fomalont, E. B. and R. A. Perley, Calibration and Editing, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific, Conf. Ser., **6**, 83–115, 1989.

Frater, R. H. and J. W. Brooks, Eds., *J. Electric. Electron. Eng. Australia*, Special Issue on the Australian Telescope, **12**, No. 2, 1992.

Hamaker, J. P., A New Theory of Radio Polarimetry with an Application to the Westerbork Synthesis Radio Telescope (WSRT), in *Workshop on Large Antennas in Radio Astronomy*, ESTEC, Noordwijk, The Netherlands, 1996.

Hamaker, J. P., Inderstanding Radio Polarimetry. IV. The Full-Coherency Analogue of Scalar Self-Calibration: Self-alignment, Dynamic Range, and Polarimetric Fidelity, *Astron. Astrophys. Suppl.*, **143**, 515–543, 2000.

Hamaker, J. P., J. D. Bregman, and R. J. Sault, Understanding Radio Polarimetry. I. Mathematical Foundations, *Astron. Astrophys. Suppl.*, **117**, 137–147, 1996.

Hamaker, J. P. and J. D. Bregman, Understanding Radio Polarimetry. III. Interpreting the IAU/IEEE Definitions of the Stokes Parameters, *Astron. Astrophys. Suppl.*, **117**, 161–165, 1996.

IAU, *Trans. Int. Astron. Union*, **15B**, 166, 1973.

IEEE, Standard Definitions of Terms for Radio Wave Propagation, Std. 211–1977, Institute of Electrical and Electronics Engineers, New York, 1977.

James, G. L., The Feed System, *J. Electric. Electron. Eng. Australia*, Special Issue on the Australia Telescope, **12**, No. 2, 137–145, 1992.

Jones, R. C., A New Calculus for the Treatment of Optical Systems. I Description and Discussion of the Calculus, *J. Opt. Soc. Am.*, **31**, 488–493, 1941.

Kemball, A. J., P. J. Diamond, and W. D. Cotton, Data Reduction Techniques for Spectral Line Polarization VLBI Observations *Astron. Astrophys. Suppl.*, **110**, 383–394, 1995.

Ko, H. C., Coherence Theory of Radio-Astronomical Measurements, *IEEE Trans. Antennas Propag.*, **AP-15**, 10–20, 1967a.

Ko, H. C., Theory of Tensor Aperture Synthesis, *IEEE Trans. Antennas Propag.*, **AP-15**, 188–190, 1967b.

Kraus, J. D. and K. R. Carver, *Electromagnetics*, 2nd ed., McGraw-Hill, New York, 1973, p. 435.

Morris, D., V. Radhakrishnan, and G. A. Seielstad, On the Measurement of Polarization Distributions over Radio Sources, *Astrophys. J.*, **139**, 551–559, 1964.

Mueller, H., The Foundations of Optics (abstract only), *J. Opt. Soc. Am.*, **38**, 661, 1948.

O'Neill, E. L. *Introduction to Statistical Optics*, Addison-Wesley, Reading, MA, 1963.

Roberts, D. H., L. F. Brown and J. F. C. Wardle, Linear Polarization Sensitive VLBI, in *Radio Interferometry: Theory, Techniques and Applications*, T. J. Cornwell and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **19**, 281–288, 1991.

Roberts, D. H., J. F. C. Wardle, and L. F. Brown, Linear Polarization Radio Imaging at Milliarcsecond Resolution, *Astrophys. J.*, **427**, 718–744, 1994.

Rohlfs, K. and T. L. Wilson, *Tools of Radio Astronomy*, 2nd ed., Springer-Verlag, Berlin, 1996.

Rowson, B., High Resolution Observations with a Tracking Interferometer, *Mon. Not. R. Astron. Soc.*, **125**, 177–188, 1963.

Sakurai, T. and S. R. Spangler, Use of the Very Large Array for Measurement of Time Variable Faraday Rotation, *Radio Science*, **29**, 635–662, 1994.

Sault, R. J., J. P. Hamaker, and J. D. Bregman, Understanding Radio Polarimetry. III. Instrumental Calibration of an Interferometer Array, *Astron. Astrophys. Suppl.*, **117**, 149–159, 1996.

Sault, R. J., N. E. B. Killeen, and M. J. Kesteven, *AT Polarization Calibration*, Aust. Tel. Tech. Doc. Ser. 39.3/015, CSIRO, Epping, NSW, 1991.

Seidelmann, P. K., Ed., *Explanatory Supplement to the Astronomical Almanac*, University Science Books, Mill Valley, CA, 1992.

Smegal, R. J., T. L. Landecker, J. F. Vaneldik, D. Routledge, and P. E. Dewdney, Aperture Synthesis Polarimetry: Application to the Dominion Astrophysical Observatory Synthesis Telescope, *Radio Sci.*, **32**, 643–656, 1997.

Wade, C. M., Precise Positions of Radio Sources, I. Radio Measurements, *Astron. J.*, **162**, 381–390, 1970.

Wardle, J. F. C., and P. P. Kronberg, Linear Polarization of Quasi-Stellar Radio Sources at 3.71 and 11.1 Centimeters, *Astrophys. J.*, **194**, 249–255, 1974.

Weiler, K. W., The Synthesis Radio Telescope at Westerbork, Methods of Polarization Measurement, *Astron. Astrophys.*, **26**, 403–407, 1973.

Weiler, K. W. and E. Raimond, Aperture Synthesis Observations of Circular Polarization, *Astron. Astrophys.*, **52**, 397–402, 1976.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

5 Antennas and Arrays

This chapter opens with a brief review of some basic considerations of antennas. The main part of the chapter is concerned with the configurations of antennas in interferometers and synthesis arrays. It is convenient to classify array designs as follows:

1. Arrays with non-tracking antennas
2. Interferometers and arrays with antennas that track the sidereal motion of a source:
   - Linear arrays
   - Arrays with open-ended arms (crosses, T-shaped arrays, and Y-shaped arrays)
   - Arrays with closed configurations (circles, ellipses, and Reuleaux triangles)
   - VLBI arrays
   - Planar arrays

Examples of these types of arrays are described and their spatial transfer functions (i.e., spatial sensitivities) are compared. Other concerns include the size and number of antennas needed in an array.

5.1 ANTENNAS

The subject of antennas is well covered in numerous books; see the bibliography at the end of this chapter. Here we mention only a few points concerning the special requirements of antennas for radio astronomy. As discussed in Chapter 1, early radio astronomy antennas operated mainly at meter wavelengths and often consisted of arrays of dipoles or parabolic-cylinder reflectors. These had large areas, but the operating wavelengths were long enough that beamwidths were usually of order 1° or more. For detection and cataloging of sources, satisfactory observations could be obtained during the passage of a source through a stationary beam. Thus it was not generally necessary for such antennas to track the sidereal motion of a source. For a more recent example of a fixed dipole array, see Koles, Frehlich, and Kojima (1994). With the narrower beams at centimeter and millimeter wavelengths, sidereal tracking is essential. Except for a few instruments designed specifically for meter-wavelength operation, the majority of

synthesis arrays use tracking antennas that incorporate equatorial or altazimuth mounts as described in Section 4.6.

The requirement for high sensitivity and high resolution in astronomical images has resulted in the development of large arrays of antennas. To make the fullest use of such instruments, they are usually designed to cover a large range of frequencies. For centimeter-wavelength instruments the coverage typically includes bands extending from a few hundred megahertz to some tens of gigahertz. As a result, the antennas are most often paraboloid or similar-type reflectors, with separate feeds for the different frequency bands. In addition to wide frequency coverage, another advantage of the paraboloid reflector is that all of the power collected is brought, essentially without loss, to a single focus, which allows full advantage to be taken of low-loss feeds and cryogenically cooled input stages to provide the maximum sensitivity.

Figure 5.1 shows several focal arrangements for parabolic antennas, of which the Cassegrain is perhaps the most often used. The Cassegrain focus offers a number of advantages. A convex hyperbolic reflector intercepts the radiation just before it reaches the prime focus and directs it to the Cassegrain focus near the vertex of the main reflector. Sidelobes resulting from spillover of the beam of the feed around the edges of the subreflector point toward the sky, for which the noise temperature is generally low. With a prime-focus feed the sidelobes resulting from spillover around the main reflector point toward the ground and thus result in a higher level of unwanted noise pickup. The Cassegrain focus also has the advantage that in all but the smallest antennas an enclosed room can be provided just behind the main reflector to accommodate the low-noise input stages of the electronics. However, the aperture of the feed for a prime-focus location is less than that for a feed at the Cassegrain focus, and as a result the feeds for the longest wavelengths are often at the prime focus.

The Cassegrain design also allows the possibility of improving the aperture efficiency by shaping the two reflectors of the antenna (Williams 1965). The principle involved can best be envisioned by considering the antenna in transmission.

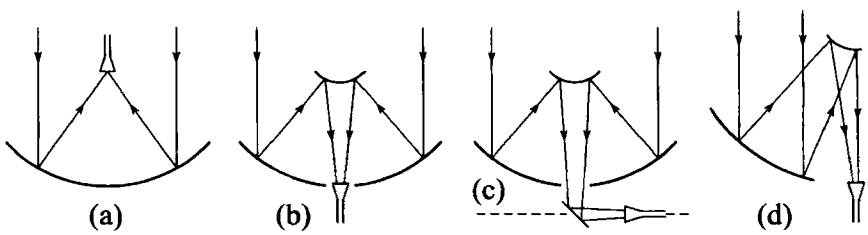


Figure 5.1 Focus arrangements of reflector antennas: (a) prime focus; (b) Cassegrain focus; (c) Naysmith focus; (d) offset Cassegrain. With the Naysmith focus the feed horn is mounted on the alidade structure below the elevation axis (indicated by the dashed line), and for a linearly polarized signal the angle of polarization relative to the feed varies with the elevation angle. In some other arrangements, for example, beam-waveguide antennas (not shown), there are several reflectors, including one on the azimuth axis, which allows the feed horn to remain fixed relative to the ground. The polarization then rotates relative to the feed for both azimuth and elevation motions.

With a conventional hyperboloid subreflector and paraboloid main reflector, the radiation from the feed is concentrated toward the center of the antenna aperture, whereas for maximum efficiency the electric field should be uniformly distributed. If the profile of the subreflector is slightly adjusted, more power can be directed toward the outer part of the main reflector, thus improving the uniformity. The main reflector must then be shaped to depart slightly from the parabolic profile to regain uniform phase across the wavefront after it leaves the main reflector. This type of shaping is used, for example, in the antennas of the VLA in New Mexico, for which the main reflector is 25 m in diameter. For the VLA the rms difference between the reflector surfaces and the best-fit paraboloid is $\sim$1 cm, so the antennas can be used with prime-focus feeds for wavelengths longer than $\sim$16 cm. Shaping is not always to be preferred since it introduces some restriction in off-axis performance, which is detrimental for multi-beam applications.

Although most antennas of synthesis arrays are tracking paraboloid reflectors, there are numerous differences in the detailed design. For example, when a number of feeds for different frequency bands are required at the Cassegrain focus, these are sometimes mounted on a turntable structure, and the one that is in use is brought to a position on the axis of the main reflector. Alternatively, the feeds may be in fixed positions on a circle centered on the vertex, and by using a rotatable subreflector of slightly asymmetric design, the incoming radiation can be focused onto the required feed.

Parabolic reflector antennas with asymmetrical feed geometry can exhibit undesirable instrumental polarization effects that would largely cancel out in a circularly symmetrical antenna. This may occur in an unblocked aperture design, as in Fig. 5.1d, or a design in which a cluster of feeds is used for operation on a number of frequency bands, where the feeds are close to, but not exactly on, the axis of the paraboloid. With crossed linearly polarized feeds the asymmetry results in strong cross-polarization sidelobes within the main beam. With opposite circularly polarized feeds the two beams are offset in opposite directions in a plane that is normal to the plane containing the axis of symmetry of the reflector and the center of the feed. This offset is a serious problem in measurements of circular polarization, since the result is obtained by taking the difference between measurements with opposite circularly polarized antennas [see Eqs. (4.35)]. For measurements of linear polarization the offset is less serious since this involves taking the product of two opposite-hand outputs, and the resulting response is symmetrical about the paraboloid axis. The effects can be largely canceled by inserting a compensating offset in a secondary reflector. For further details, see Chu and Turrin (1973) and Rudge and Adatia (1978).

A basic point concerns the accuracy of the reflector surface. Deviations of the surface from the ideal profile result in variations in the phase of the electromagnetic field as it reaches the focus. We can think of the reflector surface as consisting of many small sections that deviate from the ideal surface by ϵ, a Gaussian random variable with probability distribution

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\epsilon^2/2\sigma^2}, \tag{5.1}$$

where $\langle\epsilon\rangle = 0$, $\langle\epsilon^2\rangle = \sigma^2$, and $\langle\ \rangle$ indicates the expectation. An important relation that can be proved using Eq. (5.1) is

$$\langle e^{j\epsilon}\rangle = \langle\cos\epsilon\rangle + \langle j\sin\epsilon\rangle = \langle\cos\epsilon\rangle = \int_{-\infty}^{\infty}\cos(\epsilon)p(\epsilon)d\epsilon = e^{-\sigma^2/2}. \quad (5.2)$$

A surface deviation ϵ produces a deviation of approximately 2ϵ in the path length of a reflected ray; this approximation improves as the focal ratio is increased. Thus a deviation ϵ causes a phase shift $\phi \simeq 4\pi\epsilon/\lambda$, where λ is the wavelength. As a result, the electric field components at the focus have a Gaussian phase distribution with $\sigma_\phi = 4\pi\sigma/\lambda$. If there are N independent sections of the surface, then the collecting area, which is proportional to the square of the electric field, is given by

$$A = A_0\left\langle\left|\frac{1}{N}\sum_i e^{j\phi_i}\right|^2\right\rangle = \frac{A_0}{N^2}\sum_{i,k}\left\langle e^{j(\phi_i-\phi_k)}\right\rangle \simeq A_0\langle e^{j\phi_i}\rangle^2, \quad (5.3)$$

where A_0 is the collecting area for a perfect surface, and it has been assumed that N is large enough that terms for which $i = k$ can be ignored. Then from Eqs. (5.2) and (5.3) we obtain

$$A = A_0 e^{-(4\pi\sigma/\lambda)^2}. \quad (5.4)$$

This equation is known in radio engineering as the Ruze formula (Ruze 1966), and in some other branches of astronomy as the Strehl ratio. As an example, if $\sigma/\lambda = 1/20$, the aperture efficiency, A/A_0, is 0.67. In the case of antennas with multiple reflecting surfaces, the rms deviations can be combined in the usual root-sum-squared manner. Secondary reflectors, such as a Cassegrain subreflector, are smaller than the main reflector, and for smaller surfaces the rms deviation is usually correspondingly smaller.

Several techniques have been developed for improving the performance of paraboloid antennas. An example is the adjustment of the subreflector shape to compensate for errors in the main reflector [see, e.g., Ingalls et al. (1994), Mayer et al. (1994)]. Another improvement is in the design of the focal support structure to minimize blockage of the aperture and reduce sidelobes in the direction of the ground (Lawrence, Herbig, and Readhead 1994; Welch et al. 1996). The most common method of supporting equipment near the reflector focus is the use of a tripod or quadrupod structure. If the legs of the structure are connected to the edge of the main reflector rather than to points within the reflector aperture, they interrupt only the plane wave incident on the aperture, not the spherical wavefront between the reflector and the focus. Use of an offset-feed reflector avoids any blockage of incident wavefront in reaching the focus. However, both these methods of reducing blockage increase the difficulty in obtaining mechanical stiffness in the structure. For this reason they are more commonly used on small-diameter antennas, such as those designed for use at millimeter wavelengths.

5.2 SAMPLING THE VISIBILITY FUNCTION

Sampling Theorem

The choice of configuration of the antennas of a synthesis array is based on optimizing, in some manner, the sampling of the visibility function in (u, v) space. Thus in considering array design it is logical to start by examining the sampling requirements. These are governed by a sampling theorem of Fourier transforms (Bracewell 1958). Consider first the measurement of the one-dimensional intensity distribution of a source, $I_1(l)$. It is necessary to measure the complex visibility $\mathcal{V}$ in the corresponding direction on the ground at a series of values of the projected antenna spacing. For example, to measure an east–west profile, a possible method is to make observations near meridian transit of the source using an east–west baseline, and to vary the length of the baseline from day to day.

Figure 5.2a–c illustrates the sampling of the one-dimensional visibility function $\mathcal{V}(u)$. The sampling operation can be represented as multiplication of $\mathcal{V}(u)$ by the series of delta functions in Fig. 5.2b, which can be written

$$\left[\frac{1}{\Delta u}\right] \text{III}\left(\frac{u}{\Delta u}\right) = \sum_{i=-\infty}^{\infty} \delta(u - i\Delta u), \tag{5.5}$$

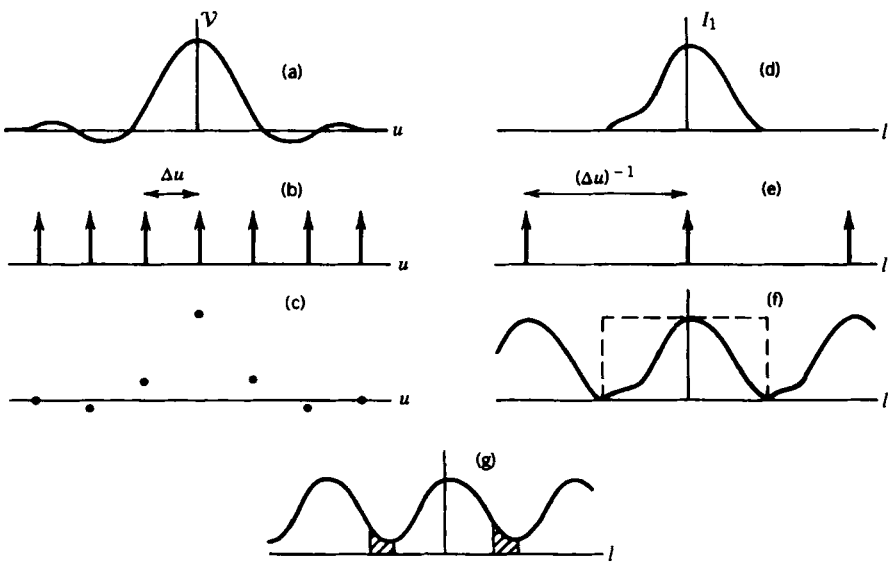


Figure 5.2 Illustration of the sampling theorem: (a) visibility function $\mathcal{V}(u)$, real part only; (b) sampling function in which the arrows represent delta functions; (c) sampled visibility function; (d) intensity function $I_1(l)$; (e) replication function; (f) replicated intensity function. Functions in (d), (e), and (f) are the Fourier transforms of those in (a), (b), and (c), respectively. (g) Replicated intensity function showing aliasing in shaded areas resulting from use of too large a sampling interval.

where the left-hand side is included to show how the series can be expressed in terms of the *shah function*, III, introduced by Bracewell and Roberts (1954). The series extends to infinity in both directions, and the delta functions are uniformly spaced with an interval Δu. The Fourier transform of Eq. (5.5) is the series of delta functions shown in Fig. 5.2e:

$$\mathrm{III}(l\Delta u) = \frac{1}{\Delta u}\sum_{p=-\infty}^{\infty}\delta\left(l - \frac{p}{\Delta u}\right). \tag{5.6}$$

In the l domain the Fourier transform of the sampled visibility is the convolution of the Fourier transform of $\mathcal{V}(u)$, which is the one-dimensional intensity $I_1(l)$, with Eq. (5.6). The result is the replication of $I_1(l)$ at intervals $(\Delta u)^{-1}$ shown in Fig. 5.2f. If $I_1(l)$ represents a source of finite dimensions, the replications of $I_1(l)$ will not overlap as long as $I_1(l)$ is nonzero only within a range of l that is no greater than $(\Delta u)^{-1}$. An example of overlapping replications is shown in Fig. 5.2g. The loss of information resulting from such overlapping is commonly referred to as *aliasing*, because the components of the function within the overlapping region lose their identity with respect to which end of the replicated function they properly belong. Avoidance of aliasing requires that the sampling interval Δu be no greater than the reciprocal of the interval in l within which $I_1(l)$ is nonzero. To be precise, we should consider the width of the source as broadened by the finite resolution of the observations, rather than the true width of the source, but this is usually only a minor effect: see discussion of *leakage** by Bracewell (2000).

The requirement for the restoration of a function from a set of samples, for example, deriving the function in Fig. 5.2a from the samples in 5.2c, is easily understood by considering the Fourier transforms in Fig. 5.2d and f. Interpolation in the u domain corresponds to removing the replications in the l domain, which can be achieved by multiplication of the function in Fig. 5.2f by the rectangular function indicated by the broken line. In the u domain this multiplication corresponds to convolution of the sampled values with the Fourier transform of the rectangular function, which is the unit area sinc function

$$\frac{\sin \pi u/\Delta u}{\pi u}. \tag{5.7}$$

If aliasing is avoided, convolution with (5.7) provides exact interpolation of the original function from the samples. Thus we can state, as a sampling theorem for the visibility, that *if the intensity distribution is nonzero only within an interval of width l_w, $I_1(l)$ is fully specified by sampling the visibility function at points spaced $\Delta u = l_w^{-1}$ in u*. In two dimensions, it is simply necessary to apply the theorem separately to the source in the l and m directions. For further discussion of the sampling theorem, see, for example, Unser (2000).

*Here the usage of the term "leakage" is different from that related to polarimetry.

Discrete Two-Dimensional Fourier Transform

The derivation of a map or image from the visibility measurements is the subject of Chapter 10, but it is important at this point to understand the form in which the visibility data are required for this transformation. The discrete form of the Fourier transform is very widely used in synthesis mapping because of the computational advantages of the fast Fourier transform (FFT) algorithm [see, e.g., Brigham (1988)]. The fast Hartley transform (Bracewell 1984) can also be used. With the discrete transform the functions $\mathcal{V}(u, v)$ and $I(l, m)$ are expressed as rectangular matrices of sampled values at uniform increments in the two variables involved. The rectangular grid points on which the intensity is obtained provide a convenient form for further data processing.

The two-dimensional form of the discrete transform for a Fourier pair f and g is defined by

$$f(p, q) = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{k=0}^{N-1} g(i, k) e^{-j2\pi ip/M} e^{-j2\pi kq/N}, \tag{5.8}$$

and the inverse is

$$g(i, k) = \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} f(p, q) e^{j2\pi ip/M} e^{j2\pi kq/N}. \tag{5.9}$$

See, for example, Oppenheim and Schafer (1975). The functions are periodic with periods of M samples in the i and p dimensions and N samples in the k and q dimensions. Evaluation of Eqs. (5.8) or (5.9) by direct computation requires approximately $(MN)^2$ complex multiplications. In contrast, if M and N are powers of 2 the FFT algorithm requires only $\frac{1}{2}MN \log_2(MN)$ complex multiplications.

The transformation between $\mathcal{V}(u, v)$ and $I(l, m)$, where I is the source intensity in two dimensions, is obtained by substituting $g(i, k) = I(i\Delta l, k\Delta m)$ and $f(p, q) = \mathcal{V}(p\Delta u, q\Delta v)$ in Eqs. (5.8) and (5.9). The relationship between the integral and discrete forms of the Fourier transform is found in several texts; see, for example, Rabiner and Gold (1975) or Papoulis (1977). The dimensions of the (u, v) plane that contain these data are $M\Delta u$ by $N\Delta v$. In the (l, m) plane the points are spaced Δl in l and Δm in m, and the map dimensions are $M\Delta l$ by $N\Delta m$. The dimensions in the two domains are related by

$$\begin{aligned} \Delta u &= (M\Delta l)^{-1}, \quad & \Delta v &= (N\Delta m)^{-1} \\ \Delta l &= (M\Delta u)^{-1}, \quad & \Delta m &= (N\Delta v)^{-1}. \end{aligned} \tag{5.10}$$

The spacing between points in one domain is the reciprocal of the total dimension in the other domain. Thus, if the size of the array in the intensity domain is chosen to be large enough that the intensity function is nonzero only within the area $M\Delta l \times N\Delta m$, then the spacings Δu and Δv in Eq. (5.10) satisfy the sampling theorem.

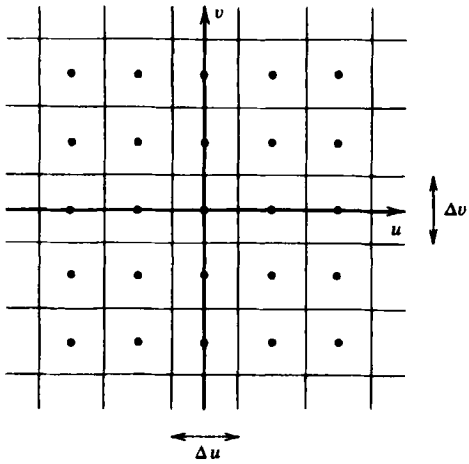


Figure 5.3 Points on a rectangular grid in the (u, v) plane at which the visibility is sampled for use with the discrete Fourier transform. As shown, the spacings Δu and Δv are equal. The division of the plane into grid cells of size $\Delta u \times \Delta v$ is also shown.

To apply the discrete transform to the synthesis mapping problem it is necessary to obtain values of $\mathcal{V}(u, v)$ at points separated by Δu in u and by Δv in v, as shown in Fig. 5.3. However, the measurements are generally not made at (u, v) points on a grid since for tracking interferometers they fall on elliptical loci in the (u, v) plane, as explained in Section 4.1. Thus it is necessary to obtain the values at the grid points by interpolation or similar processes. In Fig. 5.3 the plane is divided into cells of size $\Delta u \times \Delta v$ centered on the grid points. A very simple method of determining a visibility value to assign at each grid point is to take the mean of all values that fall within the same cell. This procedure has been termed *cell averaging* (Thompson and Bracewell 1974). Better procedures are generally used; see Section 10.2 under *Weighting of the Visibility Data*. However, the cell averaging concept helps one to visualize the required distribution of the measurements; ideally there should be at least one measurement, or a small number of measurements, within each cell. Thus the baselines should be chosen so that the spacings between the (u, v) loci are no greater than the cell size, to maximize the number of cells that are intersected by a locus. Cells that contain no measurements result in holes in the (u, v) coverage, and minimization of such holes is an important criterion in array design.

5.3 INTRODUCTORY DISCUSSION OF ARRAYS

Phased Arrays and Correlator Arrays

An array of antennas can be interconnected to operate as a phased array or as a correlator array. Phased arrays were used for early solar observations, as in the system in Fig. 1.13a, and phased arrays of small antennas can be used as single

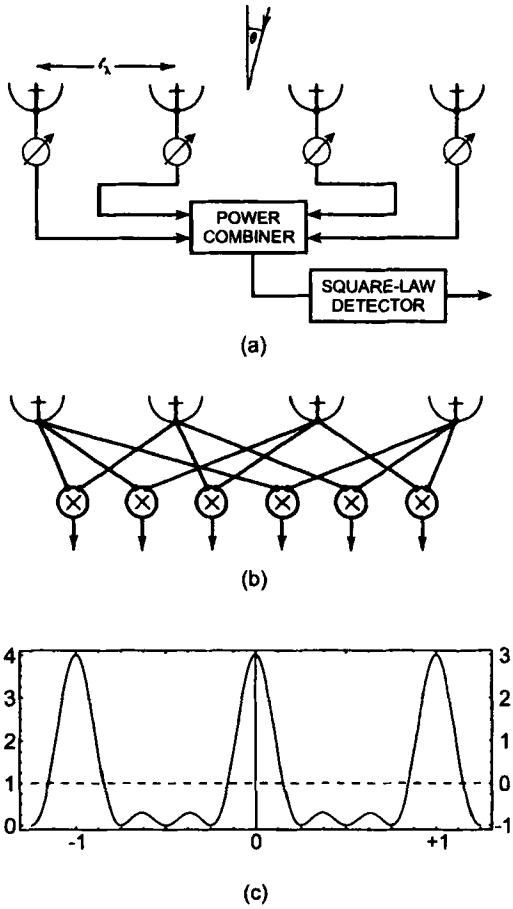


Figure 5.4 Simple four-element linear array. ℓ_λ is the unit antenna spacing measured in wavelengths, and θ indicates the angle of incidence of a signal. (a) Connected as a phased array with an adjustable phase shifter in the output of each antenna, and the combined signal applied to a square-law detector. The power combiner is a matching network in which the output is proportional to the sum of the radio-frequency input voltages. (b) The same antennas connected as a correlator array. (c) The ordinate is the response of the array: the scale at the left applies to the phased array, and at the right to the correlator array. The abscissa is proportional to θ in units of ℓ_λ^{-1} rad. The equal spacing between antennas in this simple grating array gives rise to sidelobes in the form of replications of the central beam.

elements in correlator arrays. Figure 5.4a shows a simple schematic diagram of a phased array feeding a square-law detector, in which the number of antennas, n_a, is equal to four. If the voltages at the antenna outputs are V_1, V_2, V_3, and so on, the output of the square-law detector is proportional to

$$(V_1 + V_2 + V_3 + \cdots + V_{n_a})^2. \tag{5.11}$$

Note that for n_a antennas there are $n_a(n_a - 1)$ cross-product terms of form $V_m V_n$ involving different antennas m and n, and n_a self-product terms of form V_m^2. If

the signal path (including the phase shifter) from each antenna to the detector is of the same electrical length, the signals combine in phase when the direction of the incoming radiation is given by

$$\theta = \sin^{-1}\left(\frac{N}{\ell_\lambda}\right), \tag{5.12}$$

where N is an integer, including zero, and ℓ_λ is the spacing interval measured in wavelengths. The position angles of the maxima, which represent the beam pattern of the array, can be varied by adjusting the phase shifters at the antenna outputs, and the beam pattern can be controlled and, for example, scanned to map an area of sky.

In the correlator array in Fig. 5.4b, a correlator generates the cross-product of signal voltages $V_m V_n$ for every antenna pair. The correlator outputs are equal to the cross-product terms of the phased array. These outputs can be combined to produce maxima similar to those of the phased array. If a phase shift is introduced at the output of one of the correlator array antennas, the result appears as a corresponding change in the phase of the fringes measured with the correlators connected to that antenna. Conversely, the effect of an antenna phase shift can be simulated by changing the measured phases when combining the correlator outputs. Thus a beam-scanning action can be accomplished by combining measured cross-correlations in a computer with appropriate variations in the phase. This is what happens in computing the Fourier transform of the visibility function, that is, the Fourier transform of the correlator outputs as a function of spacing. The loss of the self-product terms reduces the instantaneous sensitivity of the correlator array by a factor $(n_a - 1)/n_a$ in power, which is close to unity if n_a is large. However, at any instant, the correlator array responds to the whole field of the individual antennas, whereas the response of the phased array is determined by the narrow beam that it forms, unless it is equipped with a more complex signal-combining network that allows many beams to be formed simultaneously. Thus in mapping, the correlator array gathers data more efficiently than the phased array.

The response pattern of the correlator array to a point source is the same as that of the phased array, except for the self-product terms. The response of the phased array consists of one or more beams in the direction in which the antenna responses combine with equal phase. These are surrounded by sidelobes, the pattern and magnitude of which depend on the number and configuration of antennas. Between individual sidelobe peaks there will be nulls that can be as low as zero, but the response is never negative because the output of the square-law detector cannot go negative. Now consider subtracting the self-product terms, to simulate the response of the correlator array. Over a field of view small compared with the beamwidth of an individual antenna, each self-product term represents a constant level, and each cross-product represents a fringe oscillation. In the response to a point source, all of these terms are of equal magnitude. Subtracting the

self-products from the phased-array response causes the zero level to be shifted in the positive direction by an amount equal to $1/n_a$ of the peak level, as indicated by the broken line in Fig. 5.4c. The points that represent zeros in the phased-array response become the peaks of negative sidelobes. Thus in the response of the correlator array the positive values are decreased by a factor $(n_a - 1)/n_a$ relative to those of the phased array. In the negative direction, the response extends to a level of $-1/(n_a - 1)$ of the positive peak, but no further since this level corresponds to the zero level of the phased array. Kogan (1999) has pointed out this limitation on the magnitude of the negative sidelobes of a correlator array, and has also noted that this limit does not depend on the configuration of the individual antennas, but only on their number. Neither of these conclusions apply to the positive sidelobes. This result is strictly true only for snapshot observations [i.e., those in which the (u, v) coverage is not significantly increased by earth rotation], and for uniform weighting of the correlator outputs.

Finally, consider some characteristics of a phased array as in Fig. 5.4a. The power combiner is a passive network, for example, the branched transmission line in Fig. 1.13a. If a *correlated* waveform of power P is applied to each combiner input, then the output power is $n_a P$. In terms of the voltage V at each input, a fraction $1/\sqrt{n_a}$ of each *voltage* combines additively to produce an output of $\sqrt{n_a}V$, or $n_a P$ in power. Now if the input waveforms are *uncorrelated*, again each contributes $V/\sqrt{n_a}$ in voltage but the resulting *powers* combine additively (i.e., as the sum of the squared voltages), so in this case the power at the output is equal to the power P at one input. Each input then contributes only $1/n_a$ of its power to the output, and the remaining power is dissipated in the terminating impedances of the combiner inputs (i.e., radiated from the antennas if they are directly connected to the combiner). The signals from an unresolved source received in the main beam of the array are fully correlated, but the noise contributions from amplifiers at the antennas are uncorrelated. Thus, if there are no losses in the transmission lines or the combiner, the same signal-to-noise ratio at the detector is obtained by inserting an amplifier at the output of each antenna, or a single amplifier at the output of the combiner. However, such losses are often significant, and it is advantageous to use amplifiers at the antennas. Note that, if half the antennas in a phased array are pointed at a radio source and the others at blank sky, the signal power at the combiner output is one quarter of that with all antennas pointed at the source.

Spatial Sensitivity and the Spatial Transfer Function

We now consider the sensitivity of an antenna or array to the spatial frequencies on the sky. The angular response pattern of an antenna is the same in reception or transmission, and at this point it may be easier to consider the antenna in transmission. Then power applied to the terminals produces a field at the antenna aperture. A function $W(u, v)$ is equal to the autocorrelation function of $\mathcal{E}(x_\lambda, y_\lambda)$, the distribution of the electric field across the aperture, where x_λ, and y_λ are coordinates in the aperture plane of the antenna and are measured in wavelengths. Thus,

$$
\begin{aligned}
W(u, v) &= \mathcal{E}(x_\lambda, y_\lambda) \star\star \mathcal{E}^*(x_\lambda, y_\lambda) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{E}(x_\lambda, y_\lambda) \mathcal{E}^*(x_\lambda - u, y_\lambda - v)\, dx_\lambda\, dy_\lambda.
\end{aligned}
\tag{5.13}
$$

The double-pentagram symbol represents two-dimensional autocorrelation. The integral in Eq. (5.13) is proportional to the number of ways, suitably weighted by the field intensity, in which a specific spacing vector (u, v) can be found within the antenna aperture. In reception, $W(u, v)$ is a measure of the sensitivity of the antenna to different spatial frequencies. In effect, the antenna or array acts as a spatial frequency filter, and $W(u, v)$ is widely referred to as the *transfer function* by analogy with the usage of this term in filter theory. $W(u, v)$ has also been called the spectral sensitivity function (Bracewell 1961, 1962), which refers to the spectrum of spatial frequencies (not the radio frequencies) to which the array responds. We use the terms *spatial transfer function* and *spatial sensitivity* when discussing $W(u, v)$. The area of the (u, v) plane over which measurements can be made [i.e., the support of $W(u, v)$, defined as the closure of the domain within which $W(u, v)$ is nonzero] is referred to as the *spatial frequency coverage*, or the (u, v) coverage.

Consider the response of the antenna or array to a point source. Since the visibility of a point source is constant over the (u, v) plane, the measured spatial frequencies are proportional to $W(u, v)$. Thus the point source response $\mathcal{A}(l, m)$ is the Fourier transform of $W(u, v)$. This result is formally derived by Bracewell and Roberts (1954). [Recall from Eq. (2.15) that the point-source response is the mirror image of the antenna power pattern, $\mathcal{A}(l, m) = A(-l, -m)$, but this distinction is seldom of practical importance since the functions concerned are usually symmetrical.] The spatial transfer function $W(u, v)$ is an important feature in this chapter, and Fig. 5.5 further illustrates its place in the interrelationships between functions involved in radio imaging.

Figure 5.6a shows an interferometer in which the antennas do not track and are represented by two rectangular areas. We shall assume that $\mathcal{E}(x_\lambda, y_\lambda)$ is uniformly distributed over the apertures, such as in the case of arrays of uniformly excited dipoles. First suppose that the output voltages from the two apertures are summed and fed to a power-measuring receiver, as in some early instruments. The three rectangular areas in Fig. 5.6b represent the autocorrelation function of the aperture distributions, that is, the spatial transfer function. Note that the autocorrelation of the two apertures contains the autocorrelation of the individual apertures (the central rectangle in Fig. 5.6b) plus the cross-correlation of the two apertures (the shaded rectangles). If the two antennas are combined using a correlator instead of a receiver that responds to the total received power, the spatial sensitivity is represented only by the shaded rectangles since the correlator forms only the cross-products of signals from the two apertures. Thus the spatial transfer function $W(u, v)$ may not include all parts of the autocorrelation function of the aperture, depending on the interconnection of the correlators and/or detectors.

The interpretation of the spatial transfer function as the Fourier transform of the point-source response can be applied to both the adding and correlator cases.

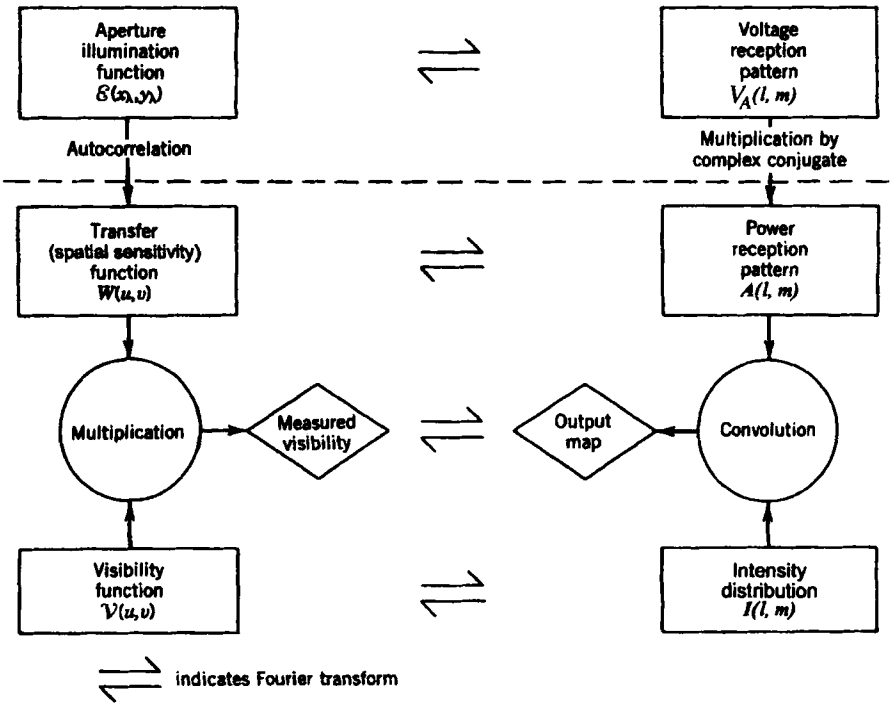


Figure 5.5 Relationships between functions involved in mapping a source. Starting at the top left, the autocorrelation of the aperture distribution of the electric field over an antenna $\mathcal{E}(x_\lambda, y_\lambda)$ gives the spatial transfer function $W(u, v)$. The measured visibility in the observation of a source is the product of the source visibility $\mathcal{V}(u, v)$ and the spatial transfer function. At the top right, the multiplication of the voltage reception pattern $V_A(l, m)$ with its complex conjugate produces the power reception pattern, $A(l, m)$. Mapping of the source intensity distribution $I(l, m)$ results in convolution of this function with the antenna power pattern. The Fourier transform relationships between the quantities in the (x_λ, y_λ) and (u, v) domains, and those in the (l, m) domain, are indicated. When the spatial sensitivity is built up by earth rotation, as in tracking arrays, it cannot, in general, be described as the autocorrelation function of any field distribution. Only the part of the diagram below the broken line applies in such cases.

For example, for the correlator implementation of the interferometer in Fig. 5.6a, the response to a point source is the Fourier transform of the function represented by the shaded areas. This Fourier transform is

$$\left[\frac{\sin \pi x_{\lambda 1} l}{\pi x_{\lambda 1} l}\right]^2 \left[\frac{\sin \pi y_{\lambda 1} m}{\pi y_{\lambda 1} m}\right]^2 \cos 2\pi D_\lambda l, \tag{5.14}$$

where $x_{\lambda 1}$ and $y_{\lambda 1}$ are the aperture dimensions and D_λ is the aperture separation, all measured in wavelengths. The sinc-squared functions represent the power pattern of the uniformly illuminated rectangular apertures, and the cosine term repre-

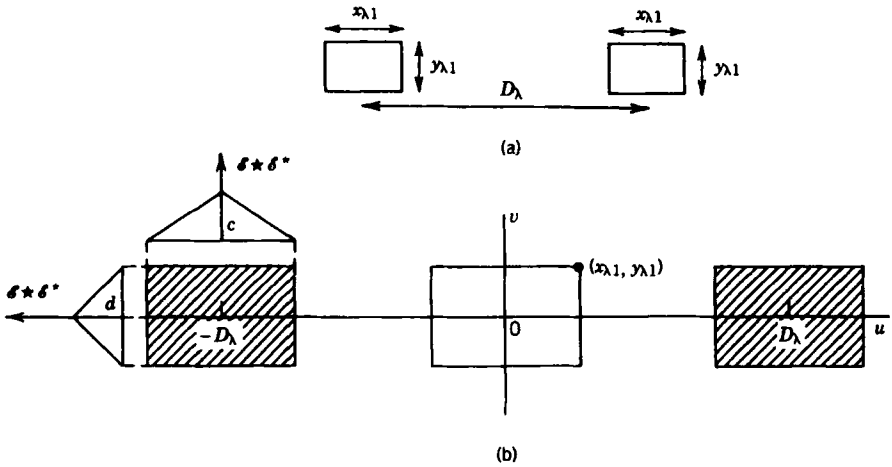


Figure 5.6 The two apertures in (a) represent a two-element interferometer, the spatial transfer function of which is shown in (b). The shaded areas contain the spatial sensitivity components that result from the cross-correlation of the signals from the two antennas. If the field distribution is uniform over the apertures, the magnitude of the spatial sensitivity is linearly tapered. This is indicated by c and d, which represent cross sections of the spatial transfer function.

sents the fringe pattern. In early instruments the relative magnitude of the spatial sensitivity was controlled only by the field distribution over the antennas, but image processing by computer enables the magnitude to be adjusted after an observation has been made.

Some commonly used configurations of antenna arrays, and the boundaries of their autocorrelation functions, are shown in Fig. 5.7. The autocorrelation functions indicate the instantaneous spatial sensitivity for a continuous aperture in the form of the corresponding figure. Equation (5.13) shows that the autocorrelation function is the integral of the product of the field distribution with its complex conjugate displaced by u and v. By investigating the values of u and v for which the two aperture figures overlap, it is easy to determine the boundary within which the spatial transfer function is nonzero using graphical procedures described by Bracewell (1961, 1995). It is also possible to identify ridges of high autocorrelation that occur for displacements at which the arms of figures such as those in Fig. 5.7a, b, or c are aligned. In the case of the ring, Fig. 5.7g, the autocorrelation function is proportional to the area of overlap at the two points where the ring intersects with its displaced replication. This area is approximately proportional to the reciprocal of the sine of the angle between the tangents to the rings at an intersection point, and is shown by the curve in Fig. 5.7j, in which the abscissa runs from the center to the edge of the autocorrelation circle. There is a broad minimum in the spatial sensitivity when the angle of the tangents is $\pi/2$, which, for a ring of unit radius, occurs at $\sqrt{u^2 + v^2} = \sqrt{2}$. When the aperture is not completely filled, that is, when the figure represents an array of

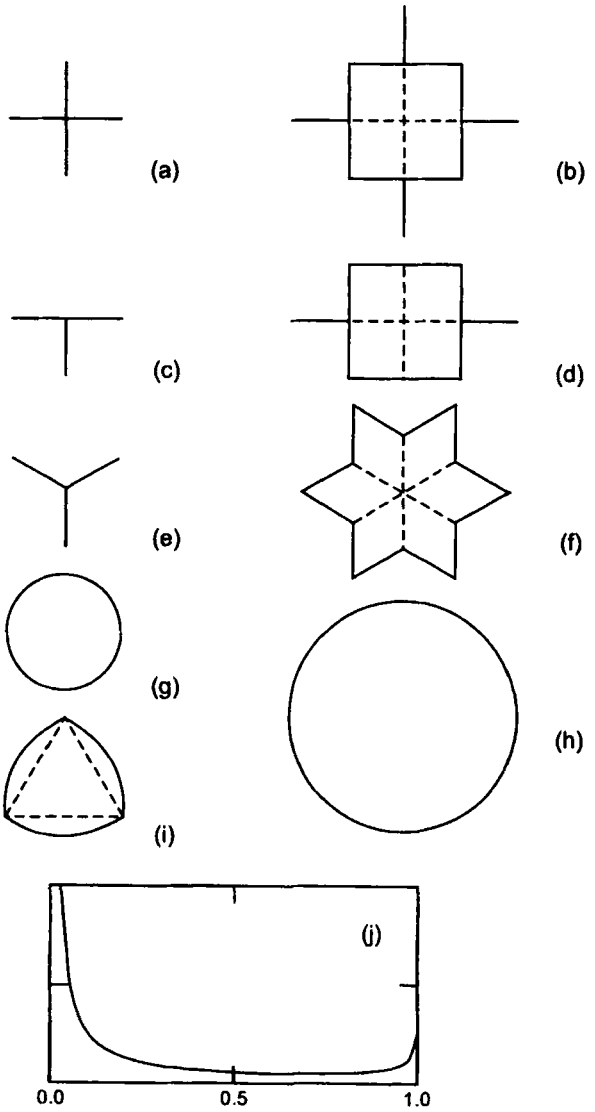


Figure 5.7 Configurations for array apertures and the boundaries within which the corresponding autocorrelation functions are nonzero. The configurations represent the aperture (x_λ, y_λ) plane and the autocorrelations, the spatial frequency (u, v) plane. (a) The cross and (b) its autocorrelation boundary. (c) The T-array and (d) its autocorrelation boundary. (e) The equiangular Y and (f) its autocorrelation boundary. The broken lines in (b), (d), and (f) indicate ridges of high autocorrelation value. (g) The ring and (h) its autocorrelation boundary. The autocorrelation function of the ring is circularly symmetrical and (j) shows the radial profile of the function from the center to the edge of the circle in (h). (i) The Reuleaux triangle. The broken lines indicate an equilateral triangle, and the circular arcs that form the Reuleaux triangle have radii centered on the vertices of the triangle. The autocorrelation of the Reuleaux triangle is bounded by the same circle shown in (h).

discrete antennas, the spatial sensitivity takes the form of samples of the autocorrelation function. For example, for a cross of uniformly spaced antennas, the square in Fig. 5.7b would be represented by a matrix pattern within the square boundary. The configurations shown in Fig. 5.7 are discussed in more detail later.

Meter-Wavelength Cross and T Arrays

A cross and its autocorrelation function are shown in Fig. 5.7a and b. It is assumed that the width of the arms is finite but small compared with the length of the arms. In the case of the Mills cross (Mills 1963) described briefly in Chapter 1, the outputs of the two arms go to a single cross-correlating receiver, so the spatial sensitivity is represented by the square in Fig. 5.7b. The narrow extensions on the centers of the sides of the square represent parts of the autocorrelation functions of the individual arms, which are not formed in the cross-correlation of the arms. However, they are formed if the arms consist of lines of individual antennas for which the cross-correlation is formed for pairs on the same arm as well as those on crossed arms. The case for a T-shaped array is similar and is shown in Fig. 5.7c and d. Again, if only the cross-correlation between the east–west arm and the half-length, north–south arm is formed, then the spatial frequency coverage is represented by the square component of the autocorrelation. The equivalence between the spatial transfer function of such a cross and a T can be understood by noting that for any pair of points in the aperture of a cross, for example, one on the east arm and one on the north arm, there is a corresponding pair on the west and south arms for which the spacing vector is identical. Thus any one of the four half-length arms can be removed without reducing the (u, v) coverage of the spatial transfer function.

If the sensitivity (i.e., the collecting area per unit length) is uniform along the arms for a cross or a corresponding T, then the weighting of the spatial sensitivity is uniform over the square (u, v) area; note that it does not taper linearly from the center as in the example in Fig. 5.6. At the edge of the square area the spatial sensitivity falls to zero in a distance equal to the width of the arms. Such a sharp edge, resulting from the uniform sensitivity, results in strong sidelobes. Therefore an important feature of the Mills cross design was a Gaussian taper of the coupling of the elements along the arms to reduce the sensitivity to about 10% at the ends. This greatly reduced local maxima in the response resulting from sidelobes outside the main beam, at the expense of some broadening of the beam.

Figure 1.12a shows an implementation of a T array that is an example of a non-tracking correlator interferometer in which a small antenna is moved in steps, with continuous coverage, to simulate a larger aperture; see Blythe (1957), Ryle, Hewish, and Shakeshaft (1959), and Ryle and Hewish (1960). The spatial frequency coverage is the same as would be obtained in a single observation with an antenna of aperture equal to that simulated by the movement of the small antenna, although the magnitude of the spatial sensitivity is not exactly the same. The term *aperture synthesis* was introduced to describe such observations.

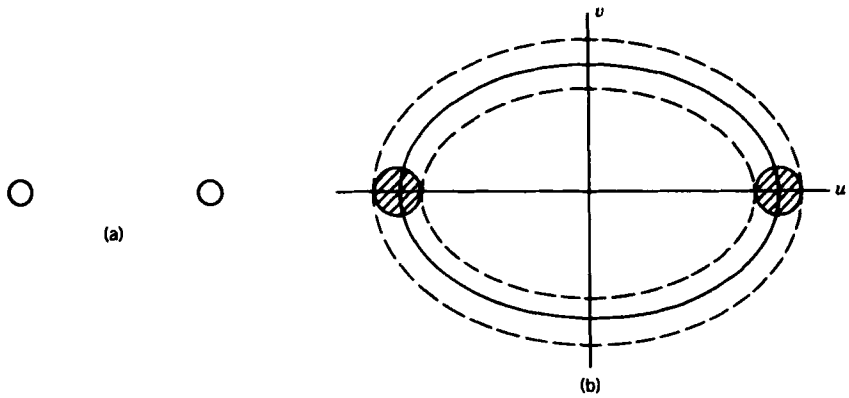


Figure 5.8 (a) The aperture of an east–west, two-element interferometer. The corresponding spatial frequency coverage for cross-correlated signals is shown by the shaded areas in (b). If the antennas track the source, the spacing vector traces out an elliptical locus (the solid line) in the (u, v) plane. The area between the broken lines in (b) indicates the spatial frequencies that contribute to the measured values. The spacing between the broken lines is determined by the cross-correlation of the antenna apertures.

5.4 SPATIAL TRANSFER FUNCTION OF A TRACKING ARRAY

The range of spatial frequencies that contribute to the output of an interferometer with tracking antennas is illustrated in Fig. 5.8b. The two shaded areas represent the cross-correlation of the two apertures of an east–west interferometer for a source on the meridian. As the source moves in hour angle, the changing (u, v) coverage is represented by a band centered on the spacing locus of the two antennas. Recall from Section 4.1 that the locus for an earth-based interferometer is an arc of an ellipse, and that since $\mathcal{V}(-u, -v) = \mathcal{V}^*(u, v)$, any pair of antennas measures visibility along two arcs symmetric about the (u, v) origin, both of which are included in the spatial transfer function.

Because the antennas track the source, the antenna beams remain centered on the same point in the source under investigation, and the array measures the product of the source intensity distribution and the antenna pattern. Another view of this effect is obtained by considering the radiation received by small areas of the apertures of two antennas, the centers of which are A_1 and A_2 in Fig. 5.9. The antenna apertures encompass a range of spacings from $u - d_\lambda$ to $u + d_\lambda$ wavelengths, where d_λ is the antenna diameter measured in wavelengths. If the antenna beams remain fixed in position as a source moves through them, then the correlator output is a combination of fringe components with frequencies from $\omega_e(u - d_\lambda)\cos\delta$ to $\omega_e(u + d_\lambda)\cos\delta$, where ω_e is the angular velocity of the earth and δ is the declination of the source. To examine the effect when the antennas track the source, consider the point B which, because of the tracking, has a component of motion toward the source equal to $\omega_e \Delta u \cos\delta$ wavelengths per second. This causes a corresponding Doppler shift in the signal re-

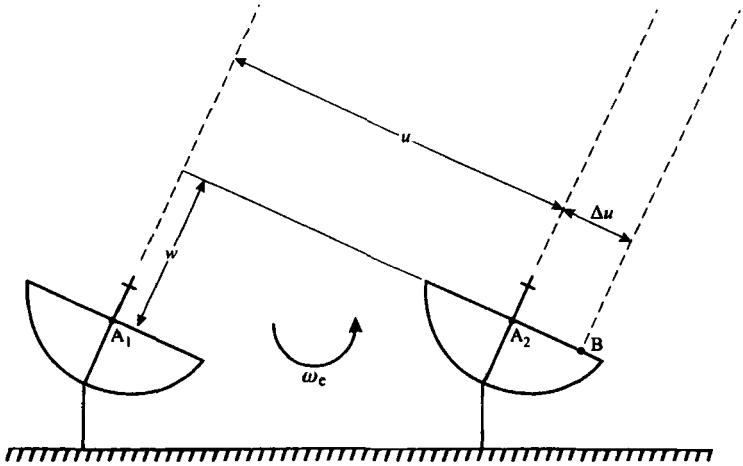


Figure 5.9 Illustration of the effect of tracking on the fringe frequency at the correlator output. The u component of the baseline is shown, and the v component is omitted since it does not affect the fringe frequency. The curved arrow indicates the tracking motion of the antennas.

ceived at B. To obtain the fringe frequency for waves arriving at A_1 and B, we subtract the Doppler shift from the non-tracking fringe frequency and obtain $[\omega_e(u + \Delta u)\cos\delta] - (\omega_e \Delta u \cos\delta) = (\omega_e u \cos\delta)$. The fringe frequency when tracking is thus the same as for the central points A_1 and A_2 of the apertures. (This is true for any pair of points; choosing one point at an antenna center in the example above slightly simplifies the discussion.) Thus if the antennas track, the contributions from all pairs of points within the apertures appear at the same fringe frequency at the correlator output. As a result, such contributions cannot be separated by Fourier analysis of the correlator output waveform and information on how the visibility varies over the range $u - d_\lambda$ to $u + d_\lambda$ is lost. However, if the antenna motion differs from a purely tracking one, the information is, in principle, recoverable. In imaging sources wider than the antenna beams, an additional scanning motion to cover the source is added to the tracking motion. In effect, this scanning allows the visibility to be sampled at intervals in u and v that are fine enough for the extended width of the source. This technique, known as *mosaicking*, is described in Section 11.6.

To accommodate the effects that result when the antennas track the source position, the normalized antenna pattern is treated as a modification to the intensity distribution. The intensity distribution then becomes $A_N(l, m)I(l, m)$, as explained in Section 3.1. The spatial transfer function $W(u, v)$ for a pair of tracking antennas is represented at any instant by a pair of two-dimensional delta functions ${}^2\delta(u, v)$ and ${}^2\delta(-u, -v)$. For an array of antennas the resulting spatial transfer function is represented by a series of delta functions weighted in proportion to the magnitude of the instrumental response. As the earth rotates, these delta functions generate the ensemble of elliptical spacing loci. The loci represent the spatial transfer function of a tracking array.

Consider observation of a source $I(l, m)$, for which the visibility function is $\mathcal{V}(u, v)$, with normalized antenna patterns $A_N(l, m)$. Then if $W(u, v)$ is the spatial transfer function, the measured visibility is

$$[\mathcal{V}(u, v) ** \overline{A}_N(u, v)]W(u, v), \tag{5.15}$$

where the double asterisk indicates two-dimensional convolution and the bar denotes the Fourier transform. The Fourier transform of (5.15) gives the measured intensity:

$$[I(l, m)A_N(l, m)] ** \overline{W}(l, m). \tag{5.16}$$

If we observe a point source at the (l, m) origin, where $A_N = 1$, expression (5.16) becomes the point-source response $b_0(l, m)$. We then obtain

$$b_0(l, m) = [{}^2\delta(l, m)A_N(l, m)] ** \overline{W}(l, m) = \overline{W}(l, m), \tag{5.17}$$

where ${}^2\delta(l, m)$ represents the point source. Here again, the point-source response is the Fourier transform of the spatial transfer function. In the tracking case the spatial frequencies that contribute to the measurement are represented by $W(u, v) ** \overline{A}_N(u, v)$. Note that $\overline{A}_N(u, v)$ is twice as wide as the corresponding antenna aperture in the (x, y) domain.

The term *aperture synthesis* is sometimes extended to include observations that involve hour-angle tracking. However, it is not possible to define an exactly equivalent antenna aperture for a tracking array. For example, consider the case of two antennas with an east–west baseline tracking a source for a period of 12 h. The spatial transfer function is an ellipse centered on the origin of the (u, v) plane, with zero sensitivity within the ellipse (except for a point at the origin that could be supplied by a measurement of total power received in an antenna). The equivalent aperture would be a function, the autocorrelation of which is the same elliptical ring as the spatial transfer function. No such aperture function exists, and thus the term aperture synthesis can only loosely be applied to describe most observations that include hour-angle tracking.

Desirable Characteristics of the Spatial Transfer Function

As a first step in considering the layout of the antennas it is useful to consider the desired spatial (u, v) coverage [see, e.g., Keto (1997)]. For any specific observation, the optimum (u, v) coverage clearly depends on the expected intensity distribution of the source under study, since one would prefer to concentrate the capacity of the instrument in (u, v) regions where the visibility is nonzero. However, most large arrays are used for a wide range of astronomical objects, so some compromise approach is required. Since, in general, astronomical objects are aligned at random in the sky, there is no preferred direction for the highest resolution. Thus it is logical to aim for visibility measurements that extend over a circular area centered on the (u, v) origin.

As described in Section 5.2, the visibility data are usually interpolated onto a rectangular grid for convenience in Fourier transformation, and if approximately equal numbers of measurements are used for each grid point, they can be given equal weights in the transformation. Uneven weighting results in loss of sensitivity, since some values then contain a larger component of noise than others. From this viewpoint one would like the natural weighting (i.e., the weighting of the measurements that results from the array configuration without further adjustment) to be as uniform as possible within the circular area.

For a general-purpose array it is difficult to improve on the circularity of the measurement area. However, there are exceptions to the uniformity of the measurements within the circle. As mentioned above, in the Mills cross uniform coupling of the radiating elements along the arms would result in uniform spatial sensitivity. To reduce sidelobes, a Gaussian taper of the coupling was introduced, resulting in a similar taper in the spatial sensitivity. This was particularly important because at the frequencies for which this type of instrument was constructed, typically in the range 85–408 MHz, source confusion is a serious problem, as noted in Chapter 1. Sidelobe responses can be mistaken for sources and can also mask genuine sources. For a spatial sensitivity function of uniform rectangular character, the beam has a sinc function ($\sin \pi x/\pi x$) profile, for which the first sidelobe has a relative strength of 0.217. For a uniform, circular, spatial transfer function the beam has a profile of the form $J_1(\pi x)/\pi x$ for which the first sidelobe has a relative strength of 0.132. Sidelobes for a uniform circular (u, v) coverage are less than for a rectangular one, but would still be a problem in conditions of source confusion. Thus the uniform weighting may not be optimum for conditions of high source density, such as those found at low frequencies.

Holes in the Spatial Frequency Coverage

Consider a circular (u, v) area of diameter a_λ wavelengths in which there are no holes in the data; that is, the visibility data interpolated onto a rectangular grid for Fourier transformation has no missing values. Then for uniform weighting, the synthesized beam, which is obtained from the Fourier transform of the gridded transfer function, has the form $J_1(\pi a_\lambda \theta)/\pi a_\lambda \theta$, where θ is the angle measured from the beam center. If centrally concentrated weighting is used, the beam is a smoothed form of this function. Let us refer to the (u, v) area described above as the complete (u, v) coverage and the resulting beam as the complete response. Now if some data are missing, the actual (u, v) coverage is equal to the complete coverage minus the (u, v) hole distribution. By the additive property of Fourier transforms, the corresponding synthesized beam is equal to the complete response minus the Fourier transform of the hole distribution. The holes add an unwanted component to the complete response, in effect adding sidelobes to the synthesized beam. From Parseval's theorem the rms amplitude of the hole-induced sidelobes is proportional to the rms value of the missing spatial sensitivity represented by the holes. Other sidelobes also occur as a result of the oscillations in the $J_1(\pi a_\lambda \theta)/\pi a_\lambda \theta$ profile of the complete response, but there is clearly a contribution from the holes.

5.5 LINEAR TRACKING ARRAYS

We now consider interferometers or arrays in which the locations of the antennas are confined to a straight line. We have seen that for pairs of antennas with east–west spacings, the tracking loci in the (u, v) plane are a series of ellipses centered on the (u, v) origin. To obtain complete ellipses it is necessary that the tracking covers a range of 12 h in hour angle. If the antenna spacings of an east–west array increase in uniform increments, the spatial sensitivity is represented by a series of concentric ellipses with uniform increments in their axes. The angular resolution obtained is inversely proportional to the width of the (u, v) coverage in the corresponding direction; the width in the v direction is equal to that in the u direction times the sine of the declination, δ. East–west linear arrays containing spacings at multiples of a basic interval have found wide use in radio astronomy, particularly for observations at $|\delta|$ greater than $\sim 30°$.

In the simplest type of linear array the antennas are spaced at uniform intervals ℓ_λ (see Fig. 1.13a). This type of array is sometimes known as a grating array, by analogy with an optical diffraction grating. If there are n_a antennas, such an array output contains $(n_a - 1)$ combinations with the unit spacing, $(n_a - 2)$ with twice the unit spacing, and so on. Thus short spacings are highly redundant, and one is led to seek other ways to configure the antennas to provide larger numbers of different spacings for a given n_a. Note, however, that redundant observations can be used as an aid in calibration of the instrumental response and effects of the atmosphere, so some degree of redundancy is arguably beneficial (Hamaker, O'Sullivan, and Noordam 1977).

An antenna configuration with no redundant spacings that was used by Arsac (1955) is shown in Fig. 5.10a. The six possible pair combinations all have different spacings. With more than four antennas there is always either some redundancy or some missing spacings. A five-element, *minimum-redundancy* configuration devised by Bracewell (1966) is shown in Fig. 5.10b. Moffet (1968) listed examples of minimum-redundancy arrays of up to 11 elements, and solutions for larger arrays are discussed by Ishiguro (1980). Moffet defined two classes. These are restricted arrays in which all spacings up to the maximum spacing, $n_{\max}\ell_\lambda$ (that is, the total length of the array), are present, and general arrays in which

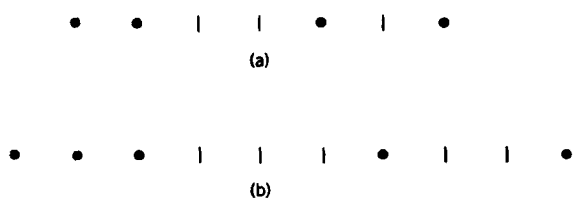


Figure 5.10 Two linear array configurations in which the antennas are represented by filled circles. (a) Arsac's (1955) configuration containing all spacings up to six times the unit spacing, with no redundancy. (b) Bracewell's (1966) configuration containing all spacings up to nine times the unit spacing, with the unit spacing occurring twice.

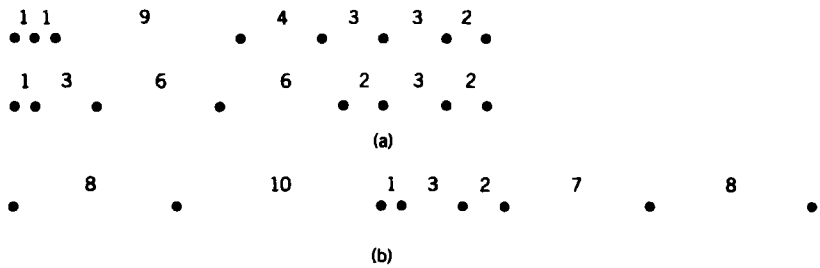


Figure 5.11 Eight-element, minimum-redundancy, linear arrays: the numbers indicate spacings in multiples of the unit spacing. (a) Two arrays that uniformly cover the range of 1 to 23 times the unit spacing. (b) An array that uniformly covers 1 to 24 times the unit spacing, but has a length of 39 times the unit spacing. The extra spacings are 8, 31 (twice), and 39 times the unit spacing. From Moffet (1968), ©1968 IEEE.

all spacings up to some particular value are present, and also some longer ones. Examples for eight elements are shown in Fig. 5.11. A measure of redundancy for a linear array is given by the expression

$$\tfrac{1}{2}n_a(n_a - 1)/n_{\max}, \tag{5.18}$$

which is the number of antenna pairs divided by the number of unit spacings in the longest spacing. This is equal to 1.0 and 1.11 for the configurations in Fig. 5.10a and 5.10b, respectively. A study in number theory by Leech (1956) indicates that for large numbers of elements this redundancy factor approaches 4/3. An example of a linear, minimum-redundancy array that uses the configuration in Fig. 5.10b is described by Bracewell et al. (1973).

The ability to move a small number of elements adds greatly to the range of performance of an array. Figure 5.12 shows the arrangement of the three antennas of the Cambridge One-Mile Telescope (Ryle 1962). Antennas 1 and 2 are fixed, and their outputs are correlated with that from antenna 3 which can be moved on a rail track. In each position of antenna 3 the source under observation is tracked for 12 h, and visibility data are obtained over two elliptical loci in the (u, v) plane.

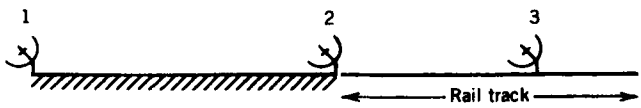


Figure 5.12 The Cambridge One-Mile Radio Telescope. Antennas 1 and 2 are at fixed locations, and the signals they receive are each correlated with the signal from antenna 3, which can be located at various positions along a rail track. The fixed antennas are 762 m apart and the rail track is a further 762 m long. The unit spacing is equal to the increment of the position of antenna 3, and all multiples up to 1524 m can be obtained.

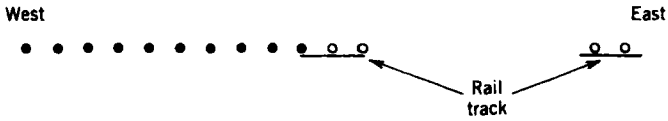


Figure 5.13 Antenna configuration of the Westerbork Synthesis Radio Telescope. The 10 filled circles represent antennas at fixed locations, and the 4 open circles represent antennas that are movable on rail tracks. Forty correlators are used to combine the signals from each of the fixed antennas with the signals from each of the movable ones. The diameter of the antennas is 25 m and the spacing of the fixed antennas is 144 m. As originally constructed, the array contained only the 12 western antennas, and the 2 at the east end were added later to double the range of spacings.

The observation is repeated as antenna 3 is moved progressively along the track, and the increments in the position of this antenna determine the spacing of the elliptical loci in the (u, v) plane. From the sampling theorem (Section 5.2), the required (u, v) spacing is the reciprocal of the angular width, in radians, of the source under investigation. The ability to vary the incremental spacing adds versatility to the array and reduces the number of antennas required. The configuration of a larger instrument of this type, the Westerbork Synthesis Radio Telescope (Baars and Hooghoudt 1974, Högbom and Brouw 1974, Raimond and Genee 1996), is shown in Fig. 5.13. Here ten fixed antennas are combined with four movable ones, and the rate of gathering data is approximately 20 times greater than with the three-element array.

The sampling of the visibility function at points on concentric, equispaced ellipses results in the introduction of ringlobe responses. These may be understood by noting that for a linear array the instantaneous spacings are represented in one dimension by a series of δ functions, as shown in Fig. 5.14a. If the array contains all multiples of the unit spacings up to $N\ell_\lambda$, and if the corresponding visibility measurements are combined with equal weights, the instantaneous response is a

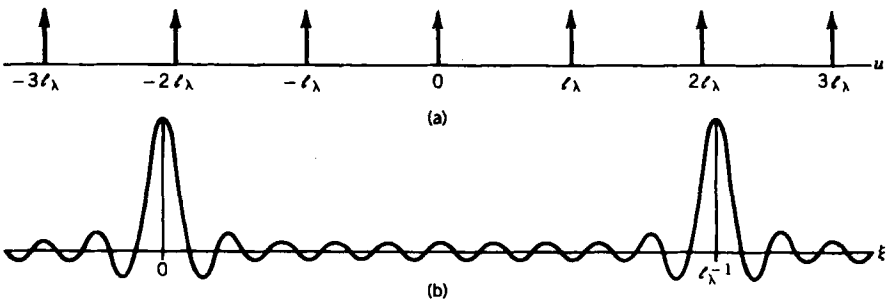


Figure 5.14 Part of a series of δ functions representing the instantaneous distribution of spacings for a uniformly spaced linear array with equal weight for each spacing. (b) Part of the corresponding series of fan beams that constitute the instantaneous response. Parts (a) and (b) represent the left- and right-hand sides of Eq.(5.19), respectively.

series of fan beams, each with a profile of sinc-function form, as in Fig. 5.14b. This follows from the Fourier transform relationship for a truncated series of delta functions:

$$\sum_{i=-N}^{N} \delta(u - i\ell_\lambda) \rightleftharpoons \frac{\sin[(2N+1)\pi\ell_\lambda l]}{\pi\ell_\lambda l} * \sum_{k=-\infty}^{\infty} \delta\left(l - \frac{k}{\ell_\lambda}\right). \tag{5.19}$$

Here $\rightleftharpoons$ represents Fourier transformation, and the delta functions on the left-hand side represent the spacings in the u domain. The series on the left is truncated, and can be envisaged as selected from an infinite series by multiplication with a rectangular window function. The right-hand side represents the beam pattern in which the Fourier transform of the window function is replicated by convolution with delta functions. As the earth's rotation causes the spacing vectors to sweep out ellipses in the (u, v) plane, the corresponding rotation of the array relative to the sky can be visualized as causing a central fan beam to rotate into a narrow pencil beam, while its neighbors give rise to lower-level, ring-shaped responses concentric with the central beam, as shown in Fig. 5.15. This general argument gives the correct spacing of the ringlobes, the profile of which is modified from the sinc-function form.

If the spatial sensitivity in the (u, v) plane is a series of circular delta functions of radius $q, 2q, \ldots, Nq$, the profile of the kth ringlobe is of the form

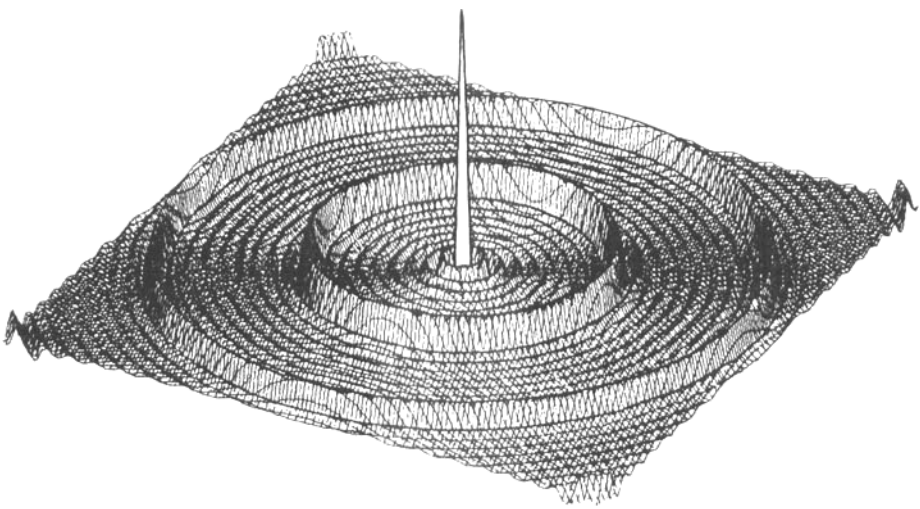


Figure 5.15 Example of ringlobes. The response of an array for which the spatial transfer function is a series of nine circles concentric with the (u, v) origin, resulting, for example, from observations with an east–west linear array with 12 h tracking at a high declination. The radii of these circles are consecutive integral multiples of the unit antenna spacing. The weighting corresponds to the principal response discussed in Section 10.2 under *Weighting of the Visibility Data*. From Bracewell and Thompson (1973).

$$\operatorname{sinc}^{1/2}\left[2\left(N+\tfrac{1}{2}\right)(qr-k)\right], \tag{5.20}$$

where $r = \sqrt{l^2+m^2}$. The function $\operatorname{sinc}^{1/2}(\chi)$ is plotted in Fig. 5.16 and is the half-order derivative of $\sin \pi\chi/\pi\chi$. It can be computed using Fresnel integrals (Bracewell and Thompson 1973).

The application of the sampling theorem (Section 5.2) to the choice of incremental spacing requires that the latter be no greater than the reciprocal of the source width. In terms of ringlobes, this condition ensures that the minimum ringlobe spacing is no less than the source width. Thus, if the sampling theorem is followed, the main-beam response to a source just avoids being overlapped by a ringlobe response to the same source. In arrays such as those in Figs. 5.12 and 5.13, ringlobes can be effectively suppressed if the movable antennas are positioned in steps slightly less than the antenna diameter, in which case the ringlobe lies outside the primary antenna beam. Note, however, that the first spacing cannot be less than the antenna diameter, and the missing low-spacing measurements may have to be obtained by other means (see the discussion of mosaicking in Section 11.6). Ringlobes can also be greatly reduced by image-processing techniques such as the CLEAN algorithm which is described in Section 11.2.

Although the elliptical loci in the (u, v) plane are spaced at equal intervals, the natural weighting of the data for an east–west linear array is not uniform, because in any interval of time the antenna-spacing vectors move a distance proportional to their length. In the projection of the (u, v) plane onto the equatorial plane of the earth, which is discussed in Section 4.2 as the (u', v') plane, the spacing

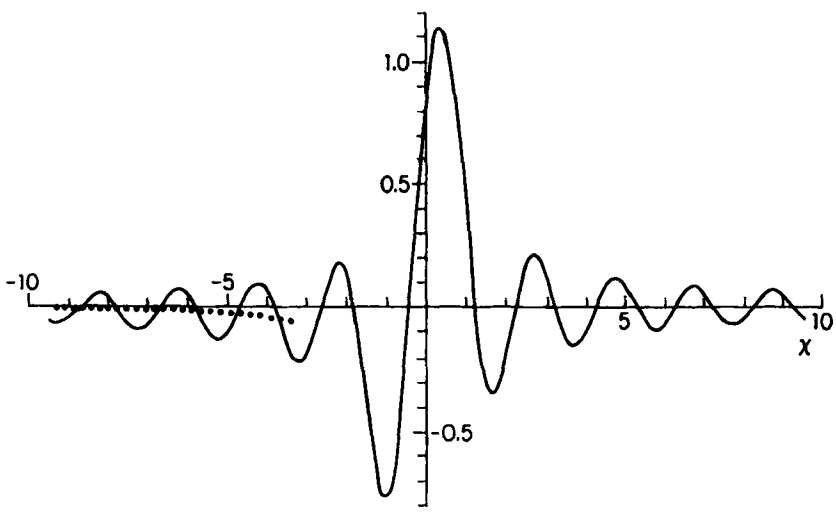


Figure 5.16 Cross section of a ringlobe in the principal response to a point source of an east–west array with uniform increments in antenna spacing. The left-hand side is the inside of the ring and the right is the outside. The dotted line indicates a negative mean level of the oscillations on the inner side. From Bracewell and Thompson (1973).

vectors rotate at constant angular velocity, and the density of measured points is proportional to

$$q'^{-1} = (u'^2 + v'^2)^{-1/2} = (u^2 + v^2\mathrm{cosec}^2\delta^2)^{-1/2}. \tag{5.21}$$

In the (u, v) plane the density of measurements, averaged over an area of dimensions comparable to the unit spacing of the antennas, is inversely proportional to $\sqrt{u^2 + v^2\mathrm{cosec}^2\delta}$. Along a straight line through the (u, v) origin the density is inversely proportional to $\sqrt{u^2 + v^2}$.

5.6 TWO-DIMENSIONAL TRACKING ARRAYS

As noted previously, the spatial frequency coverage for an east–west linear array becomes severely foreshortened in the v dimension for observations near the celestial equator. For such observations a configuration of antennas is required in which the Z component of the antenna spacing, as defined in Section 4.1, is comparable to the X and Y components. This is achieved by including spacings with azimuths other than east–west. The configuration is then two-dimensional. An array located at an intermediate latitude and designed to operate at low declinations can cover the sky from the pole to declinations of about 30° into the opposite celestial hemisphere. This range includes about 70% of the total sky, that is, almost three times as much as that of an east–west array. Since the Z component is not zero, the elliptical (u, v) loci are broken into two parts as shown in Fig. 4.4. As a result, the pattern of the (u, v) coverage is more complex than is the case for an east–west linear array, and the ringlobes that result from uniform spacing of the loci are replaced by more complex sidelobe structure. In two dimensions the choice of a minimum-redundancy configuration of antennas is not as simple as for a linear array. A first step is to consider the desired spatial transfer function $W(u, v)$. There is no known analytical way to go from $W(u, v)$ to the antenna configuration, but iterative methods of finding an optimum, or near-optimum, solution are available.

First consider the effect of tracking a source across the sky, and suppose that for a source near the zenith the *instantaneous* spatial frequency coverage results in approximately uniform sampling within a circle centered on the (u, v) origin. At any time during the period of tracking of the source, the (u, v) coverage is the zenith coverage projected onto the plane of the sky, with some degree of rotation that depends on the hour angle and declination of the source. The projection results in foreshortening of the coverage from a circular to an elliptical area, still centered on the (u, v) origin, and this foreshortening is least at meridian transit. The effect of observing over a range of hour angle can be envisaged as averaging a range of elliptical (u, v) areas that suffer some rotation of the major axis. At the center of the (u, v) plane there will be an area that remained within the foreshortened coverage over the whole observation, and if the instantaneous coverage is uniform, then it will remain uniform within this area. Outside the area, the foreshortening will cause the coverage to taper off smoothly. These effects depend on the declination of the source and the range of hour-angle tracking.

Practical experience indicates that some tapering of the visibility measurements is seldom a serious problem. Thus it can be expected that two-dimensional arrays in which the number of antennas is large enough to provide good instantaneous (u, v) coverage will also provide good performance when used with hour angle tracking.

Open-Ended Configurations

For configurations with open-ended arms such as the cross, T, and Y, the spatial frequency coverage is shown in Fig. 5.7. The spatial frequency coverage of the cross and T has four-fold symmetry in both cases; we ignore the effect of the missing small extensions on the top and bottom sides of the square for the T. The spatial frequency coverage of the equiangular Y array (120° between adjacent arms) has six-fold symmetry. (n-fold symmetry denotes a figure that is unchanged by rotation through $2\pi/n$. For a circle, n becomes infinite, and other figures approach circular symmetry as n increases.) The autocorrelation function of the equiangular Y is closer to circular symmetry than that of a cross or T. In this respect a five-armed array, as suggested by Hjellming (1989), would be better still, but more expensive.

As an example of the open-ended configuration, we examine some details of the design of the VLA (Thompson et al. 1980; Napier, Thompson, and Ekers 1983). This instrument is located at latitude 34° N in New Mexico and is able to track objects as far south as −30° for almost 7 h without going below 10° in elevation. Performance specifications called for mapping with full resolution down to at least −20° declination and for obtaining a map in no more than 8 h of observation without moving antennas to new locations. In designing the array, comparison of the performance of various antenna configurations was accomplished by computing the spatial transfer function with tracking over an hour angle range ±4 h at various declinations. In judging the merit of any configuration the basic concern was to minimize sidelobes in the synthesized beam. It was found that the percentage of holes in the (u, v) coverage was a consistent indication of the sidelobe levels of the synthesized beam, and to judge between different configurations, it was not always necessary to calculate the detailed response (NRAO 1967, 1969). For a given number of antennas, the equiangular Y was found to be superior to the cross and T; see Fig. 5.17.

Inverting the Y has no effect on the beam, but if the antennas have the same radial disposition on each arm, the performance near zero declination is improved by rotating the array so that the nominal north or south arm makes an angle of about 5° with the north–south direction. Without this rotation the baselines between corresponding antennas on the other two arms are exactly east–west, and for $\delta = 0°$ the spacing loci degenerate to straight lines that are coincident with the u axis and become highly redundant. The total number of antennas, 27, was chosen from a consideration of (u, v) coverage and sidelobe levels, and resulted in peak sidelobes at least 16 dB below the main-beam response, except at $\delta = 0°$ where earth rotation is least effective. The 27 antennas provide 351 pair combinations.

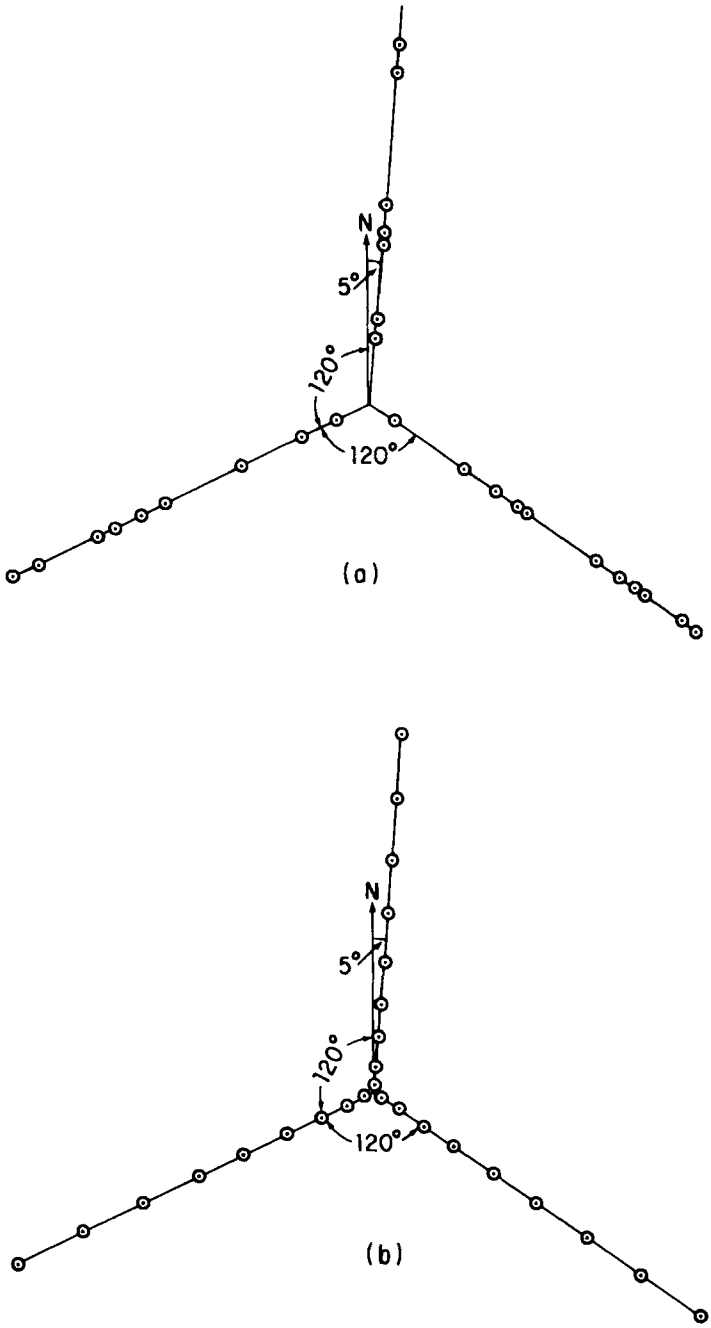


Figure 5.17 (a) Proposed antenna configuration for the VLA that resulted from Mathur's (1969) computer-optimized design. (b) Power-law design (Chow 1972) adopted for the VLA. From Napier, Thompson, and Ekers (1983), ©1983 IEEE.

The positions of the antennas along the arms provide another set of variables that can be adjusted to optimize the spatial transfer function. Figure 5.17 shows two approaches to the problem. Configuration (a) was obtained by using a pseudodynamic computation technique (Mathur 1969), in which arbitrarily chosen initial conditions were adjusted by computer until a near-optimum (u, v) coverage was reached. Configuration (b) shows a power-law configuration derived by Chow (1972). This analysis led to the conclusion that a spacing in which the distance of the nth antenna on an arm is proportional to n^{α} would provide good (u, v) coverage. Comparison of the empirically optimized configuration with the power-law spacing with $\alpha \simeq 1.7$ showed the two to be essentially equal in performance. The power-law result was chosen largely for reasons of economy. A requirement of the design was that four sets of antenna stations be provided to vary the scale of the spacings in four steps, to allow a choice of resolution and field of view for different astronomical objects. By making α equal to the logarithm to the base 2 of the scale factor between configurations, the location of the nth station for one configuration coincides with that of the $2n$th station for the next-smaller configuration. The total number of antenna stations required was thereby reduced from 108 to 72. Plots of the spatial frequency coverage are shown in Fig. 5.18. The snapshot in Fig. 5.18d shows the instantaneous coverage, which is satisfactory for mapping simple structure in strong sources.

Closed Configurations

The discussion here will largely follow that of Keto (1997). Returning to the proposed criterion of uniform distribution of measurements within a circle in the (u, v) plane, we note that a configuration of antennas around a circle (a ring array) provides a useful starting point since the distribution of antenna spacings cuts off sharply in all directions at the circle diameter. This is shown in Fig. 5.7g and h. We begin by considering the instantaneous (u, v) coverage for a source at the zenith. This is shown in Fig. 5.19a for 21 equally spaced antenna locations indicated by triangles. There are 21 antenna pairs at the unit spacing, uniformly distributed in azimuth, and each of these is represented by two points in the (u, v) plane. The same statement can be made for any other paired spacings around the circle. As a result, the spatial transfer function consists of points that lie on a pattern of circles and radial lines. Note also that as the spacings approach the full diameter of the circle the distance between antennas increases only very slowly. For example, the direct distance between antennas spaced 10 intervals around the circle is very little more than that for antennas at 9 intervals. Thus there is an increase in the density of measurements at the longest spacings (the points along any radial line become more closely spaced) as well as a marked increase toward the center. Note that the density of points closely follows the radial profile of the autocorrelation function in Fig. 5.7j, except close to the origin since Fig. 5.19 includes only cross-correlations between antennas.

One way of obtaining a more uniform distribution is to randomize the spacings of the antennas around the circle. The (u, v) points are then no longer constrained to lie on the pattern of circles and lines, and Fig. 5.19b shows an example in which

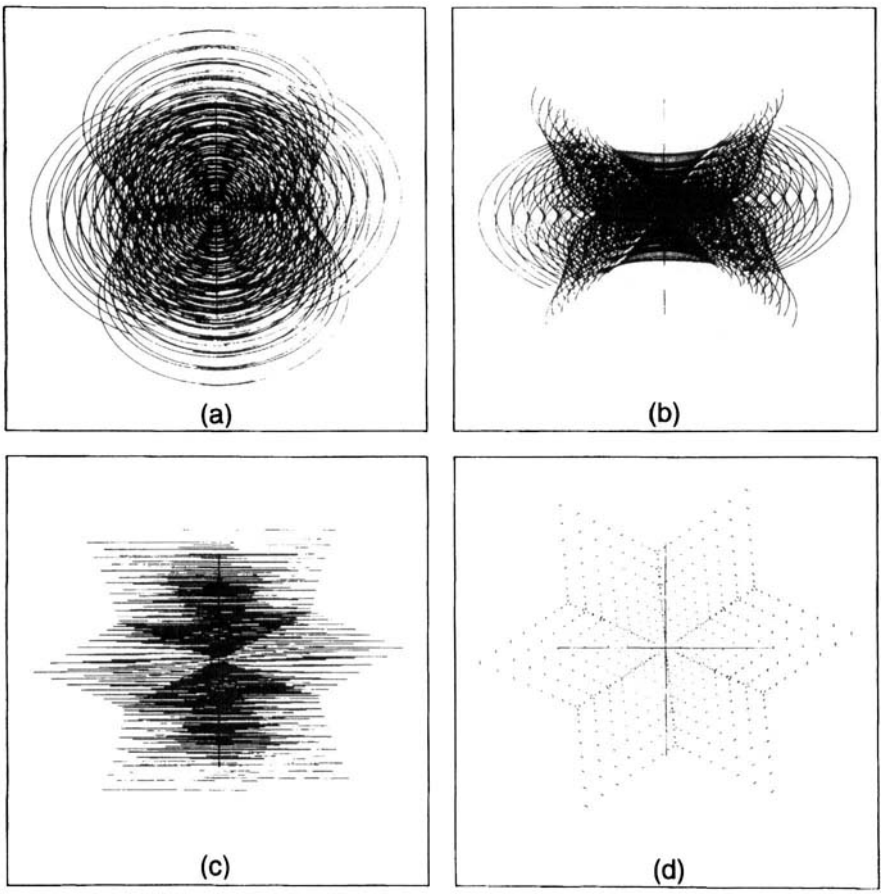


Figure 5.18 Spatial frequency coverage for the VLA with the power-law configuration of Fig. 5.17b: (a) $\delta = 45°$; (b) $\delta = 30°$; (c) $\delta = 0°$; (d) snapshot at zenith. The range of hour angle is ± 4 h or as limited by a minimum pointing elevation of 9°, and ± 5 min for the snapshot. The lengths of the (u, v) axes from the origin represent the maximum distance of an antenna from the array center, that is, 21 km for the largest configuration. From Napier, Thompson, and Ekers (1983), ©1983 IEEE.

a partial optimization has been obtained by computation using a neural-net algorithm. Keto (1997) discusses various algorithms for optimizing the uniformity of the spatial sensitivity. An earlier investigation of circular arrays by Cornwell (1988) also resulted in good uniformity within a circular (u, v) area. In this case an optimizing program based on simulated annealing was used, and the spacing of the antennas around the circle shows various degrees of symmetry that result in patterns resembling crystalline structure in the (u, v) spacings.

Optimizing the antenna configurations can also be considered more broadly, and Keto (1997) notes that the cutoff in spacings at the same value for all di-

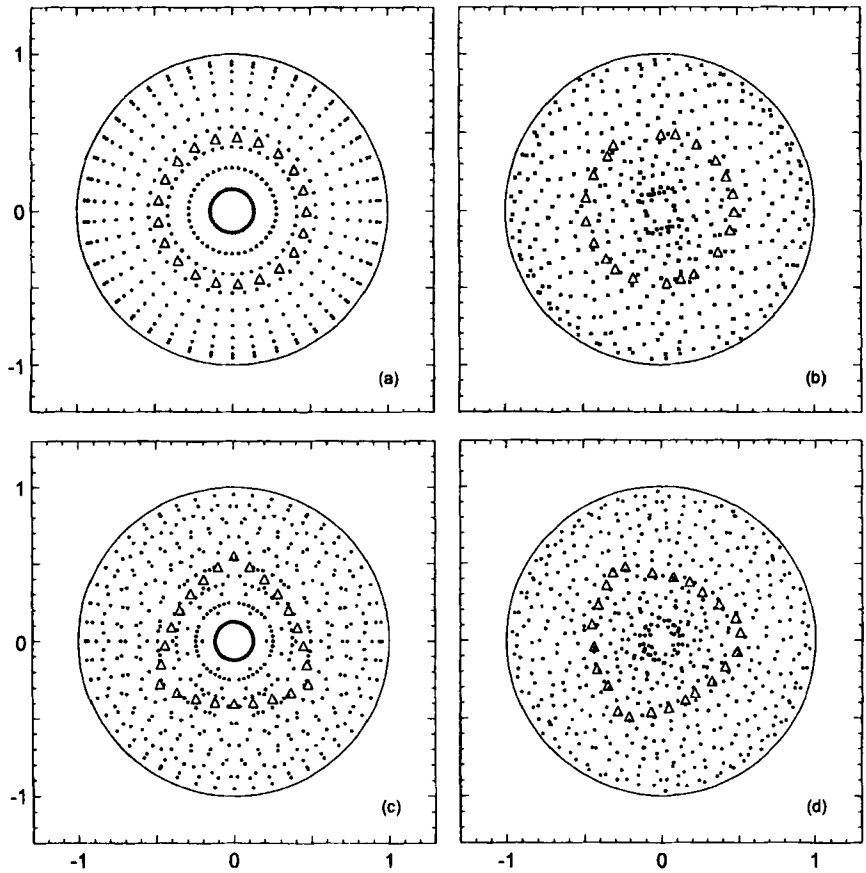


Figure 5.19 (a) A circular array with 21 uniformly spaced antennas indicated by the triangles, and the instantaneous spatial frequency coverage indicated by the points. The scale of the diagrams is the same for both the antenna positions and the spatial frequency coordinates u and v. (b) The array and spatial frequency coverage as in (a) but after adjustment of the antenna positions around the circle to improve the uniformity of the coverage. (c) An array of 24 antennas equally spaced around a Reuleaux triangle, and the corresponding spatial frequency coverage. (d) The array and spatial sensitivity as in (c) with adjustment of the antenna spacing to optimize the uniformity of the coverage. From Keto (1997), ©1997 American Astron. Soc.

rections is not unique to the circular configuration. There are other figures, such as the Reuleaux triangle, for which the width is constant in all directions. The Reuleaux triangle is shown in Fig. 5.7i, and consists of three equal circular arcs indicated by the solid lines. The total perimeter is equal to that of a circle of diameter equal to one of the sides of the equilateral triangle shown by the broken lines. Similar figures can be constructed for any regular polygon with an odd number of sides, and a circle represents such a figure for which the number tends to infinity. The Reuleaux triangle is the least symmetrical of this family of figures. Other

facts about the Reuleaux triangle and similar figures can be found in Rademacher and Toeplitz (1957).

Since the optimization of the circular array in Fig. 5.19b results in a reduction in the symmetry, it may be expected that an array based on the Reuleaux triangle would provide better uniformity in the spatial frequency coverage than the circular array. This is indeed the case, as can be seen by comparing Fig. 5.19a and c for both of which the antenna spacing is uniform. The circular array with irregular antenna spacings in Fig. 5.19b was obtained by starting with a circular array and allowing antenna positions to be moved small distances. In this case the program was not allowed to reach a fully optimized solution. Allowing the optimization to run to convergence results in antennas at irregular spacings around a Reuleaux triangle, as shown in Fig. 5.19d. This result does not depend on the starting configuration. Comparison of Figs. 5.19b and d shows that the difference between the circle and the Reuleaux triangle is much less marked when they have both been subjected to some randomization of the antenna positions around the figure, although a careful comparison shows the uniformity in Fig. 5.19d to be a little better than in b.

Figure 5.20 shows the spatial frequency coverage for an array in an optimized Reuleaux triangle configuration. The tracking range is $\sim \pm 3$ h of hour angle, and the latitude is equal to that of the VLA. Comparison of these figures with corresponding ones for the VLA in Fig. 5.18 shows that the Reuleaux triangle produces spatial frequency coverage that is closer to the uniformly sampled circular area than does the equiangular Y configuration. As indicated in Fig. 5.7, the autocorrelation function of a figure with linear arms contains high values in directions where the arms of overlapping figures line up. This effect contributes to the lack of uniformity in the spatial sensitivity of the Y array. Curvature of the arms or quasirandom lateral deviations of the antennas from the arms helps to smear the sharp structure in the spatial transfer function. The high values along radial lines do not occur in the autocorrelation function of a circle or similar closed figure, which is one reason why configurations of this type provide more uniform spatial frequency coverage.

Despite some less-than-ideal features of the equiangular Y, the VLA produces astronomical images of very high quality. Thus, although the circularity and uniformity of the spatial frequency coverage is a useful criterion, this is not a highly critical factor. So long as the measurements cover the range of u and v for which the visibility is high enough to be measurable, and the source is strong enough that any loss in sensitivity resulting from nonuniform weighting can be tolerated, excellent results can be obtained. The Y array has a number of practical advantages over a closed configuration. When several scaled configurations are required to allow for a range of angular resolution, the alternative locations lie along the same arms, whereas with the circle or Reuleaux triangle, separate scaled configurations are required. The flexibility of the Y array is particularly useful in VLA observations at southern declinations for which the projected spacings are seriously foreshortened in the north–south direction. For such cases it is possible to move the antennas on the north arm onto the positions for the next-larger configuration, and thereby substantially compensate for the foreshortening.

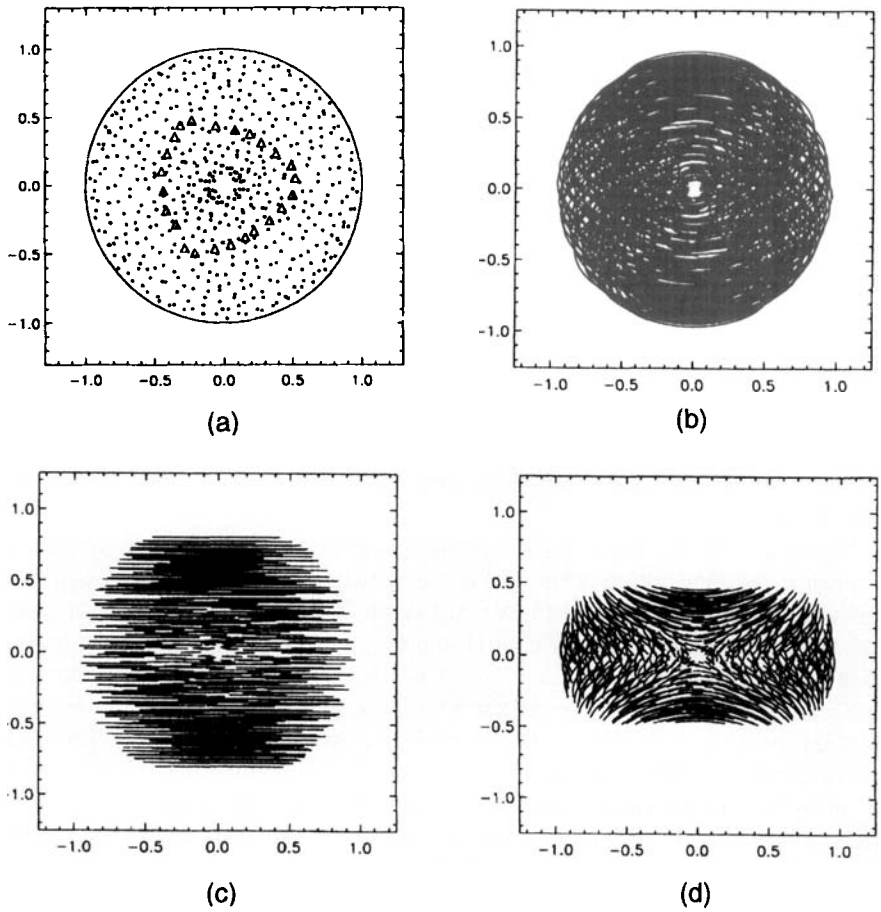


Figure 5.20 Spatial frequency coverage for a closed configuration of 24 antennas optimized for uniformity of measurements in the snapshot mode: (a) snapshot at zenith; (b) $\delta = +30°$; (c) $\delta = 0°$; (d) $\delta = -28°$. The triangles in (a) indicate the positions of the antennas. The tracking is calculated for an array at 34° latitude to simplify comparison with the VLA (Fig. 5.18). For each declination shown the tracking range is the range of hour angle for which the source elevation is greater than 25°. From Keto (1997), ©1997 American Astron. Soc.

Some further interesting examples of important configurations are given below.

- The compact array of the Australia Telescope is an east–west linear array of six antennas, all movable on rail track (Frater, Brooks, and Whiteoak 1992).
- The UTR-2 is a T-shaped array of large-diameter, broadband dipoles built by the Ukrainian Academy of Sciences near Grakovo, Ukraine (Braude et al. 1978). The frequency range of operation is 10–25 MHz. Several smaller antennas of similar type have been constructed at distances up to approximately 900 km from the Grakovo site, and are used for VLBI observations.

- An array of 720 conical spiral antennas in a T-shaped configuration operating in the frequency range 15–125 MHz was constructed at Borrego Springs, California (Erickson, Mahoney, and Erb 1982).
- The Mauritius Radio Telescope, near Bras d'eau, Mauritius, is a T-shaped array of helix antennas operating at 150 MHz. The east–west arm is 2 km long. The south arm is 880 m long and is synthesized by moving a group of antennas on trolleys. The array is similar in principle to the one in Fig. 1.12a. It is intended to cover a large portion of the southern hemisphere.
- The GMRT (Giant Meter-wave Radio Telescope) near Pune, India, consists of 30 antennas, 16 of which are in a Y-shaped array with curved arms approximately 15 km long. The remaining 14 are in a quasirandom cluster in the central 2 km (Swarup et al. 1991). The antennas are 45 m in diameter and are at fixed locations. The highest operating frequency is approximately 1.6 GHz.
- A circular array with 96 uniformly spaced antennas was constructed at Culgoora, Australia, for observations of the sun (Wild 1967). This was a multibeam, scanning, phased array rather than a correlator array, consisting of 96 antennas uniformly spaced around a circle of diameter 3 km and operating at 80 and 160 MHz. To suppress unwanted sidelobes of the beam, Wild (1965) devised an ingenious phase-switching scheme termed J^2 synthesis. The spatial sensitivity of this ring array was analyzed by Swenson and Mathur (1967).
- The Submillimeter Array (SMA) of the Smithsonian Astrophysical Observatory and Academia Sinica of Taiwan, located on Mauna Kea, Hawaii, is the first array to be built using a Reuleaux triangle configuration (Moran 1998).

VLBI Configurations

In VLBI arrays the layout of antennas results from considerations of both (u, v) coverage and practical operating requirements. Important factors involve proximity to existing observatories for technical support services, and access to transportation centers for return of tapes to the correlator facility. Ranges of hour angle and declination that are simultaneously observable from the widely spaced locations must also be considered. Although the locations of the widely spaced antennas of a VLBI array deviate significantly from a plane, the angular widths of the sources under observation are generally sufficiently small that the small-field approximation can be used in deriving the radio image, as discussed in Section 3.1. Similar considerations apply to long-baseline arrays which operate in a connected-element mode using radio links to transmit IF and reference signals. An example of this type is the Multielement Radio-linked Interferometer Network (MERLIN) of the Jodrell Bank Observatory, England, which consists of six antennas with baselines up to 233 km (Thomasson 1986).

Any suitable radio telescope with an atomic frequency standard, phase-locked oscillators, and appropriate receiving and recording systems can be used in a VLBI experiment. Several organized networks have been set up to coordinate

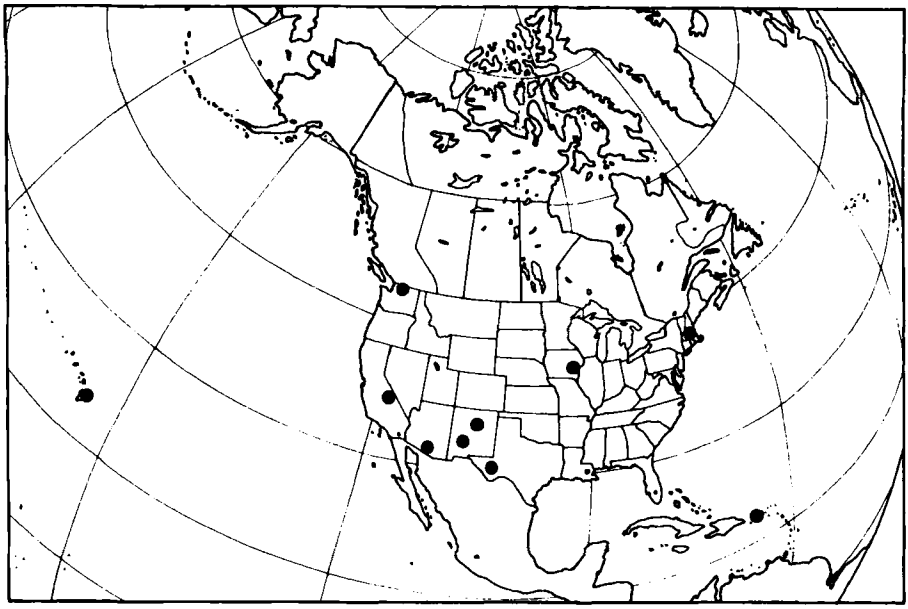
joint experiments between different observatories. For the first two decades after the inception of the VLBI technique, observations were mainly joint ventures between different institutions. Consideration of arrays dedicated solely to VLBI occurred as early as 1975 (Swenson and Kellermann 1975), but construction of such instruments did not begin for another decade. A study of antenna locations for a VLBI array has been described by Seielstad, Swenson, and Webber (1979). To obtain a single index as a measure of the performance of any configuration, the spatial transfer function was computed for a number of declinations. The fraction of appropriately sized (u, v) cells containing measurements was then weighted in proportion to the area of sky at each declination and averaged. Maximizing the index, in effect, minimizes the number of holes (unfilled cells). Other studies have involved computing the response to a model source, synthesizing a map, and comparing the result with the model.

The design of an array dedicated to VLBI, the Very-Long-Baseline Array (VLBA) of the United States, is described by Napier et al. (1994). The antenna locations are listed in Table 5.1 and shown in Fig. 5.21a. A discussion of the choice of sites is given by Walker (1984). Antennas in Hawaii and St. Croix provide long east–west baselines. Massachusetts to Saint Croix is the longest north–south spacing. A site in Alaska would be further north, but would be of limited benefit because it would provide only restricted accessibility for sources at southern declinations. An additional site within the southern hemisphere would enhance the (u, v) coverage at southern declinations. The south-eastern region of the United States is avoided because of the higher levels of water vapor in the atmosphere. Intermediate north–south baselines are provided by the drier west coast area. The Iowa site fills in a gap between Massachusetts and the southwestern sites. The short spacings are centered on the VLA to allow the possibility of development of real-time linkage to it, and as a result the spatial frequency coverage shows a degree of central concentration. This enables the array to make measurements on a wider range of source sizes than would be possible with the same number of

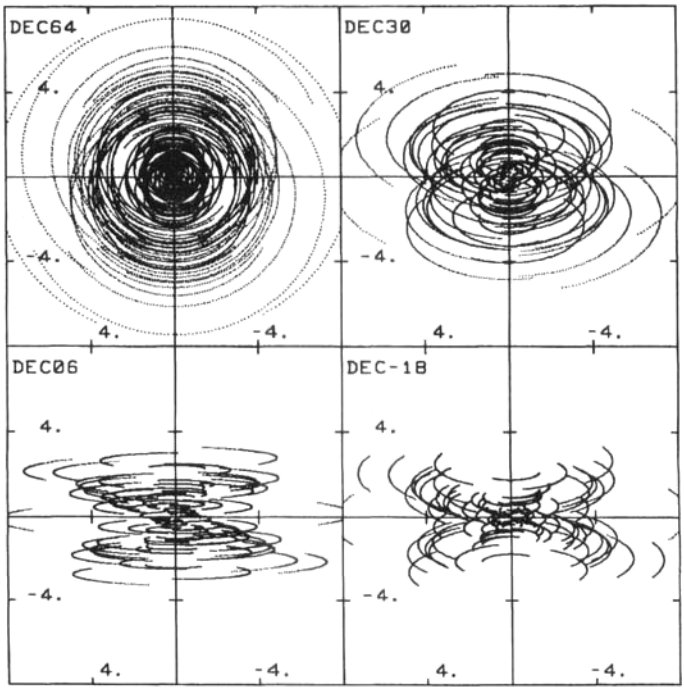
TABLE 5.1 Locations of Antennas in the VLBA[a]

Location	N. Latitude (deg	min	sec)	W. Longitude (deg	min	sec)	Elevation (m)
St. Croix, VI	17	45	30.57	64	35	02.61	16
Hancock, NH	42	56	00.96	71	59	11.69	309
N. Liberty, IA	41	46	17.03	91	34	26.35	241
Fort Davis, TX	30	38	05.63	103	56	39.13	1615
Los Alamos, NM	35	46	30.33	106	14	42.01	1967
Pie Town, NM	34	18	03.61	108	07	07.24	2371
Kitt Peak, AZ	31	57	22.39	111	36	42.26	1916
Owens Valley, CA	37	13	54.19	118	16	33.98	1207
Brewster, WA	48	07	52.80	119	40	55.34	255
Mauna Kea, HI	19	48	15.85	155	27	28.95	3720

[a]Data from Napier et al. (1994), © 1994 IEEE.



(a)



(b)

Figure 5.21 Very-Long-Baseline Array in the United States: (a) locations of the 10 antennas, and (b) spatial frequency coverage (spacings in thousands of kilometers) for declinations of 64°, 30°, 6°, and −18°, in which the observing time at each antenna is determined by an elevation limit of 10°. From Walker (1984).

antennas and more uniform coverage. However, this results in some sacrifice in capability for mapping complex sources.

Orbiting VLBI Antennas

A logical step in the development of VLBI from ground-based arrays is the addition of antennas in space (Preston et al. 1983, Burke 1984). The combination of orbiting VLBI (OVLBI) and ground-based antennas has several obvious advantages. Higher angular resolution can be achieved, and the ultimate limit may be set by interstellar scintillation (see Section 13.6). The orbital motion of the spacecraft helps to fill in the coverage in the (u, v) plane, and thereby improves the detail and dynamic range in the resulting images.

Figure 5.22 shows an example of the (u, v) coverage for observations with the HALCA spacecraft (Hirabayashi et al. 1998) and a series of terrestrial antennas: one at Usuda, Japan, one at the VLA site, and the 10 VLBA antennas. The

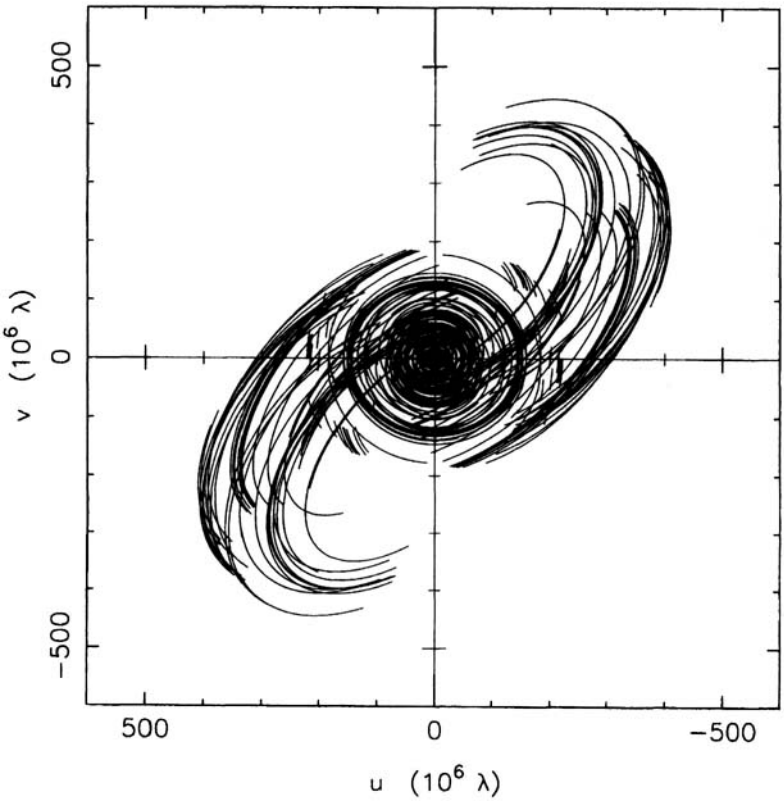


Figure 5.22 An example of spatial frequency coverage for the HALCA satellite with 12 ground-based antennas. This is for an observation of the source 1622 + 633 at 5 GHz frequency. See text for further details.

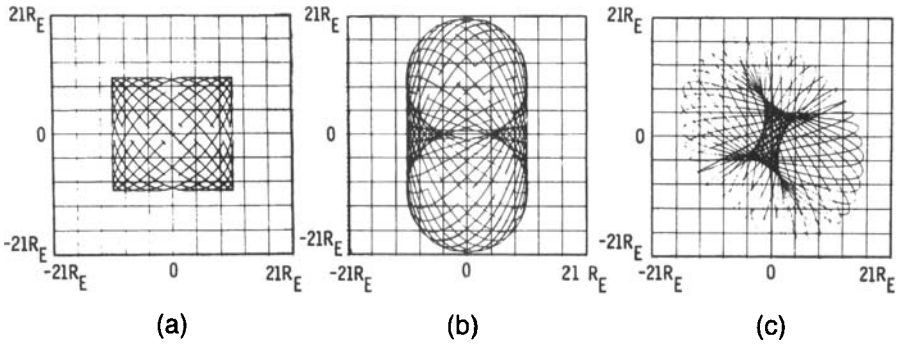


Figure 5.23 Spatial frequency coverage for two antennas on satellites with circular orbits of radius approximately ten times the earth's radius R_E: (a) source along X axis; (b) source along Y or Z axes; (c) source centered between X, Y, and Z axes. The orbits lie in the XY and XZ planes of a rectangular coordinate system. The satellite periods differ by 10% and the observing period is approximately 20 days. From Preston et al., in *Very Long Baseline Interferometry Techniques*, F. Biraud, Ed., Cepadues, France, 1983.

spacecraft orbit is inclined at an angle of 31° to the earth's equator, and the height above the earth's surface is 21,400 km at apogee and 560 km at perigee. The spacings shown are for a frequency of 5 GHz, and the units of u and v are 10^6 wavelengths; the maximum spacing is 5×10^8 wavelengths which corresponds to a fringe width of 0.4 mas. The approximately circular loci at the center of the figure represent baselines between terrestrial antennas. The orbital period is 6.3 h and the data shown correspond to an observation of duration about four orbital periods. The spacecraft orbit precesses at a rate of order 1° per day, and over the course of one to two years, the coverage of any particular source can be improved by combining several observations. Figure 5.23 shows an example of the (u, v) coverage that could be obtained between two spacecraft in circular orbits of radius about ten earth radii, with orthogonal planes that have periods differing by 10%. In practice there are likely to be restrictions on coverage resulting from the limited steerability of the astronomy and communication antennas relative to the spacecraft. It is necessary for the spacecraft to maintain an attitude in which the solar power panels remain illuminated and the communications antenna can be pointed toward the earth. A discussion of some technical points concerning orbiting VLBI is given in Section 9.10.

Planar Arrays

Studies of cosmic background radiation and the Sunyaev–Zel'dovich effect require observations with very high brightness sensitivity at wavelengths of order 1 cm and shorter: see also Section 10.4 under *Cosmic Background Anisotropy*. Unlike the sensitivity to point sources, the sensitivity to a broad feature that largely fills the antenna beam does not increase with increasing collecting area of the antenna. Thus, for cosmic background measurements large antennas are

not required. Extremely good stability is necessary to allow significant measurements at the level of a few tens of microkelvins per beam, that is, of order 10^{-6} Jy arcmin^{-2}. Special arrays have been designed for this purpose. A number of antennas are mounted on a platform, with their apertures in a common plane. The whole structure is then supported on an altazimuth mount so that the antennas can be pointed to track any position on the sky. An example of such an instrument, the Cosmic Background Imager (CBI), has been developed by A. C. S. Readhead and colleagues at Caltech (Padin et al. 2001). It consists of thirteen Cassegrain focus paraboloids, each of diameter 90 cm, which are operated in the 26–36-GHz range. In this instrument the antenna mounting frame has the shape of an irregular hexagon with three-fold symmetry and maximum dimensions of approximately 6.5 m, as shown in Fig. 5.24. For the particular type of measurements required, the planar array has a number of desirable properties compared with a single antenna of similar aperture, or a number of individually mounted antennas, as outlined below:

- The use of a number of individual antennas allows the output to be measured in the form of cross-correlations between antenna pairs. Thus the output is not sensitive to the total power of the receiver noise but only to correlated signals entering the antennas. The effects of gain variations are much less severe than in the case of a total-power receiver. Thermal noise from ground pickup in the sidelobes is substantially resolved.
- The antennas can be mounted with the closest spacing physically possible. There are then no serious gaps in the spatial frequencies measured, and structure can be mapped up to the width of the primary antenna beams. The apertures cannot block one another as the system tracks, as can occur for individually mounted antennas in close-spaced arrays.

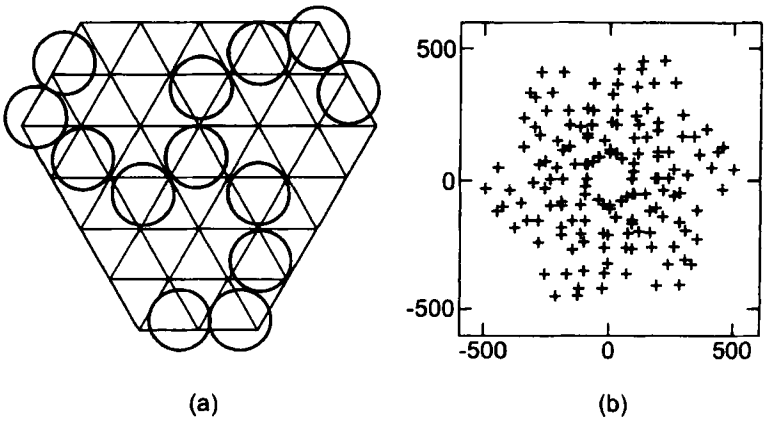


Figure 5.24 (a) Face view of the antenna platform of the Cosmic Background Imager, showing a possible configuration of the 13 antennas. (b) The corresponding antenna spacings in (u, v) coordinates for a wavelength of approximately 1 cm.

- In the array in Fig. 5.24, the whole antenna mounting platform can be rotated about an axis normal to the plane of the apertures. Thus rotation of the baselines can be controlled as desired and is independent of earth rotation. For a constant pointing direction and rotation angle relative to the sky, the pattern of (u, v) coverage remains constant as the instrument tracks. Variations in the correlator outputs with time can result from ground radiation in the sidelobes, which varies with azimuth and elevation as the array tracks. This variation can help to separate out the unwanted response.
- The close spacing of the antennas results in some cross-coupling by which spurious correlated noise is introduced into the receiving channels of adjacent antennas. However, because the antennas are rigidly mounted, the coupling does not vary constantly with time as is the case for individually mounted antennas, and it is therefore more easily calibrated out. In the CBI design the coupling is reduced to -110 to -120 dB by the use of a cylindrical shield around each antenna, and by designing the subreflector supports to minimize scattering.

At a frequency of 30 GHz, a pointing error of 1 arcsec in a 6-m baseline produces a visibility phase error of 1°. Pointing accuracy is critical, and the CBI antenna is mounted in a retractable dome to shield it from wind, which can be strong at the 5000-m-elevation site at Llano de Chajnantor, Chile.

5.7 CONCLUSIONS ON ANTENNA CONFIGURATIONS

The most accurate prediction of the performance of an array is obtained by computation of the response of the particular design to models of sources to be observed. However, in this book we are more concerned with broad comparisons of various configurations to illustrate the general considerations in array design. Some conclusions are summarized below:

- A circle centered on the (u, v) origin can be considered an optimum boundary for the distribution of measurements of visibility. Uniformity of the distribution within the circle is a further useful criterion in many circumstances. An exception is the condition where sidelobes of the synthesized beam are a serious problem, for example, in low-frequency arrays operating in conditions of source confusion, as mentioned in Chapter 1. In arrays where the scale of the configuration cannot be varied to accommodate a wide range of source dimensions, a centrally concentrated distribution allows a greater range of angular sizes to be measured with a limited number of antennas. If sensitivity to broad, low-brightness objects is important, it is preferable to have more antenna pairs with short spacings at which such sources are not highly resolved. Note that two of the largest arrays for which the antennas are not movable, the GMRT and the VLBA, each have a cluster of antennas at relatively short spacings as well as other antennas at longer spacings in order to cover a wide range of source dimensions.

- Although the effect of sidelobes on the synthesized beam can be greatly reduced by CLEAN and other image processing algorithms to be described in Chapter 11, obtaining the highest dynamic range in radio images (that is, a range of reliable intensity measurements of order 10^6 or more) requires both good spatial frequency coverage and effective image processing. Reducing holes (unsampled cells) in this coverage, which are found to be a consistent indicator of sidelobe levels, is a primary objective in array design.
- The linear array has been used for both large and small instruments and requires tracking over ± 6 h to obtain full two-dimensional coverage. It is most useful for regions of the sky within about 60° of the celestial poles and is the most economical configuration with respect to land use for road or rail track. A number of small arrays originally built as linear arrays have later been developed into crosses or T arrays.
- The equiangular Y gives the best spatial frequency coverage of the existing configurations with linear, open-ended arms. Autocorrelation functions of configurations with odd numbers of arms have higher-order symmetry than those with even numbers in which opposite arms are aligned. Curvature of the arms or random displacement of the antennas helps to smooth out the linear ridges in the (u, v) coverage (e.g., in the snapshot in Fig. 5.18). Such features are also smoothed out by hour-angle tracking and are most serious for snapshot observations.
- The circle and Reuleaux triangle provide the most uniform distributions of measurements. With uniformly spaced antennas the Reuleaux triangle provides more uniform (u, v) coverage than the circle, but varying the spacing in a quasirandom manner greatly improves both cases and reduces the difference between them; see Fig. 5.19.
- The circle can be elongated into an ellipse in the north–south direction to compensate for foreshortening toward the extremes of the declination coverage, and other configurations can be similarly extended.

5.8 OTHER CONSIDERATIONS

Up to this point we have concentrated on the configuration of antennas. Other array considerations include sensitivity, atmospheric effects, and observation of sources wider than the antenna beam. Further details of array performance that impact the design are found in chapters that follow, and a broad discussion is given by Hjellming (1989). We now briefly outline some of the more important factors.

Sensitivity

The sensitivity to a point source is proportional to the effective collecting area of an antenna multiplied by the number of antennas, that is, proportional to $n_a d^2$, where n_a is the number of antennas and d is the antenna diameter. In the case of a

source that is larger than the beams of the individual antennas, a situation that occurs mostly at millimeter wavelengths; the source can be covered by mosaicking, in which a number of pointing directions are used, as discussed in Section 11.6. The number of pointing directions is inversely proportional to the solid angle of the beam; that is, it is proportional to d^2. The sensitivity for any one pointing direction is proportional to the square root of the time spent at that direction, so overall the sensitivity to an extended source is proportional to $n_a d$. To maximize sensitivity for point sources one would maximize $n_a d^2$, but to maximize the sensitivity to surface brightness, one would maximize $n_a d$ and also use a compact configuration. Other aspects of sensitivity including system noise are discussed in Section 6.2.

A commonly used rule of thumb for the cost of an antenna is that it is proportional to d^α, where $\alpha \approx 2.7$ for values of d from a few meters to tens of meters. Thus, to obtain a large collecting area, it is cheaper to use a large number of small antennas, as long as the cost of electronics, most of which is proportional to the number of antennas, is relatively low. The cost of the correlator system is, in part, proportional to the number of antenna pairs, that is, to ${n_a}^2$, and for a wideband, multichannel correlator for spectral line observing, this can also become a significant cost item. Thus the choice of antenna size and number depends on the array parameters to be optimized and also on cost considerations.

Long Wavelengths

At low frequencies, by which we mean frequencies below a few hundred megahertz (wavelengths of $\sim$1 m and longer), the ionosphere causes serious phase fluctuations in the signals passing through it, as discussed further in Section 13.3. Calibration of this effect is particularly difficult if the excess path length varies significantly over the beam of the antennas, which can occur if the angular width of the beam is greater than that of the ionospheric irregularities. Thus, for low-frequency observations, it is advantageous to keep the beam small by using large antennas. For example, the GMRT, which operates in the range 75–1600 MHz, uses antennas of diameter 45 m (Swarup et al. 1991).

Millimeter Wavelengths

At frequencies $\sim$100 GHz and greater, antenna sizes are often reduced to the 10–20 m range to maintain surface accuracy. Nevertheless, the beamwidths are typically very narrow, for example, 25 arcsec for a 10 m antenna at 300 GHz. For observations of extended sources such as nebulae or molecular clouds under such conditions, the mosaicking technique mentioned above becomes important; see Section 11.6. Since sensitivity is then proportional to $n_a d$, optimizing the performance points even more strongly toward reducing d and increasing n_a than in the case of observing sources smaller than the beamwidth. With more antennas it is possible to obtain satisfactory (u, v) coverage with less tracking time. This also helps by reducing the need for observations at low angles of elevation for which atmospheric effects are most severe.

An important requirement in mosaicking is measuring the visibility at values of u and v smaller than the antenna diameter. This is possible, but observations at the shortest practicable baselines are necessary. The closest spacing is usually determined by the condition that neighboring antennas should be able to be pointed in different directions without danger of collision. Minimizing this spacing, which depends on the focal ratio and the design of the mount, is a consideration for millimeter-wavelength antennas. The minimum practical spacing for individually mounted antennas is about $1.25d$ (Welch et al. 1996). Some considerations of mosaicking requirements on array design are discussed by Cornwell, Holdaway, and Uson (1993).

BIBLIOGRAPHY

Balanis, C. A., *Antenna Theory Analysis and Design*, Wiley, New York, 1982, 1997.

Collin, R. E., *Antennas and Radiowave Propagation*, McGraw-Hill, New York, 1985.

Imbriale, W. A. and M. Thorburn, Eds., *Proc. IEEE*, Special Issue on Radio Telescopes, **82**, 633–823, 1994.

Johnson, R. C. and H. Jasik, Eds., *Antenna Engineering Handbook*, McGraw-Hill, New York, 1984.

Love, A. W., Ed., *Reflector Antennas*, IEEE Press, The Institute of Electrical and Electronics Engineers, New York, 1978.

Milligan, T. A., *Modern Antenna Design*, McGraw-Hill, New York, 1985.

Stutzman, W. L. and G. A. Thiele, *Antenna Theory and Design*, 2nd ed., Wiley, New York, 1998.

REFERENCES

Arsac, J., Nouveau Réseau Pour l'Observation Radioastronomique de la Brillance sur le Soleil à 9530 Mc/s, *C. R. Acad. Sci.*, **240**, 942–945, 1955.

Baars, J. W. M. and B. G. Hooghoudt, The Synthesis Radio Telescope at Westerbork, General Layout and Mechanical Aspects, *Astron. Astrophys.*, **31**, 323–331, 1974.

Blythe, J. H., A New Type of Pencil Beam Aerial for Radio Astronomy, *Mon. Not. R. Astron. Soc.*, **117**, 644–651, 1957.

Bracewell, R. N., Interferometry of Discrete Sources, *Proc. IRE*, **46**, 97–105, 1958.

Bracewell, R. N., Interferometry and the Spectral Sensitivity Island Diagram, *IRE Trans. Antennas Propag.*, **AP-9**, 59–67, 1961.

Bracewell, R. N., Radio Astronomy Techniques, in *Handbuch der Physik*, Vol. 14, S. Flugge, Ed., Springer-Verlag, Berlin, 1962, pp. 42–129.

Bracewell, R. N., *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 2000 (earlier eds. 1965, 1978).

Bracewell, R. N., Optimum Spacings for Radio Telescopes with Unfilled Apertures, in *Progress in Scientific Radio*, Report on the 15th General Assembly of URSI, Publication 1468 of the National Academy of Sciences, Washington, DC, 1966, pp. 243–244.

Bracewell, R. N., The Fast Hartley Transform, *Proc. IEEE*, **72**, 1010–1018, 1984.

Bracewell, R. N., *Two-Dimensional Imaging*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

Bracewell, R. N., R. S. Colvin, L. R. D'Addario, C. J. Grebenkemper, K. M. Price, and A. R. Thompson, The Stanford Five-Element Radio Telescope, *Proc. IEEE*, **61**, 1249–1257, 1973.

Bracewell, R. N. and J. A. Roberts, Aerial Smoothing in Radio Astronomy, *Aust. J. Phys.*, **7**, 615–640, 1954.

Bracewell, R. N. and A. R. Thompson, The Main Beam and Ringlobes of an East-West Rotation-Synthesis Array, *Astrophys. J.*, **182**, 77–94, 1973.

Braude, S. Ya., A. V. Megn, B. P. Ryabov, N. K. Sharykin, and I. N. Zhouck, Decametric Survey of Discrete Sources in the Northern Sky, *Astrophys. Space Sci.*, **54**, 3–36, 1978.

Brigham, E. O., *The Fast Fourier Transform and Its Applications*, Prentice Hall, Englewood Cliffs, N, 1988.

Burke, B. F., Orbiting VLBI: A Survey, in *VLBI and Compact Radio Sources*, R. Fanti, K. Kellermann, and G. Setti, Eds., Reidel, Dordrecht, Holland, 1984.

Chow, Y. L., On Designing a Supersynthesis Antenna Array, *IEEE Trans. Antennas Propag.*, **AP-20**, 30–35, 1972.

Chu, T.-S. and R. H. Turrin, Depolarization Effects of Offset Reflector Antennas, *IEEE Trans. Antennas Propag.*, **AP-21**, 339–345, 1973.

Cornwell, T. J., A Novel Principle for Optimization of the Instantaneous Fourier Plane Coverage of Correlation Arrays, *IEEE Trans. Antennas Propag.*, **36**, 1165–1167, 1988.

Cornwell, T. J., M. A. Holdaway, and J. M. Uson, Radio-Interferometric Imaging of Very Large Objects: Implications for Array Design, *Astron. Astrophys.*, **271**, 697–713, 1993.

Erickson, W. C., M. J. Mahoney, and K. Erb, *Astrophys. J. Suppl.*, **50**, 403–420, 1982.

Frater, R. H., J. W. Brooks, and J. B. Whiteoak, The Australia Telescope—Overview, *Proc. IREE, Aust.*, **12**, 102–112, 1992.

Hamaker, J. P., J. D. O'Sullivan, and J. E. Noordam, Image Sharpness, Fourier Optics, and Redundant Spacing Interferometry, *J. Opt. Soc. Am.*, **67**, 1122–1123, 1977.

Hirabayashi, H., and 52 coauthors, Overview and Initial Results of the Very Long Baseline Interferometry Space Observatory Program, *Science*, **281**, 1825–1829, 1998.

Hjellming, R. M., The Design of Aperture Synthesis Arrays, *Synthesis Imaging in Radio Astronomy*,R. A, Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific. Conf. Ser., **6**, 477–500, 1989.

Högbom, J. A. and W. N. Brouw, The Synthesis Radio Telescope at Westerbork, Principles of Operation, Performance and Data Reduction, *Astron. Astrophys.*, **33**, 289–301, 1974.

Ingalls, R. P., J. Antebi, J. A. Ball, R. Barvainis, J. F. Cannon, J. C. Carter, P. J. Charpentier, B. E. Corey, J. W. Crowley, K. A. Dudevoir, M. J. Gregory, F. W. Kan, S. M. Milner, A. E. E. Rogers, J. E. Salah, and M. S. Zarghamee, Upgrading of the Haystack Radio Telescope for Operation at 115 GHz, *Proc. IEEE*, **82**, 742–755, 1994.

Ishiguro, M., Minimum Redundancy Linear Arrays for a Large Number of Antennas, *Radio Sci.*, **15**, 1163–1170, 1980.

Keto, E., The Shapes of Cross-Correlation Interferometers, *Astrophys. J.*, **475**, 843–852, 1997.

Kogan, L., Level of Negative Sidelobes in an Array Beam, *Pub. Astron. Soc. Pacific*, **111**, 510–511, 1999.

Koles, W. A., R. G. Frehlich, and M. Kojima, Design of a 74-MHz Antenna for Radio Astronomy, *Proc. IEEE*, **82**, 697–704, 1994.

Lawrence, C. R., T. Herbig, and A. C. S. Readhead, Reduction of Ground Spillover in the Owens Valley 5.5-m Telescope, *Proc. IEEE*, **82**, 763–767, 1994.

Leech, J., On Representation of 1, 2 . . . , *n* by Differences, *J. London Math. Soc.*, **31**, 160–169, 1956.

Mathur, N. C., A Pseudodynamic Programming Technique for the Design of Correlator Supersynthesis Arrays, *Radio Sci.*, **4**, 235–244, 1969.

Mayer, C. E., D. T. Emerson, and J. H. Davis, Design and Implementation of an Error-Compensating Subreflector for the NRAO 12-m Radio Telescope, *Proc. IEEE*, **82**, 756–762, 1994

Mills, B. Y., Cross-Type Radio Telescopes, *Proc. IRE Aust.*, **24**, 132–140, 1963.

Moffet, A. T., Minimum-Redundancy Linear Arrays, *IEEE Trans. Antennas Propag.*, **AP-16**, 172–175, 1968.

Moran, J. M., The Submillimeter Array, in *Advanced Technology MMW, Radio and Terahertz Telescopes*, T. G. Phillips, Ed., *Proc. SPIE*, **3357**, 208–219, 1998.

Napier, P. J., D. S. Bagri, B. G. Clark, A. E. E. Rogers, J. D. Romney, A. R. Thompson, and R. C. Walker, The Very long Baseline Array, *Proc. IEEE*, **82**, 658–672, 1994.

Napier, P. J., A. R. Thompson, and R. D. Ekers, The Very Large Array: Design and Performance of a Modern Synthesis Radio Telescope, *Proc. IEEE*, **71**, 1295–1320, 1983.

National Radio Astronomy Observatory (NRAO), *A Proposal for a Very Large Array Radio Telescope*, National Radio Astronomy Observatory, Green Bank, WV, Vol. 1, Jan. 1967; Vol. 3, Jan. 1969.

Oppenheim, A. V. and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975, Ch. 3.

Padin, S., J. K. Cartwright, B. S. Mason, T. J. Pearson, A. C. S. Readhead, M. C. Shepherd, J. Sievers, P. S. Udomprasert, W. L. Holzapfel, S. T. Myers, J. E. Carlstrom, E. M. Leitch, M. Joy, L. Bronfman, and J. May, First Intrinsic Anisotropy Observations with the Cosmic Background Imager, *Astrophys. J. Letters*, **549**, L1–L5, 2001.

Papoulis, A., *Signal Analysis*, McGraw-Hill, New York, 1977, p. 74.

Preston, R. A., B. F. Burke, R. Doxsey, J. F. Jordan, S. H. Morgan, D. H. Roberts, I. I. Shapiro, The Future of VLBI Observations in Space, in *Very Long Baseline Interferometry Techniques*, F. Biraud, Ed., Cepadues, Toulouse, France, 1983, pp. 417–431.

Rabiner, L. R. and B. Gold, *Theory and Application of Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975, p. 50.

Rademacher, H. and O. Toeplitz, *The Enjoyment of Mathematics*, Princeton Univ. Press, Princeton, NJ, 1957.

Raimond, E. and R. Genee, Eds., *The Westerbork Observatory, Continuing Adventure in Radio Astronomy*, Kluwer, Dordrecht, 1996.

Rudge, A. W., and N. A. Adatia, Offset-Parabolic-Reflector Antennas: A Review, *Proc. IEEE*, **66**, 1592–1618, 1978.

Ruze, J., Antenna Tolerance Theory—A Review, *Proc. IEEE*, **54**, 633–640, 1966.

Ryle, M., The New Cambridge Radio Telescope, *Nature*, **194**, 517–518, 1962.

Ryle, M. and A. Hewish, The Synthesis of Large Radio Telescopes, *Mon. Not. R. Astron. Soc.*, **120**, 220–230, 1960.

Ryle, M., A. Hewish, and J. R. Shakeshaft, The Synthesis of Large Radio Telescopes by the Use of Radio Interferometers, *IRE Trans. Antennas Propag.*, **7**, S120–S1204, 1959.

Seielstad, G. A., G. W. Swenson Jr. and J. C. Webber, A New Method of Array Evaluation Applied to Very Long Baseline Interferometry, *Radio Sci.*, **14**, 509–517, 1979.

Swarup, G., S. Ananthakrishnan, V. K. Kapahi, A. P. Rao, C. R. Subrahmanya, and V. K. Kulkarni, The Giant Metre-Wave Radio Telescope, *Current Science* (Current Science Association and Indian Academy of Sciences), **60**, 95–105, 1991.

Swenson, G. W., Jr. and K. I. Kellermann, An Intercontinental Array—A Next-Generation Radio Telescope, *Science*, **188**, 1263–1268, 1975.

Swenson, G. W., Jr. and N. C. Mathur, The Circular Array in the Correlator Mode, *Proc. IREE Aust.*, **28**, 370–374, 1967.

Thomasson, P., MERLIN, *Quat. J. Royal Astron. Soc.*, **27**, 413–431, 1986.

Thompson, A. R. and R. N. Bracewell, Interpolation and Fourier Transformation of Fringe Visibilities, *Astron. J.*, **79**, 11–24, 1974.

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

Unser, M., Sampling—50 Years After Shannon, *Proc. IEEE*, **88**, 569–587, 2000.

Walker, R. C., VLBI Array Design, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Univ. Press, Cambridge, UK, 1984, pp. 53–65.

Welch, W. J. and 36 coauthors, The Berkeley–Illinois–Maryland Association Millimeter Array, *Pub. Astron. Soc. Pacific*, **108**, 93–103, 1996.

Wild, J. P., A New Method of Image Formation with Annular Apertures and an Application in Radio Astronomy, *Proc. R. Soc. A*, **286**, 499–509, 1965.

Wild, J. P., Ed., *Proc. IRE Aust.*, Special Issue on the Culgoora Radioheliograph, Vol. 28, No. 9, 1967.

Williams, W. F., High Efficiency Antenna Reflector, *Microwave J.*, **8**, 79–82, 1965 (reprinted in Love (1978); see Bibliography).

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

6 Response of the Receiving System

This chapter is concerned with the response of the receiving system that accepts the signals from the antennas, amplifies and filters them, and measures the cross-correlations for the various antenna pairs. We show how the basic parameters of the system affect the output. Some of the effects were introduced in earlier chapters, and here we present a more detailed development that leads to consideration of system design in Chapters 7 and 8.

6.1 FREQUENCY CONVERSION, FRINGE ROTATION, AND COMPLEX CORRELATORS

Frequency Conversion

In practically all receiving systems in radio astronomy the frequencies of the signals received at the antennas are changed by mixing with a local oscillator signal. This feature, referred to as *frequency conversion* (or heterodyne frequency conversion), enables the major part of the signal processing to be performed at intermediate frequencies that are most appropriate for amplification, transmission, filtering, delaying, recording, and similar processes.

Frequency conversion takes place in a mixer, in which the signal to be converted plus a local oscillator waveform are applied to a circuit element with a nonlinear voltage–current response. This element may be a diode as shown in Fig. 6.1a. The current i through the diode can be expressed as a power series in the applied voltage V:

$$i = a_0 + a_1 V + a_2 V^2 + a_3 V^3 + \cdots . \tag{6.1}$$

Now let V consist of the sum of a local oscillator voltage $b_1 \cos(2\pi \nu_{\mathrm{LO}} t + \theta_{\mathrm{LO}})$ and a signal, of which one Fourier component is $b_2 \cos(2\pi \nu_s t + \phi_s)$. The second-order term in V then gives rise to a product in the mixer output of the form

$$\begin{aligned} & b_1 \cos(2\pi \nu_{\mathrm{LO}} t + \theta_{\mathrm{LO}}) \\ & \times b_2 \cos(2\pi \nu_s t + \phi_s) = \tfrac{1}{2} b_1 b_2 \cos\left[2\pi(\nu_s + \nu_{\mathrm{LO}})t + \phi_s + \theta_{\mathrm{LO}}\right] \\ & \qquad + \tfrac{1}{2} b_1 b_2 \cos\left[2\pi(\nu_s - \nu_{\mathrm{LO}})t + \phi_s - \theta_{\mathrm{LO}}\right] . \end{aligned} \tag{6.2}$$

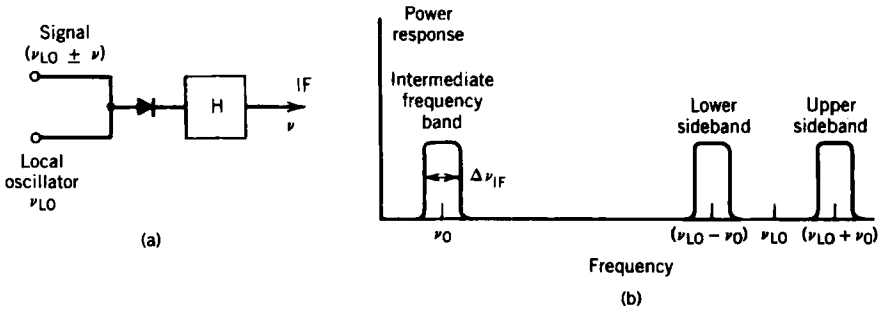


Figure 6.1 Frequency conversion in a radio receiving system. (a) Simplified diagram of a mixer and a filter H that defines the intermediate-frequency (IF) band. The nonlinear element shown is a diode. (b) Signal spectrum showing upper and lower sidebands that are converted to the intermediate frequency. Frequency ν_0 is the center of the IF band.

Thus the current through the diode contains components at the sum and difference of ν_s and ν_{LO}. Other terms in (6.1) lead to other components, such as $3\nu_{LO} \pm \nu_s$, but the filter H shown in Fig. 6.1 passes only the wanted output spectrum, and with proper design unwanted combinations can be prevented from falling within the filter passband. Usually the signal voltage is much smaller than the local oscillator voltage, so harmonics and intermodulation products (i.e., spurious signals that arise as a result of cross-products of different frequency components within the input signal band) are small compared with the wanted terms containing ν_{LO}.

In most cases of frequency conversion the signal frequency is being reduced, and the second term on the right-hand side in Eq. (6.2) is the important one. The filter H then defines an intermediate-frequency (IF) band centered on ν_0, as shown in Fig. 6.1b. Signals from within the bands centered on $\nu_{LO} - \nu_0$ and $\nu_{LO} + \nu_0$ are converted and admitted by the filter. These bands are known as the *lower* and *upper sidebands*, as shown, and if only a single sideband is wanted, the other can often be removed by a suitable filter inserted before the mixer. In some cases both sidebands are accepted, resulting in a double-sideband response.

Response of a Single-Sideband System

Figure 6.2 shows a basic receiving system for two antennas, m and n, of a synthesis array. Here we are interested in the effects of frequency conversion that were omitted from the earlier discussions. The time difference τ_g between the arrival at the antennas of the signals from a radio source varies continuously as the earth rotates and the antennas track the source across the sky. An instrumental delay τ_i is continuously adjusted to compensate for the geometric delay τ_g, so that the signals arrive simultaneously at the correlator. The receiving channels through which the signals pass contain amplifiers and filters, the overall amplitude (voltage) responses of which are $H_m(\nu)$ and $H_n(\nu)$ for antennas m and n. Here ν represents a frequency at the correlator input; the corresponding frequency at the antenna is

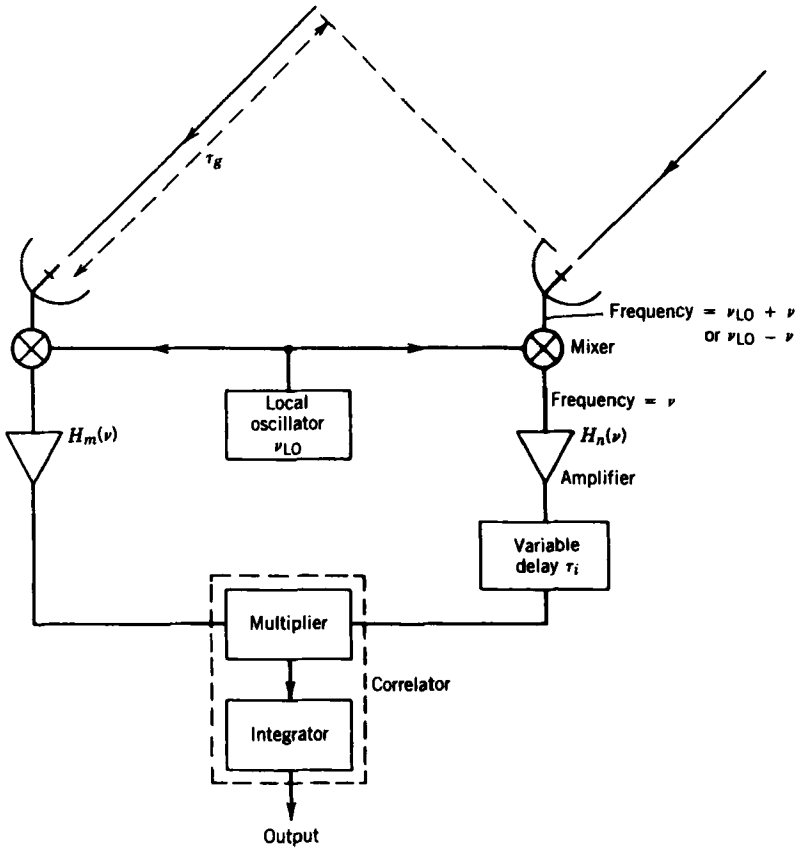


Figure 6.2 Basic receiving system for two antennas of a synthesis array. The variable delay τ_i is continuously adjusted under computer control to compensate for the geometric path delay τ_g. The frequency response functions $H_m(\nu)$ and $H_n(\nu)$ represent the overall bandpass characteristics of the amplifiers and filters in the signal channels.

$\nu_{LO} \pm \nu$. The voltage waveforms that are processed by the receiving system result from cosmic noise and system noise; we consider the usual case in which these processes are constant across the receiver passband. The spectra at the correlator inputs are thus determined mainly by the response of the receiving system. Let ϕ_m be the phase change in the signal path through antenna m resulting from τ_g and the local oscillator phase, and let ϕ_n be the corresponding phase change in the signal for the path through antenna n, including τ_i. ϕ_m and ϕ_n, together with the instrumental phase resulting from the amplifiers and filters, represent the phases of the cosmic signal at the correlator inputs. Negative values of these parameters indicate phase lag (signal delay). The response to a source for which the visibility is $\mathcal{V}(u, v) = |\mathcal{V}|e^{j\phi_v}$ is most easily obtained by returning to Eq. (3.5) and replacing the phase difference $2\pi \mathbf{D}_\lambda \cdot \mathbf{s}_0$ by the general term $\phi_n - \phi_m$. Then the response at the correlator output resulting from a frequency band of width $d\nu$ can

be written as

$$dr = \mathcal{R}e\{A_0|\mathcal{V}|H_m(\nu)H_n^*(\nu)e^{j(\phi_n-\phi_m-\phi_\nu)}\,d\nu\}, \tag{6.3}$$

and the response from the full system passband is

$$r = \mathcal{R}e\left\{A_0|\mathcal{V}|\int_{-\infty}^{\infty} H_m(\nu)H_n^*(\nu)e^{j(\phi_n-\phi_m-\phi_\nu)}\,d\nu\right\}, \tag{6.4}$$

where we have included both positive and negative frequencies in the integral and assumed that $\mathcal{V}$ does not vary significantly over the observing bandwidth. Equation (6.4) represents the real part of the complex cross-correlation, and we explain how to obtain both real and imaginary parts from the correlator later in this section.

Upper-Sideband Reception

For upper-sideband reception a filter or amplifier at the receiver input selects frequencies in a band defined by the correlator input spectrum (frequency ν) plus $\nu_{\rm LO}$. We now express the phases ϕ_m and ϕ_n in terms of the phases encountered by the signals in Fig. 6.2. The signal entering antenna m traverses the geometric delay τ_g at a frequency $\nu_{\rm LO}+\nu$, and thus suffers a phase shift $2\pi(\nu_{\rm LO}+\nu)\tau_g$. At the mixer its phase is also decreased by the local oscillator phase θ_m. Thus we obtain

$$\phi_m(\nu) = -2\pi(\nu_{\rm LO}+\nu)\tau_g - \theta_m. \tag{6.5}$$

The phase of the signal entering antenna n is decreased by the local oscillator phase θ_n, and the signal then traverses the instrumental delay τ_i at a frequency ν, thus suffering a shift $2\pi\nu\tau_i$. The total phase shift for antenna n is

$$\phi_n(\nu) = -2\pi\nu\tau_i - \theta_n. \tag{6.6}$$

From Eqs. (6.4), (6.5), and (6.6) the correlator output is

$$r_u = \mathcal{R}e\left\{A_0|\mathcal{V}|e^{j[2\pi\nu_{\rm LO}\tau_g+(\theta_m-\theta_n)-\phi_\nu]}\int_{-\infty}^{\infty} H_m(\nu)H_n^*(\nu)e^{j2\pi\nu\Delta\tau}\,d\nu\right\}. \tag{6.7}$$

The real part of the integral in Eq. (6.7) is one-half the Fourier transform of the (hermitian) cross power spectrum $H_m(\nu)H_n^*(\nu)$ with respect to the delay compensation error, $\Delta\tau = \tau_g - \tau_i$, which introduces a linear phase slope across the band*. We assume that $\mathcal{V}$ does not vary significantly over the observing bandwidth. For example, if the IF passbands are rectangular with center frequency ν_0, width $\Delta\nu_{\rm IF}$, and identical phase responses, then for positive frequencies,

*Here we assume that the source is sufficiently close to the center of the field being mapped that the condition $\Delta\tau = 0$ maintains zero delay error. The effect of the variation of the delay error across a wider field of view is considered in Section 6.3.

$$|H_m(\nu)| = |H_n(\nu)| = \begin{cases} H_0, & |\nu - \nu_0| < \dfrac{\Delta\nu_{\text{IF}}}{2}, \\ 0, & |\nu - \nu_0| > \dfrac{\Delta\nu_{\text{IF}}}{2}. \end{cases} \tag{6.8}$$

Using the equality in Eq. (A3.6) of Appendix 3.1 for the hermitian function $H_m H_n$, we can write

$$\begin{aligned} \int_{-\infty}^{\infty} H_m(\nu) H_n^*(\nu) e^{j2\pi\nu\Delta\tau}\, d\nu &= 2\mathcal{R}e \left\{ \int_{\nu_0-(\Delta\nu_{\text{IF}}/2)}^{\nu_0+(\Delta\nu_{\text{IF}}/2)} H_0^2 e^{j2\pi\nu\Delta\tau}\, d\nu \right\} \\ &= 2H_0^2 \Delta\nu_{\text{IF}} \left[\frac{\sin(\pi\Delta\nu_{\text{IF}}\Delta\tau)}{\pi\Delta\nu_{\text{IF}}\Delta\tau} \right] \cos 2\pi\nu_0\Delta\tau. \end{aligned} \tag{6.9}$$

In the general case we define an instrumental gain factor $G_{mn} = |G_{mn}|e^{j\phi_G}$ as follows:

$$\begin{aligned} A_0 \int_{-\infty}^{\infty} H_m(\nu) H_n^*(\nu) e^{j2\pi\nu\Delta\tau}\, d\nu &= G_{mn}(\Delta\tau) e^{j2\pi\nu_0\Delta\tau} \\ &= |G_{mn}(\Delta\tau)| e^{j(2\pi\nu_0\Delta\tau + \phi_G)}. \end{aligned} \tag{6.10}$$

The variation of G_{mn} with $\Delta\tau$ causes the delay pattern effect discussed in earlier chapters. The phase ϕ_G results from the difference in the phase responses of the amplifiers and filters. The local oscillator phases θ_m and θ_n are not included within the general instrumental phase term ϕ_G because they enter into the upper and lower sidebands with different signs.

Substituting Eq. (6.10) into Eq. (6.7), we obtain for upper-sideband reception

$$r_u = |\mathcal{V}||G_{mn}(\Delta\tau)| \cos\left[2\pi(\nu_{\text{LO}}\tau_g + \nu_0\Delta\tau) + (\theta_m - \theta_n) - \phi_v + \phi_G\right]. \tag{6.11}$$

The term $2\pi\nu_{\text{LO}}\tau_g$ in the cosine function results in a quasisinusoidal oscillation as the source moves through the fringe pattern. The phase of this oscillation depends on the delay error $\Delta\tau$, the relative phases of the local oscillator signals, the phase responses of the signal channels, and the phase of the visibility function. The frequency of the output oscillation $\nu_{\text{LO}}\, d\tau_g/dt$ is often referred to as the *natural fringe frequency*. The oscillations result because the signals traverse the delays τ_g and τ_i at different frequencies, that is, at the input radio frequency for τ_g and at the intermediate frequency for τ_i, and these two frequencies differ by ν_{LO}. Thus, even if these two delays are identical they introduce different phase shifts, and they increase or decrease progressively as the earth rotates.

Lower-Sideband Reception

Consider now the situation where the frequencies accepted from the antenna are those in the lower sideband, at ν_{LO} minus the correlator input frequencies. The

phases are

$$\phi_m = 2\pi(\nu_{\text{LO}} - \nu)\tau_g + \theta_m \tag{6.12}$$

and

$$\phi_n = -2\pi\nu\tau_i + \theta_n. \tag{6.13}$$

The signs of these terms and of ϕ_v differ from those in the upper-sideband case because increasing the phase of the signal at the antenna here decreases the phase at the correlator. The expression for the correlator output is

$$r_\ell = \mathcal{R}e\left\{A_0|\mathcal{V}|e^{-j[2\pi\nu_{\text{LO}}\tau_g+(\theta_m-\theta_n)-\phi_v]}\int_{-\infty}^{\infty} H_m(\nu)H_n^*(\nu)e^{j2\pi\,\Delta\tau}\,d\nu\right\}. \tag{6.14}$$

Proceeding as in the upper-sideband case, we obtain

$$r_\ell = |\mathcal{V}||G_{mn}(\Delta\tau)|\cos\left[2\pi(\nu_{\text{LO}}\tau_g - \nu_0\,\Delta\tau) + (\theta_m - \theta_n) - \phi_v - \phi_G\right]. \tag{6.15}$$

Multiple Frequency Conversions

In an operational system the signals may undergo several frequency conversions between the antennas and the correlators. Operation with multiple frequency conversions is essentially the same as with the systems considered above. A frequency conversion in which the output is at the lower sideband (i.e., the local oscillator frequency minus the input frequency) results in a reversal of the signal spectrum in which frequencies at the high end at the input appear at the low end at the output, and vice versa. If there is no net reversal (that is, an even number of lower-sideband conversions), Eq. (6.11) applies, except that ν_{LO} must be replaced by a combination of local oscillator frequencies sometimes known as the signed-sum of the local oscillator frequencies because some frequencies enter it with positive signs and some with negative ones. Similarly, the oscillator phase terms θ_m and θ_n are replaced by corresponding combinations of oscillator phases. If there is a net reversal of the frequency band, Eq. (6.15) applies with similar modifications.

Delay Tracking and Fringe Rotation

Adjustment of the compensating delay τ_i of Fig. 6.2 is usually accomplished under computer control, the required delay being a function of the antenna positions and the position of the phase center of the field under observation. This can be achieved by designating one antenna of the array as the delay reference and adjusting the instrumental delays of other antennas so that, for an incoming wavefront from the phase reference direction, the signals intercepted by the different antennas all arrive at the correlator simultaneously.

To control the frequency of the sinusoidal fringe variations in the correlator output, a continuous phase change can be inserted into one of the local oscillator

signals. Equations (6.11) and (6.15) show that the fringe frequency can be reduced to zero by causing $\theta_m - \theta_n$ to vary at a rate that maintains constant, modulo 2π, the term $[2\pi \nu_{LO} \tau_g + (\theta_m - \theta_n)]$. This requires adding a frequency $2\pi \nu_{LO}(d\tau_g/dt)$ to θ_n or subtracting it from θ_m. Note that $d\tau_g/dt$ can be evaluated from Eq. (4.9) in which w, the third component of the interferometer baseline, is equal to $c\tau_g$ measured in wavelengths; for example, for an east–west antenna spacing of 1 km, the maximum value of $d\tau_g/dt$ is 2.42×10^{-10}, so the fringe frequencies are generally small compared with the radio frequencies involved. Reduction of the output frequency reduces the quantity of data to be processed, since each correlator output must be sampled at least twice per cycle of the output frequency (the Nyquist rate) to preserve the information, as discussed in Section 8.2. With antenna spacings required for angular resolution of millarcsecond order, which occur in VLBI, the natural fringe frequency, $\nu_{LO}\, d\tau_g/dt$, can exceed 10 kHz. For an array with more than one antenna pair it is possible to reduce each output frequency to the same fraction of its natural frequency, or to zero. Reduction to zero frequency is generally the preferred practice and is often referred to as *fringe stopping*. Some special technique, such as the use of a complex correlator, described in the following subsection, is then required to extract the amplitude and phase of the output.

Simple and Complex Correlators

A method of measuring the amplitude and phase of the correlator output signal when the fringe frequency at the correlator output is reduced to zero is shown in Fig. 6.3. Two correlators are used: one that multiplies the signals in the manner considered above, and another that has a quadrature phase-shift network in one input. This network shifts the phase of each frequency component in the input band by $\pi/2$, and the output is thus the Hilbert transform of the input. For signals of finite bandwidth the phase shift is not equivalent to a delay. The phase shift can also be effected by feeding the signal into two separate mixers and converting it with two local oscillators in phase quadrature. The output of the second correlator can be obtained by replacing $H_m(\nu)$ by $H_m(\nu)e^{-j\pi/2}$. From Eq. (6.10) the result

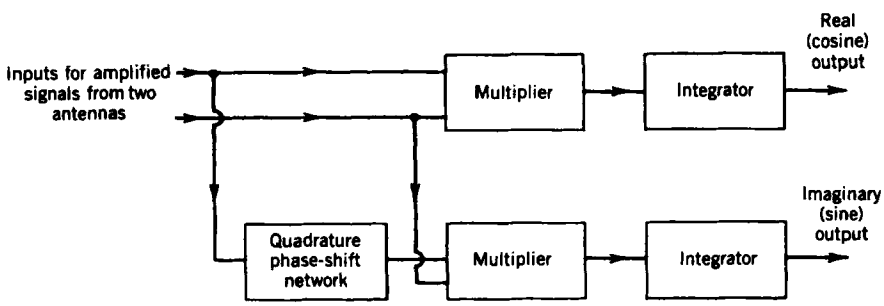


Figure 6.3 Use of two correlators to measure the real and imaginary parts of the visibility. This device is called a complex correlator.

is to add $-\pi/2$ to ϕ_G, and thus in Eqs. (6.11) and (6.15) the cosine function is replaced by $\pm$sine. Another way of comparing the two correlator outputs in Fig. 6.3 is to note that the output of the real correlator, omitting constant factors, is

$$r_{\text{real}} = \mathcal{R}e\left\{\mathcal{V}\int_{-\infty}^{\infty} H_m(\nu)H_n^*(\nu)\,d\nu\right\} = \mathcal{R}e\{\mathcal{V}\}\int_{-\infty}^{\infty} H_m(\nu)H_n^*(\nu)\,d\nu, \quad (6.16)$$

where the integral is real since $H_m(\nu)$ and $H_n^*(\nu)$ are hermitian (real part even, imaginary part odd), and thus $H_m(\nu)H_n^*(\nu)$ is hermitian. The output of the imaginary correlator is proportional to

$$r_{\text{imag}} = \mathcal{R}e\left\{\mathcal{V}\int_{-\infty}^{\infty} H_m(\nu)H_n^*(\nu)e^{-j\pi/2}d\nu\right\} = \mathcal{I}m\{\mathcal{V}\}\int_{-\infty}^{\infty} H_m(\nu)H_n^*(\nu)\,d\nu. \quad (6.17)$$

Thus the two outputs respond to the real and imaginary parts of the visibility $\mathcal{V}$.

The combination of two correlators and the quadrature network is usually referred to as a *complex correlator*, and the two outputs as the cosine and sine, or real and imaginary, outputs. (When necessary to emphasize the distinction we can refer to a single multiplier and integrator as a *simple* or *single-multiplier correlator*.) For continuum observations the compensating delay is adjusted so that $\Delta\tau = 0$ and the fringe rotation maintains the condition $2\pi\nu_{\text{LO}}\tau_i + (\theta_m - \theta_n) = 0$. Thus the cosine and sine outputs represent the real and imaginary parts of $G_{mn}\mathcal{V}(u, v)$. With the use of the complex correlator, the rotation of the earth, which sweeps the fringe pattern across the source, is no longer a necessary feature in the measurement of visibility. An important feature of the complex correlator is that the noise fluctuations in the cosine and sine outputs are independent, as discussed in Section 6.2 [see text following Eq. (6.50)].

Spectral correlator systems, in which a number of correlators are used to measure the correlation as a function of time offset or "lag" [i.e., τ in Eq. (3.27)], are discussed in Section 8.7. The correlation as a function of τ measured using correlators with a quadrature phase shift in one input is the Hilbert transform of the same quantity measured without the quadrature phase shifts (Lo et al. 1984). Thus, unlike the case where the correlation is measured for $\tau = 0$ only, here it is only necessary to use simple correlators since sine outputs would provide no additional information. See also case 8 in the text associated with Table 6.1.

Response of a Double-Sideband System

A double-sideband receiving system is one in which both the upper- and lower-sideband responses are accepted. From Eqs. (6.11) and (6.15) the output is

$$\begin{aligned} r_d = r_u + r_\ell = 2|\mathcal{V}||G_{mn}(\Delta\tau)|\cos(2\pi\nu_0\,\Delta\tau + \phi_G) \\ \times\cos\left[2\pi\nu_{\text{LO}}\tau_g + (\theta_m - \theta_n) - \phi_v\right]. \end{aligned} \quad (6.18)$$

There is a significant difference from the single-sideband cases. The phase of the fringe-frequency term, which is the cosine function containing the term $2\pi\nu_{\mathrm{LO}}\tau_g$, is no longer dependent on $\Delta\tau$ or ϕ_G, but instead these quantities appear in the term that controls the fringe amplitude:

$$|G_{mn}(\Delta\tau)|\cos(2\pi\nu_0\Delta\tau + \phi_G). \tag{6.19}$$

If the delay τ_i is held constant, $\Delta\tau$ varies continuously, resulting in cosinusoidal modulation of the fringe oscillations through the cosine term in (6.19). Also, as shown in Fig. 6.4, the cross-correlation (fringe amplitude) falls off more rapidly because of the cosine term in (6.19) than it does in the single-sideband case, in which it depends only on $G_{mn}(\Delta\tau)$. The required precision in matching the geometric and instrumental delays is correspondingly increased. The lack of dependence of the fringe phase on the phase response of the signal channel occurs because the latter has equal and opposite effects on the signals from the two sidebands.

The response of a double-sideband system with a complex correlator is given by Eq. (6.18) for the cosine output, and for the sine output it is obtained by replacing ϕ_G by $\phi_G - \pi/2$:

$$\begin{aligned}(r_d)_{\mathrm{sine}} = 2|\mathcal{V}||G_{mn}(\Delta\tau)|\sin(2\pi\nu_0\Delta\tau + \phi_G)\\ \times\cos[2\pi\nu_{\mathrm{LO}}\tau_g + (\theta_m - \theta_n) - \phi_v].\end{aligned} \tag{6.20}$$

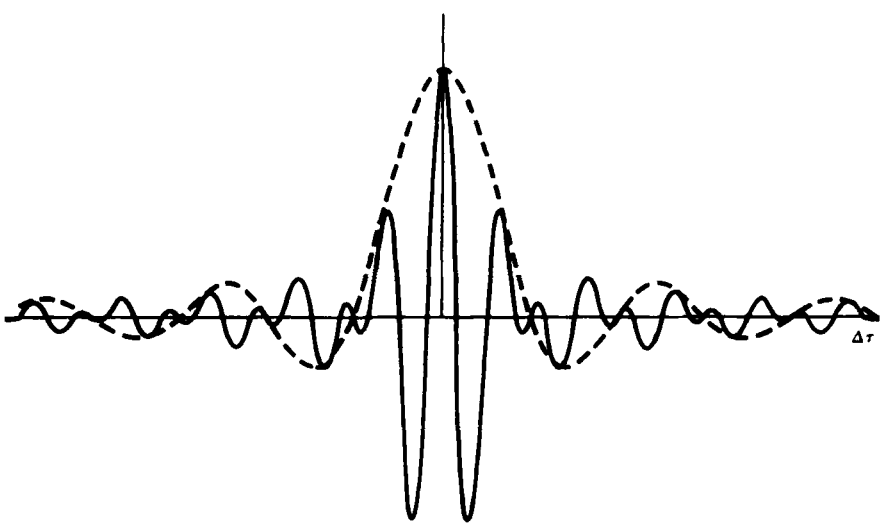


Figure 6.4 Example of the variation of the fringe amplitude as a function of $\Delta\tau$ for a double-sideband system (full line). In this case the centers of the two sidebands are separated by three times the IF bandwidth, that is, $\nu_0 = 1.5\Delta\nu_{\mathrm{IF}}$, and the IF response is rectangular. The broken line shows the equivalent function for a single-sideband system with the same IF response.

If the term $2\pi\nu_0\,\Delta\tau+\phi_G$ is adjusted to maximize either the real output [Eq. (6.18)] or the imaginary output [Eq. (6.20)], the other will be zero. Thus for continuum observations in which the signal is of equal strength in both sidebands, the complex correlator offers no increase in sensitivity. However, it can be useful for observations in the sideband-separation mode described later.

To help visualize the difference between single- and double-sideband interferometer systems, Fig. 6.5 illustrates the correlator outputs in the complex plane. The single-sideband case is shown in Fig. 6.5a. The output of the complex correlator is represented by the vector **r**. If the fringes are not stopped, the vector **r** rotates through 2π each time the geometric delay τ_g changes by one wavelength. (That is, one wavelength at the local oscillator frequency if the instrumental delay is tracking the geometric delay.) The projections of the radial vector on the real and imaginary axes indicate the real and imaginary outputs of the complex correlator, which are two fringe-frequency sinusoids in phase quadrature. If the fringes are stopped, **r** remains fixed in position angle. Figure 6.5b represents the double-sideband case. Vectors $\mathbf{r}_u$ and $\mathbf{r}_\ell$ represent the output components from the upper and lower sidebands. Here the variation of τ_g causes $\mathbf{r}_u$ and $\mathbf{r}_\ell$ to rotate in opposite directions. To verify this statement, note that the real parts of the correlator output are given in Eqs. (6.11) and (6.15), and the corresponding imaginary parts are obtained by replacing ϕ_G by $\phi_G - \pi/2$. Then with $(\theta_m - \theta_n) = 0$ (no fringe rotation), consider the effect of a small change in τ_g.

The contrarotating vectors representing the two sidebands at the correlator output coincide at an angle determined by instrumental phase, which we represent by the line AB in Fig. 6.5b. Thus the vector sum oscillates along this line, and the fringe-frequency sinusoids at the real and imaginary outputs of the correlator are in phase. Now suppose that we adjust the phase term $(2\pi\nu_0\,\Delta\tau + \phi_G)$ in Eq. (6.18) to maximize the fringe amplitude at the real output. This action has the

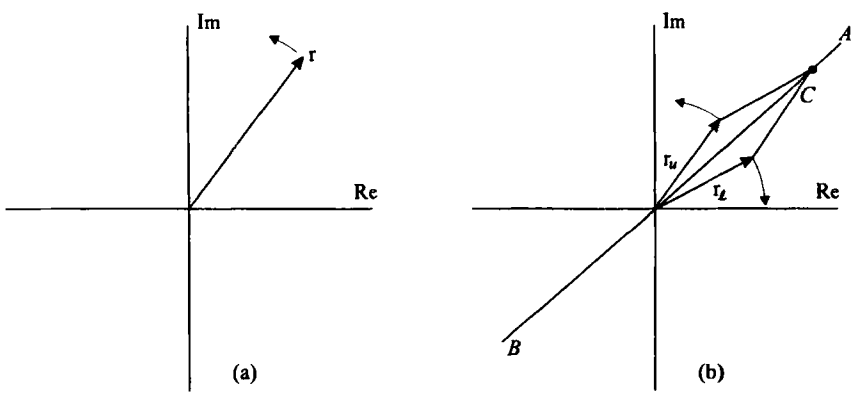


Figure 6.5 Representation in the complex plane of the output of a correlator with (a) a single-sideband and (b) a double-sideband receiving system. The point C in (b) represents the sum of the upper- and lower-sideband outputs of the correlator.

effect of rotating the line AB to coincide with the real axis. The imaginary output of the complex correlator then contains no signal, only noise. From Eq. (6.18) it can be seen that the visibility phase ϕ_v is represented by the phase of the vector that oscillates in amplitude along the real axis. The phase can be recovered by letting the fringes run and fitting a sinusoid to the waveform at the real output. If the fringes are stopped, it is possible to determine the amplitude and phase of the fringes by $\pi/2$ switching of the local oscillator phase at one antenna. In Eq. (6.18) this phase switch action can be represented by $\theta_m \to (\theta_m - \pi/2)$, which results in a change of the second cosine function to a sine, thus enabling the argument in square brackets to be determined. However, in such a case the data representing the cosine and sine components of the output are not measured simultaneously, so the effective data-averaging time is half that for the single-sideband, complex-correlator case. In Fig. 6.5b, a $\pi/2$ switch of the local oscillator phase results in a rotation of $\mathbf{r}_u$ and $\mathbf{r}_\ell$ by $\pi/2$ in opposite directions, so the vector sum of the two sideband outputs remains on the line AB. Relative sensitivities of different systems are discussed in Section 6.2; see Table 6.1 and associated text.

Double-Sideband System with Multiple Frequency Conversions

The response with multiple frequency conversions is more complicated for a double-sideband interferometer than for a single-sideband one and is illustrated by considering the system in Fig. 6.6. Note that for the case where the IF signal undergoes a number of single-sideband frequency conversions after the first mixer, the second mixer of each antenna in Fig. 6.6 can be considered to represent several mixers in series, and ν_2 to be equal to the sum of the local oscillator frequencies with appropriate signs to take account of upper- or lower-sideband conversions. The signal phase terms are determined by considerations similar to those described in the derivation of Eqs. (6.5) and (6.6). Thus we obtain

$$\phi_m = \mp 2\pi(\nu_1 \pm \nu_2 \pm \nu)\tau_g \mp \theta_{m1} - \theta_{m2} \tag{6.21}$$

and

$$\phi_n = -2\pi(\nu_2 + \nu)\tau_{i1} - 2\pi\nu\tau_{i2} \mp \theta_{n1} - \theta_{n2}, \tag{6.22}$$

where the upper signs correspond to upper-sideband conversion at both the first and second mixers for each antenna; and the lower signs, to lower-sideband conversion at the first mixer for each antenna and upper-sideband conversion at the second. We then proceed as in the previous examples; that is, use Eqs. (6.21) and (6.22) to substitute for ϕ_m and ϕ_n in Eq. (6.4), separate out the integral of $H_m H_n^*$ with respect to frequency, ν, as in Eq. (6.7), and substitute for the integral using Eq. (6.10). The results are

$$\begin{aligned} r_u = {} & |\mathcal{V}||G_{mn}(\Delta\tau)|\cos[2\pi\nu_1\tau_g + 2\pi\nu_2(\tau_g - \tau_{i1}) + 2\pi\nu_0\,\Delta\tau + (\theta_{m1} - \theta_{n1}) \\ & + (\theta_{m2} - \theta_{n2}) - \phi_v + \phi_G] \end{aligned} \tag{6.23}$$

and

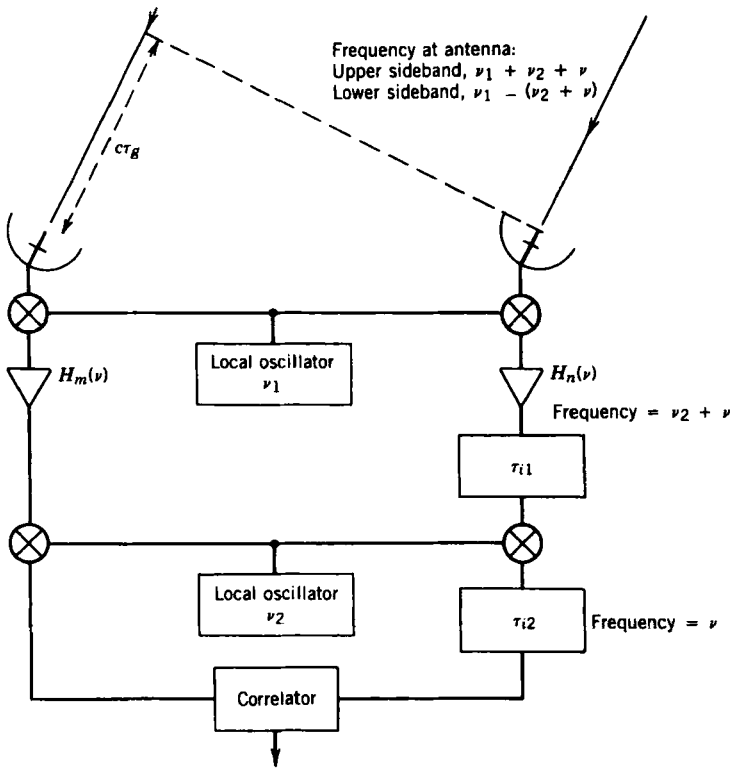


Figure 6.6 Receiving system for two antennas that incorporates two frequency conversions, the first being double-sideband and the second upper-sideband. Two compensating delays, τ_{i1} and τ_{i2}, are included so that in deriving the response for a double-sideband system the effect of the position of the delay relative to the first mixer can be investigated. In practice only one compensating delay is required. The overall frequency responses H_m and H_n are specified as functions of ν, which is the corresponding frequency at the correlator input.

$$r_\ell = |\mathcal{V}||G_{mn}(\Delta\tau)|\cos[2\pi\nu_1\tau_g - 2\pi\nu_2(\tau_g - \tau_{i1}) - 2\pi\nu_0\,\Delta\tau + (\theta_{m1} - \theta_{n1}) - (\theta_{m2} - \theta_{n2}) - \phi_v - \phi_G]. \tag{6.24}$$

The double-sideband response is

$$\begin{aligned} r_d &= r_u + r_\ell \\ &= 2|\mathcal{V}||G_{mn}(\Delta\tau)|\cos\left\{2\pi\left[\nu_2(\tau_{i1} - \tau_g) - \nu_0\,\Delta\tau\right] - (\theta_{m2} - \theta_{n2}) - \phi_G\right\} \\ &\quad \times \cos\left[\nu_1\tau_g + (\theta_{m1} - \theta_{n1}) - \phi_v\right], \end{aligned} \tag{6.25}$$

where $\Delta\tau = \tau_g - \tau_{i1} - \tau_{i2}$. Note that the phase of the output fringe pattern, given by the second cosine term, depends only on the phase of the first local oscillator. Thus, in the implementation of fringe rotation, the phase shift must be applied to

this oscillator. The first cosine term in Eq. (6.25) affects the fringe amplitude and two cases must be considered, as follows:

1. The delay τ_{i1}, at the intermediate frequency immediately following the double-sideband mixer, is used as the compensating delay, and $\tau_{i2} = 0$. Then in the first cosine function in Eq. (6.25), $\tau_{i1} - \tau_g \simeq 0$, and ϕ_G should be small if the frequency responses of the two channels are similar. It is necessary only to equalize θ_{m2} and θ_{n2} to maximize the amplitude of the fringe-frequency term. This is similar to the single conversion case in Eq. (6.18).
2. The delay τ_{i2}, located after the last mixer, is used as the compensating delay, and $\tau_{i1} = 0$. (This is the case in any array in which the compensating delays are implemented digitally, which includes most large arrays.) Then a continuously varying phase shift is required in θ_{m2} or θ_{n2} of Eq. (6.25) to keep the value of the first cosine function close to unity as τ_g varies. This phase shift does not affect the phase of the output fringe oscillations, only the amplitude [see, e.g., Wright et al. (1973)].

Fringe Stopping in a Double-Sideband System

Consider two antennas of an array as shown in Fig. 6.6, and the case where the instrumental delay that compensates for τ_g is the one immediately preceding the correlator, so that $\tau_{i1} = 0$. One can think of interferometer fringes as being caused by a Doppler shift in the signal at one antenna, which results in a beat frequency when the signals are combined in the correlator. Suppose that the geometric delay, τ_g, in the signal path to antenna m (on the left-hand side of the diagram) is increasing with time, that is, antenna m is moving away from the source relative to antenna n. Then a signal at frequency ν_{RF} at the wavefront from a source appears at frequency $\nu_{\mathrm{RF}}(1 - d\tau_g/dt)$ when received at antenna m. If the signal is in the upper sideband, its frequency at the correlator input will be

$$\nu_{\mathrm{RF}}\left(1 - \frac{d\tau_g}{dt}\right) - \nu_1 - \nu_2. \tag{6.26}$$

To stop the fringes, we need to apply a corresponding decrease to the frequency of the signal from antenna n so that the signals arrive at the correlator at the same frequency. To do this we increase the frequencies of the two local oscillators for antenna n by the factor $(1 + d\tau_g/dt)$. Note that this is equivalent to adding $2\pi(d\tau_g/dt)\nu_1$ to θ_{n1} and $2\pi(d\tau_g/dt)\nu_2$ to θ_{n2}, which are the rates of change of the oscillator phases required to maintain each of the two cosine functions in Eq. (6.25) at constant value. The corresponding signal from antenna n traverses the delay τ_{i2} at a frequency $\nu_{\mathrm{RF}} - (\nu_1 + \nu_2)(1 + d\tau_g/dt)$, and since the delay is continuously adjusted to equal τ_g, the signal suffers a reduction in frequency by a factor $(1 - d\tau_g/dt)$. Thus at the correlator input the frequency of the antenna-n signal is

$$\left[\nu_{\mathrm{RF}} - (\nu_1 + \nu_2)\left(1 + \frac{d\tau_g}{dt}\right)\right]\left(1 - \frac{d\tau_g}{dt}\right), \tag{6.27}$$

which is equal to (6.26) when second-order terms in $d\tau_g/dt$ are neglected. (Recall that for, e.g., a 1-km baseline the highest possible value of $d\tau_g/dt$ is 2.42×10^{-10}.) For the lower sideband, (6.26) and (6.27) apply if the signs of both ν_{RF} and ν_1 are reversed and again the frequencies at the correlator input are equal. Thus the overall effect is that the fringes are stopped for *both* sidebands.

Relative Advantages of Double- and Single-Sideband Systems

The principal reason for using double-sideband reception in interferometry is that in certain cases the lowest receiver noise temperatures are obtained by using input stages that are inherently double-sideband devices. At millimeter and shorter wavelengths (frequencies greater than ~100 GHz), it is difficult to make low-noise amplifiers, and receiving systems often use a mixer of the superconductor-insulator-superconductor (SIS) type [see, e.g., Tucker and Feldman (1985)] as the input stage followed by a low-noise IF amplifier. Both the mixer and the IF amplifier are cryogenically cooled to obtain superconductivity in the mixer and to minimize the amplifier noise. If a filter is placed between the antenna and the mixer to cut out one sideband, the received signal power is halved, but there is no reduction in the receiver noise generated in the mixer and IF stages. Thus the signal-to-noise ratio in the IF stages is reduced, and in this case the best continuum sensitivity may be obtained if both sidebands are retained. As a historical note, double-sideband systems were used at centimeter wavelengths during the 1960s and early 1970s [see, e.g., Read (1961)], sometimes with a degenerate type of parametric amplifier as the low-noise input stage. These amplifiers were inherently double-sideband devices and their use in interferometry is discussed by Vander Vorst and Colvin (1966).

Double-sideband systems have a number of disadvantages. Increased accuracy of delay setting is required, frequency and phase adjustment on more than one local oscillator is likely to be required, interpretation of spectral line data is complicated if there are lines in both sidebands, and the width of the required interference-free band of the radio spectrum is doubled. Also, the smearing effect of a finite bandwidth, to be discussed in Section 6.3, is increased. These problems have stimulated the development of schemes by which the responses for upper and lower sidebands can be separated.

Sideband Separation

To illustrate the method by which the responses for the two sidebands can be separated at the correlator output of a double-sideband receiving system, we examine the sum of the upper- and lower-sideband responses from Eqs. (6.11) and (6.15). This is

$$\begin{aligned} r_d = r_u + r_\ell = |\mathcal{V}||G_{mn}(\Delta\tau)|\{&\cos[2\pi(\nu_{\mathrm{LO}}\tau_g + \nu_0\Delta\tau) + \theta_{mn} - \phi_v + \phi_G] \\ &+ \cos\left[2\pi(\nu_{\mathrm{LO}}\tau_g - \nu_0\Delta\tau) + \theta_{mn} - \phi_v - \phi_G\right]\}, \end{aligned} \tag{6.28}$$

where $\theta_{mn} = \theta_m - \theta_n$. Equation (6.28) represents the real output of a complex correlator. We rewrite Eq. (6.28) as

$$r_d = |\mathcal{V}||G_{mn}|(\cos\Psi_u + \cos\Psi_\ell), \tag{6.29}$$

where Ψ_u and Ψ_ℓ represent the corresponding expressions in square brackets in Eq. (6.28). The responses considered above represent the normal output of the interferometer, which we call condition 1. The expression for the imaginary output of the correlator is obtained by replacing ϕ_G by $\phi_G - \pi/2$. Consider a second condition in which a $\pi/2$ phase shift is introduced into the first local oscillator signal of antenna m, so that θ_{mn} becomes $\theta_{mn} - \pi/2$. The correlator outputs for the two conditions are obtained from Eqs. (6.28) and (6.29):

$$\left.\begin{aligned} r_1 &= |\mathcal{V}||G_{mn}|(\cos\Psi_u + \cos\Psi_\ell) \\ r_2 &= |\mathcal{V}||G_{mn}|(\sin\Psi_u - \sin\Psi_\ell) \end{aligned}\right\} \quad \text{condition 1} \tag{6.30}$$

$$\left.\begin{aligned} r_3 &= |\mathcal{V}||G_{mn}|(\sin\Psi_u + \sin\Psi_\ell) \\ r_4 &= |\mathcal{V}||G_{mn}|(-\cos\Psi_u + \cos\Psi_\ell) \end{aligned}\right\} \quad \begin{aligned}&\text{condition 2} \\ &(\theta_{mn} \to \theta_{mn} - \pi/2),\end{aligned} \tag{6.31}$$

where r_1 and r_3 represent the real outputs of the correlator, and r_2 and r_4 the imaginary outputs. Thus the upper-sideband response, expressed in complex form, is

$$|\mathcal{V}||G_{mn}|(\cos\Psi_u + j\sin\Psi_u) = \tfrac{1}{2}[(r_1 - r_4) + j(r_2 + r_3)]. \tag{6.32}$$

Similarly, the lower-sideband response is

$$|\mathcal{V}||G_{mn}|(\cos\Psi_\ell + j\sin\Psi_\ell) = \tfrac{1}{2}[(r_1 + r_4) - j(r_2 - r_3)]. \tag{6.33}$$

If the $\pi/2$ phase shift is periodically switched into and out of the local oscillator signal, the upper- and lower-sideband responses can be obtained as indicated by Eqs. (6.32) and (6.33).

A similar implementation of sideband separation that makes use of fringe frequencies is attributable to B. G. Clark. This method is based on the fact that a small frequency shift in the first local oscillator adds the same frequency shift to the fringes at the correlator for both sidebands, but a similar shift in a later local oscillator adds to the fringe frequency for one sideband but subtracts from it for the other. Consider two antennas of an array in which the fringes have been stopped as in the discussion associated with expressions (6.26) and (6.27). Now suppose that we increase the frequency of the first local oscillator at antenna n by a frequency $\delta\nu$, and decrease the frequency of the second local oscillator by the same amount. The fringe frequency for the upper-sideband signal will be unchanged; that is, the fringes will remain stopped. For the lower sideband the signal frequencies after the second mixer will be decreased by $2\delta\nu$. The lower-sideband output will consist of fringes at frequency $2\delta\nu(1 - d\tau_g/dt) \approx 2\,\delta\nu$, and will be averaged to a small residual if $(2\,\delta\nu)^{-1}$ is small compared with the integration period at the correlator output, or if an integral number of fringe cycles fall

within such an integration period. If the frequency of the second local oscillator is increased by $\delta\nu$ instead of decreased, the lower sideband will be stopped and the upper one averaged out. To apply this scheme to an array of n_a antennas, the offset must be different for each antenna, and this can be achieved by using an offset $n\,\delta\nu$ for antenna n, where n runs from 0 to $n_a - 1$. An advantage of this sideband-separating scheme is that it can be implemented using the variable local oscillators required for fringe stopping, and no other special hardware is needed. Unlike the $\pi/2$ phase-switching scheme, one sideband is lost in this method. However, as mentioned above, sideband separation schemes of this type separate only the correlated component of the signal, and not the noise. To separate the noise, the SIS mixers at the receiver inputs can be mounted in a sideband-separating circuit of the type described in Appendix 7.1. In such cases the isolation of the sidebands achieved in the mixer circuit may be only ~15 dB, which is sufficient to remove most of the noise contributed by an unwanted sideband, but not sufficient to remove strong spectral lines. The Clark technique described above is nicely suited to increasing the suppression of an unwanted sideband that has already suffered limited rejection at the mixer.

Fringe-frequency effects can also be used for sideband separation in VLBI observations. In VLBI systems the fringe rotation is usually applied during the playback operation using a local oscillator later in the signal path than the first one. Fringe rotation then has the effect of reducing the fringe frequency for one sideband and increasing it for the other. If the fringe rotation is set to stop the fringes in one sideband, then since the baselines are so long, fringes resulting from the other sideband will generally have a sufficiently high frequency to be reduced to a negligible level by the time averaging at the correlator output. The data are played back to the correlator twice, once for each sideband, with appropriate fringe rotation.

6.2 RESPONSE TO THE NOISE

The ultimate sensitivity of a receiving system is determined principally by the system noise. We now consider the response to the noise and the resulting threshold of sensitivity, beginning with the effect at the correlator output and the resulting uncertainty in the real and imaginary parts of the visibility, $\mathcal{V}$. This leads to calculation of the rms noise level in a synthesized map in terms of the peak response to a source of given flux density. Finally, we consider the effect of noise in terms of the rms fluctuations in the amplitude and phase of $\mathcal{V}$.

Signal and Noise Processing in the Correlator

Consider an observation in which the field to be mapped contains only a point source located at the phase reference position. Let $V_m(t)$ and $V_n(t)$ be the waveforms at the correlator input from the signal channels of antennas m and n. The output is

$$r = \langle V_m(t) V_n(t) \rangle, \tag{6.34}$$

where all three functions are real, and the expectation denoted by the angular brackets is approximated in practice by a finite time average.† To determine the relative power levels of the signal and noise components of r, we determine their power spectra by first calculating the autocorrelation functions. The autocorrelation of the signal product in Eq. (6.34) is

$$\rho_r(\tau) = \langle V_m(t)V_n(t)V_m(t-\tau)V_n(t-\tau)\rangle. \tag{6.35}$$

This expression can be evaluated using the following fourth-order moment relation:‡

$$\langle z_1 z_2 z_3 z_4\rangle = \langle z_1 z_2\rangle\langle z_3 z_4\rangle + \langle z_1 z_3\rangle\langle z_2 z_4\rangle + \langle z_1 z_4\rangle\langle z_2 z_3\rangle, \tag{6.36}$$

where z_1, z_2, z_3, and z_4 are joint Gaussian random variables with zero mean. Thus,

$$\begin{aligned}\rho_r(\tau) &= \langle V_m(t)V_n(t)\rangle\langle V_m(t-\tau)V_n(t-\tau)\rangle \\ &\quad + \langle V_m(t)V_m(t-\tau)\rangle\langle V_n(t)V_n(t-\tau)\rangle \\ &\quad + \langle V_m(t)V_n(t-\tau)\rangle\langle V_m(t-\tau)V_n(t)\rangle \\ &= \rho_{mn}^2(0) + \rho_m(\tau)\rho_n(\tau) + \rho_{mn}(\tau)\rho_{mn}(-\tau),\end{aligned} \tag{6.37}$$

where ρ_m and ρ_n are the unnormalized autocorrelation functions of the two signals V_m and V_n, respectively, and ρ_{mn} is their cross-correlation function. Each V term is the sum of a signal component s and a noise component n, and to examine how these components contribute to the correlator output, we substitute them in Eq. (6.37). Products of uncorrelated terms, that is, products of signal and noise voltages, or noise voltages from different antennas, have an expectation of zero, and omitting them, we obtain

$$\begin{aligned}\rho_r(\tau) &= \langle s_m(t)s_n(t)\rangle\langle s_m(t-\tau)s_n(t-\tau)\rangle \\ &\quad + \langle s_m(t)s_m(t-\tau) + n_m(t)n_m(t-\tau)\rangle\langle s_n(t)s_n(t-\tau) + n_n(t)n_n(t-\tau)\rangle \\ &\quad + \langle s_m(t)s_n(t-\tau)\rangle\langle s_m(t-\tau)s_n(t)\rangle,\end{aligned} \tag{6.38}$$

where the three lines on the right-hand side correspond to the three terms on the last line of Eq. (6.37). To determine the effect of the frequency response of the receiving system on the various terms of $\rho(\tau)$, we need to convert them to power spectra. By the Wiener–Khinchin relation we should therefore examine the Fourier transforms of each term on the right-hand sides of Eqs. (6.37) and (6.38).

†The result for a total-power (single antenna) system given in Eq. (1.7) can also be derived by starting from Eq. (6.34), in this case by putting $V_m = V_n$ and proceeding through the analysis that follows.

‡This relation is a special case of a more general expression for the expectation of the product of N such variables, which is zero if N is odd and a sum of pair products if N is even. A form of Eq. (6.36) can be found in Lawson and Uhlenbeck (1950), Middleton (1960), and Wozencraft and Jacobs (1965).

The first term from Eq. (6.37), $\rho_{mn}^2(0)$, is a constant, and its Fourier transform is a delta function at the origin in the frequency domain, multiplied by $\rho_{mn}^2(0)$. From Eq. (6.38) we see that $\rho_{mn}^2(0)$ involves only the signal terms, which it is convenient to express as antenna temperatures. By the integral theorem of Fourier transforms, $\rho_{mn}(0)$ is the infinite integral of the Fourier transform of $\rho_{mn}(\tau)$, and thus the Fourier transform of $\rho_{mn}^2(0)$ is

$$k^2 T_{Am} T_{An} \left[\int_{-\infty}^{\infty} H_m(\nu) H_n^*(\nu)\, d\nu \right]^2 \delta(\nu), \tag{6.39}$$

where k is Boltzmann's constant, T_{Am} and T_{An} are the components of antenna temperature resulting from the source [see Eq. (1.5)], and $H_m(\nu)$ and $H_n(\nu)$ are the frequency responses of the signal channels.

The Fourier transform of the second term of Eq. (6.37), $\rho_m(\tau)\rho_n(\tau)$, is the convolution of the transforms of ρ_m and ρ_n, that is

$$k^2 (T_{Sm} + T_{Am})(T_{Sn} + T_{An}) \int_{-\infty}^{\infty} H_m(\nu) H_m^*(\nu) H_n(\nu' - \nu) H_n^*(\nu' - \nu)\, d\nu, \tag{6.40}$$

where T_{Sm} and T_{Sn} are the system temperatures. Note that the magnitude of this term is proportional to the product of the total noise temperatures.

The Fourier transform of the third term of Eq. (6.37), $\rho_{mn}(\tau)\rho_{mn}(-\tau)$, is the convolution of the transforms of $\rho_{mn}(\tau)$ and $\rho_{mn}(-\tau)$, and the latter is the complex conjugate of the former, since ρ_{mn} is real. Thus the Fourier transform of $\rho_{mn}(\tau)\rho_{mn}(-\tau)$ is

$$k^2 T_{Am} T_{An} \int_{-\infty}^{\infty} H_m(\nu) H_n^*(\nu) H_m^*(\nu' - \nu) H_n(\nu' - \nu) d\nu. \tag{6.41}$$

In expression (6.39), as in (6.37), only the antenna temperatures appear, since the receiver noise for different antennas makes no contribution to the cross-correlation.

Expression (6.39) represents the signal power in the correlator output, and (6.40) and (6.41) represent the noise. The effect of the time averaging at the correlator output can be modeled in terms of a filter that passes frequencies from 0 to $\Delta\nu_{\text{LF}}$. The output bandwidth $\Delta\nu_{\text{LF}}$ is less than the correlator input bandwidth by several or many orders of magnitude. Therefore, the spectral density of the output noise can be assumed to be equal to its value at zero frequency, that is, for $\nu' = 0$ in (6.40) and (6.41). From these considerations, and because $H_m(\nu)$ and $H_n(\nu)$ are hermitian, the ratio of the signal voltage to the rms noise voltage after averaging at the correlator output is

$$\mathcal{R}_{\text{sn}} = \frac{\sqrt{T_{Am} T_{An}} \int_{-\infty}^{\infty} H_m(\nu) H_n^*(\nu)\, d\nu}{\sqrt{(T_{Am} + T_{Sm})(T_{An} + T_{Sn}) + T_{Am} T_{An}} \sqrt{2\Delta\nu_{\text{LF}} \int_{-\infty}^{\infty} |H_m(\nu)|^2 |H_n(\nu)|^2\, d\nu}}, \tag{6.42}$$

where $2\,\Delta\nu_{\mathrm{LF}}$ is the equivalent bandwidth after averaging, with negative frequencies included. It is unusual for $\mathcal{R}_{\mathrm{sn}}$, the estimate of the signal-to-noise ratio at the output of a simple correlator, to be required to an accuracy better than a few percent. Indeed, it is usually difficult to specify T_S to any greater accuracy since the effects of ground radiation and atmospheric absorption on T_S vary as the antennas track. Thus, it is usually satisfactory to approximate $H_m(\nu)$ and $H_n(\nu)$ by identical rectangular functions of width $\Delta\nu_{\mathrm{IF}}$. Also, in sensitivity calculations one is concerned most often with sources near the threshold of detectability, for which $T_A \ll T_S$. With these simplifications Eq. (6.42) becomes

$$\mathcal{R}_{\mathrm{sn}} = \sqrt{\frac{T_{Am}T_{An}}{T_{Sm}T_{Sn}}}\sqrt{\frac{\Delta\nu_{\mathrm{IF}}}{\Delta\nu_{\mathrm{LF}}}}. \tag{6.43}$$

Figure 6.7 shows the signal and noise spectra for the rectangular bandpass approximation. Note that the input spectra $|H_m(\nu)|^2$ and $|H_n(\nu)|^2$ contain both positive and negative frequencies and are symmetric about the origin in ν. Thus, the out-

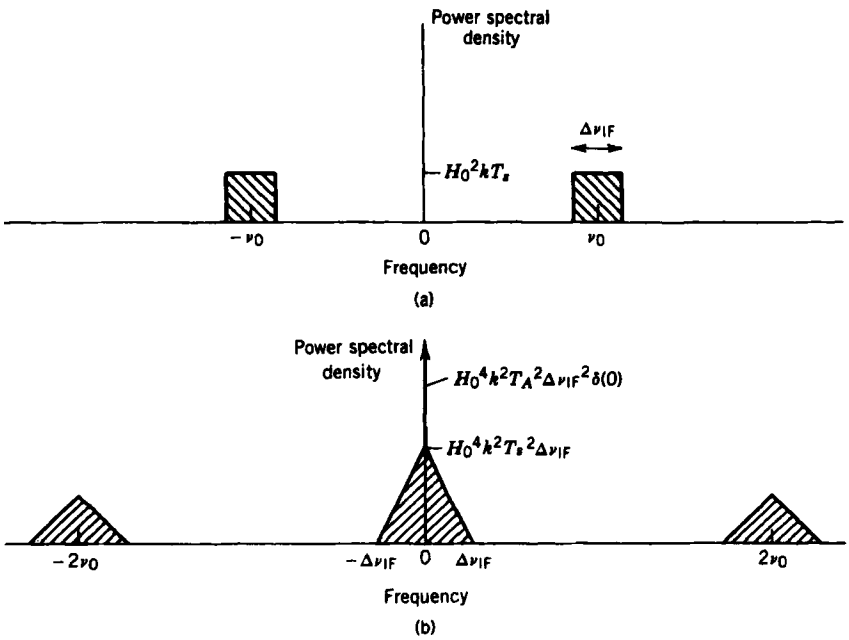


Figure 6.7 Spectra of (a) the input and (b) the output waveforms of a correlator. The input passbands are rectangular of width $\Delta\nu_{\mathrm{IF}}$. Shown in (b) is the complete spectrum of signals generated in the multiplication process, including noise bands at twice the input frequency. Only frequencies very close to zero are passed by the averaging circuit at the correlator output. These include the wanted signal, the spectrum of which has the form of a delta function and is represented by the arrow. It is assumed that $T_A \ll T_S$.

put noise spectrum can be described as proportional to either the convolution or the cross-correlation function of $|H_m(\nu)|^2$ and $|H_n(\nu)|^2$.

The output bandwidth is related to the data averaging time τ_a since the averaging can be described as convolution in the time domain with a rectangular function of unit area and width τ_a. The power response of the averaging circuit as a function of frequency is the square of the Fourier transform of the rectangular function, that is, $\sin^2(\pi\tau_a\nu)/(\pi\tau_a\nu)^2$. The equivalent bandwidth, including both positive and negative frequencies, is

$$2\Delta\nu_{\mathrm{LF}} = \int_{-\infty}^{\infty} \frac{\sin^2(\pi\tau_a\nu)}{(\pi\tau_a\nu)^2}\, d\nu = \frac{1}{\tau_a}. \tag{6.44}$$

Then from Eq. (6.43) we obtain

$$\mathcal{R}_{\mathrm{sn}} = \sqrt{\left(\frac{T_{Am}T_{An}}{T_{Sm}T_{Sn}}\right) 2\Delta\nu_{\mathrm{IF}}\tau_a}. \tag{6.45}$$

Note that $2\,\Delta\nu_{\mathrm{IF}}\tau_a$ is the number of independent samples of the signal in time τ_a, as mentioned in Section 1.2 under *Reception of Cosmic Signals*.

If the source is unpolarized, each antenna responds to half the total flux density S, and the received power density is

$$kT_A = \tfrac{1}{2}AS, \tag{6.46}$$

where A is the effective collecting area of the antenna. For identical antennas and system temperatures we obtain, from Eqs. (6.45) and (6.46),

$$\mathcal{R}_{\mathrm{sn}} = \frac{AS}{kT_S}\sqrt{\frac{\Delta\nu_{\mathrm{IF}}\tau_a}{2}}. \tag{6.47}$$

Similar derivations of this result can be found in the work of Blum (1959), Colvin (1961), and Tiuri (1964). Usually the result in Eq. (6.47), in which we have assumed $T_S \gg T_A$, is the one needed. At the other extreme, which may be encountered in observations of very strong, unresolved sources for which $T_A \gg T_S$, we have $\mathcal{R}_{\mathrm{sn}} = \sqrt{\Delta\nu_{\mathrm{IF}}\tau_a}$. The signal-to-noise ratio is determined by the fluctuations in signal level, and is independent of the areas of the antennas. Anantharamaiah et al. (1989) give a discussion of noise levels in the observation of very bright sources.

From Fig. 6.7 we can see how the factor $\sqrt{\Delta\nu_{\mathrm{IF}}\tau_a}$ in Eq. (6.47), which enables very high sensitivity to be achieved in radio astronomy, arises. The noise within the correlator results from beats between components in the two input bands and thus extends in frequency up to $\Delta\nu_{\mathrm{IF}}$. The triangular noise spectrum in Fig. 6.7 is simply proportional to the number of beats per unit frequency interval. However, only the very small fraction of this noise that falls within the output bandwidth is retained after the averaging. Note that the signal bandwidth $\Delta\nu_{\mathrm{IF}}$ that is important

here is the bandwidth at the correlator input. In a double-sideband system this is only one-half of the total input bandwidth at the antenna.

One other factor that affects the signal-to-noise ratio should be introduced at this point. If the signals are quantized and digitized before entering the correlators, an efficiency factor η_Q related to the quantization must be included, and Eq. (6.47) becomes

$$\mathcal{R}_{\rm sn} = \frac{AS\eta_Q}{kT_S}\sqrt{\frac{\Delta\nu_{\rm IF}\tau_a}{2}}, \tag{6.48}$$

or in terms of antenna temperature,

$$\mathcal{R}_{\rm sn} = \frac{T_A\eta_Q}{T_S}\sqrt{2\Delta\nu_{\rm IF}\tau_a}. \tag{6.49}$$

Values of η_Q vary between 0.637 and 1 and are discussed in Chapter 8; see Table 8.1. In VLBI observing, other losses affect the signal-to-noise ratio, and η_Q is replaced by a general loss factor η discussed in Section 9.7; see Eq. (9.156).

Noise in the Measurement of Complex Visibility

To understand precisely what $\mathcal{R}_{\rm sn}$ represents, note that in deriving Eqs. (6.48) and (6.49) no delay was introduced between the signal components at the correlator, and the phase responses of the signal channels were assumed to be identical. Thus the source is in the central fringe of the interferometer pattern, and in this particular case the response is the peak fringe amplitude, which represents the modulus of the visibility. To express the rms noise level at the correlator output in terms of the flux density σ of an unresolved source for which the peak fringe amplitude produces an equal output, we put $\mathcal{R}_{\rm sn} = 1$ in Eq. (6.48) and replace S by σ:

$$\sigma = \frac{\sqrt{2}kT_S}{A\eta_Q}\Big/\sqrt{\Delta\nu_{\rm IF}\tau_a}, \tag{6.50}$$

where σ is in units of W m^{-2} Hz^{-1}. Consider the case of an instrument with a complex correlator in which the output oscillations are slowed to zero frequency as described earlier. The noise fluctuations in the real and imaginary outputs are uncorrelated as we now show. Suppose that the antennas are pointed at blank sky so that the only inputs to the correlators in Fig. 6.3 are the noise waveforms n_m, n_n, and n_m^H, where the last is the Hilbert transform of n_m produced by the quadrature phase shift. The expectation of the product of the real and imaginary outputs is $\langle n_m n_n n_m^H n_n\rangle$, which can easily be shown to be zero by using Eq. (6.36) and noting that the expectations $\langle n_m n_n\rangle$, $\langle n_m n_m^H\rangle$, and $\langle n_m^H n_n\rangle$ must all be zero. Thus the noise from the real and imaginary outputs is uncorrelated.[§]

[§]The noise in the correlator outputs is composed of an ensemble of components of frequency $|\nu_m - \nu_n|$, where ν_m and ν_n are frequency components of the correlator inputs n_m and n_n. Components of the imaginary output are shifted in frequency by $\pm\pi/2$ relative to the corresponding components of the real output. Note that for any pair of input components, the sign of this shift in the imaginary output takes opposite values depending on whether $\nu_m > \nu_n$ or $\nu_m < \nu_n$. As a result, the noise waveforms at the correlator outputs are not a Hilbert transform pair, and one cannot be derived from the other.

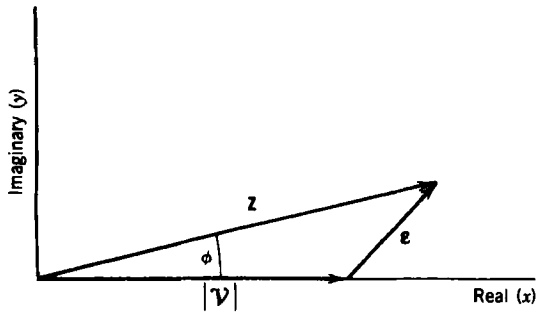


Figure 6.8 Complex quantity $\mathbf{Z}$, which is the sum of the modulus of the true complex visibility $|\mathcal{V}|$ and the noise $\boldsymbol{\varepsilon}$. The noise has real and imaginary components of rms amplitude σ, and ϕ is the phase deviation resulting from the noise.

The signal and noise components in the measurement of the complex visibility are shown in Fig. 6.8 as vectors in the complex plane. Here $\mathcal{V}$ represents the visibility as it would be measured in the absence of noise, and $\mathbf{Z}$ represents the sum of the visibility and noise, $\mathcal{V} + \boldsymbol{\varepsilon}$. We consider $\mathbf{Z}$ and $\boldsymbol{\varepsilon}$ to be vectors whose components correspond to the real and imaginary parts of the corresponding quantities. The noise in both components of $\mathbf{Z}$ has an rms amplitude σ. In practice, we must combine the real and imaginary outputs of the correlator to measure the visibility, and the resulting rms uncertainty in the measurement is

$$\varepsilon_{\text{rms}} = \sqrt{\langle \mathbf{Z} \cdot \mathbf{Z} \rangle - \langle \mathbf{Z} \rangle^2} = \sqrt{\langle \boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon} \rangle} = \sqrt{2}\,\sigma, \tag{6.51}$$

since $\langle \boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon} \rangle = \langle \varepsilon_x^2 \rangle + \langle \varepsilon_y^2 \rangle = 2\sigma^2$, where ε_x and ε_y are the components of $\boldsymbol{\varepsilon}$. If the measurement is made using only a single-multiplier correlator, one can periodically introduce a quadrature phase shift at one input, thus obtaining real and imaginary outputs, each for half the observing time. Then the data are half those that would be obtained with a complex correlator, and the noise in the visibility measurement is greater by $\sqrt{2}$. The same result is obtained by recording the single-multiplier output with a nonzero fringe frequency and fitting a sine curve. If the position of an unresolved source is known, it is possible to stop the fringes to give the maximum output, and thus measure the visibility amplitude with the same sensitivity as when using a complex correlator. However, this does not measure the *complex* visibility and is not generally useful.

Signal-to-Noise Ratio in a Synthesized Map

Having determined the noise-induced error in the visibility, the next step is to consider the signal-to-noise ratio in a map or image. Consider an array with n_p antenna pairs and suppose that the visibility data are averaged for time τ_a and that the whole observation covers a time interval τ_0. The total number of independent data points in the (u, v) plane is therefore

$$n_d = n_p \frac{\tau_0}{\tau_a}. \tag{6.52}$$

In mapping an unresolved source at the field center for which the visibility data combine in phase, we should thus expect the signal-to-noise ratio in the map to be greater than that in Eqs. (6.48) and (6.49) by a factor $\sqrt{n_p \tau_0/\tau_a}$. This simple consideration gives the correct result for the case in which the data are combined with equal weights. We now derive the result for the more general case of arbitrarily weighted data.

The ensemble of measured data can be represented by

$$\sum_{i=1}^{n_d} \left[{}^2\delta(u-u_i, v-v_i)(\mathcal{V}_i+\varepsilon_i) + {}^2\delta(u+u_i, v+v_i)(\mathcal{V}_i^*+\varepsilon_i^*)\right], \tag{6.53}$$

where ${}^2\delta$ is the two-dimensional delta function and ε_i is the complex noise contribution to the ith measurement. Each such data point appears at two (u, v) locations, reflected through the origin of the (u, v) plane. Before taking the Fourier transform of the data in Eq. (6.53), each data point is assigned a weight w_i (the choice of weighting factors is discussed in Section 10.2 under *Weighting of the Visibility Data*). To simplify the calculation we assume that the source is unresolved and located at the phase reference point of the map, and therefore produces a constant real visibility $\mathcal{V}$ equal to its flux density S. The intensity at the center of the map is then

$$I_0 = \frac{\sum_{i=1}^{n_d} w_i(\mathcal{V}+\varepsilon_{\mathcal{R}i})}{\sum w_i}, \tag{6.54}$$

where $\varepsilon_{\mathcal{R}i}$ is the real part of ε_i. Note that the imaginary part of ε_i vanishes at the map origin when the conjugate components are summed. For neighboring points in the map the same rms level of noise is distributed between the real and imaginary parts of ε. The expectation of I_0 is

$$\langle I_0\rangle = \mathcal{V} = S, \tag{6.55}$$

since $\langle\varepsilon_{\mathcal{R}i}\rangle = 0$. The variance of the estimate of the intensity, σ_m^2, is

$$\sigma_m^2 = \langle I_0^2\rangle - \langle I_0\rangle^2 = \frac{\sum w_i^2\langle\varepsilon_{\mathcal{R}i}^2\rangle}{\left(\sum w_i\right)^2}. \tag{6.56}$$

Equation (6.56) is derived directly from Eq. (6.54) using the fact that the noise terms from different (u, v) locations are uncorrelated, that is, $\langle\varepsilon_{\mathcal{R}i}\varepsilon_{\mathcal{R}j}\rangle = 0$, for $i \neq j$. We define the mean weighting factor w_{mean} and rms weighting factor w_{rms} by the equations

$$w_{\text{mean}} = \frac{1}{n_d}\sum w_i \tag{6.57}$$

and

$$w_{\text{rms}}^2 = \frac{1}{n_d}\sum w_i^2. \tag{6.58}$$

The noise contribution [see Eq. (6.51)] is the same for each (u, v) point and is equal to $\langle \varepsilon_{\mathcal{R}i}^2 \rangle = \sigma^2$, where σ is given by Eq. (6.50). Thus, the signal-to-noise ratio can be calculated from Eqs. (6.55), (6.56), (6.57), and (6.58) as

$$\frac{\langle I_0 \rangle}{\sigma_m} = \frac{S\sqrt{n_d}}{\sigma} \frac{w_{\text{mean}}}{w_{\text{rms}}}. \tag{6.59}$$

For an array with complex correlators we have, from Eq. (6.50),

$$\frac{\langle I_0 \rangle}{\sigma_m} = \frac{AS\eta_Q\sqrt{n_d \Delta\nu_{\text{IF}}\tau_a}}{\sqrt{2}kT_S} \frac{w_{\text{mean}}}{w_{\text{rms}}}. \tag{6.60}$$

If combinations of all pairs of antennas are used, $n_p = \frac{1}{2}n_a(n_a - 1)$, where n_a is the number of antennas. Since $n_d = n_p\tau_0/\tau_a$, we obtain

$$\frac{\langle I_0 \rangle}{\sigma_m} = \frac{AS\eta_Q\sqrt{n_a(n_a - 1)\Delta\nu_{\text{IF}}\tau_0}}{2kT_S} \frac{w_{\text{mean}}}{w_{\text{rms}}}. \tag{6.61}$$

To express the rms noise level in terms of flux density we put $I_0/\sigma_m = 1$ in Eq. (6.61). S then represents the flux density of a point source for which the peak response is equal to the rms noise level, and we can write

$$S_{\text{rms}} = \frac{2kT_S}{A\eta_Q\sqrt{n_a(n_a - 1)\Delta\nu_{\text{IF}}\tau_0}} \frac{w_{\text{rms}}}{w_{\text{mean}}}. \tag{6.62}$$

If all the weighting factors w_i are equal, $w_{\text{mean}}/w_{\text{rms}} = 1$, and this situation is referred to as the use of *natural weighting*. In such a case the signal-to-noise ratio given by Eq. (6.61) is equal to the corresponding sensitivity for a total-power receiver combined with an antenna of aperture $\sqrt{n_a(n_a - 1)}\,A$, which approaches n_aA as n_a becomes large. For an analysis of the sensitivity of single-antenna systems, see, for example, Tiuri (1964), Tiuri and Räisänen (1986).

We have considered the *point-source sensitivity* in Eq. (6.62). In the case of a source that is wider than the synthesized beam, it is useful to know the *brightness sensitivity*. The flux density (in W m^{-2} Hz^{-1}) received from a broad source of mean intensity I (W m^{-2} Hz^{-1} sr^{-1}) across the synthesized beam is $I\Omega$, where Ω sr is the effective solid angle of the synthesized beam. Thus the intensity level that is equal to the rms noise is S_{rms}/Ω. Note that the brightness sensitivity decreases as the synthesized beam becomes smaller, so compact arrays are best for detecting broad, faint sources. However, to measure the intensity of a uniform background, a measurement of the total power received in an antenna is required because a correlator interferometer does not respond to such a background.

The ratio $w_{\text{mean}}/w_{\text{rms}}$ is less than unity except when the weighting is uniform. Although the signal-to-noise ratio depends on the choice of weighting, in practice this dependence is not highly critical. The use of natural weighting maximizes the sensitivity for detection of a point source in a largely blank field but can

also substantially broaden the synthesized beam. The advantage in sensitivity is usually small. For example, if the density of data points is inversely proportional to the distance from the (u, v) origin, as is the case for an east–west array with uniform increments in antenna spacing, the weighting factors required to obtain effective uniform density of data result in $w_{\text{mean}}/w_{\text{rms}} = 2\sqrt{2}/3 = 0.94$. In this case the natural weighting results in an undesirable beam profile in which the response remains positive for large angular distances from the beam axis and dies away only slowly.

Various methods of Fourier transformation of visibility data are reviewed in Chapter 10, and the results derived in Eqs. (6.61) and (6.62) can be applied to these by using the appropriate values of w_{mean} and w_{rms}. Convolution of the visibility data in the (u, v) plane to obtain values at points on a rectangular grid is a widely used process. In general, the data at adjacent grid points are then not independent, and a tapering of the signal and noise is introduced into the map. Aliasing can also cause the signal-to-noise ratio to vary across the map. (These effects are explained in Fig. 10.5 and the associated discussion.) In such cases the results derived here apply near the origin of the map, where the effects of tapering and aliasing are unimportant. The rms noise level over the map can be obtained by the application of Parseval's theorem to the noise in the visibility data.

In practice a number of factors that affect the signal-to-noise ratio are difficult to determine precisely. For example, T_S varies somewhat with antenna elevation. There are also a number of effects that can reduce the response to a source without reducing the noise, but these are important only for sources not near the (l, m) origin of a map. These include the smearing resulting from the receiving bandwidth and from visibility averaging, discussed later in this chapter, and the effect of non-coplanar baselines, discussed in Sections 3.1 and 11.8.

Note also that in many instruments two oppositely polarized signals (with crossed linear or opposite circular polarizations) are received and processed using separate IF amplifiers and correlators. For unpolarized sources, the overall signal-to-noise ratio is then $\sqrt{2}$ greater than the values derived above, which include only one signal from each antenna.

Noise in Visibility Amplitude and Phase

In synthesis mapping we are usually concerned with data in the form of the real and imaginary parts of $\mathcal{V}$, but sometimes it is necessary to work with the amplitude and phase. Given that the real and imaginary parts of $\mathcal{V}$ are accompanied by Gaussian noise of standard deviation σ, what are the probability distributions of the amplitudes and phases? The answers are well known, and we do not derive them here. The sum of the visibility and noise is represented by $\mathbf{Z} = Ze^{j\phi}$, where we choose the real axis so that the phase ϕ is measured with respect to the phase of $\mathcal{V}$, as in Fig. 6.8. Then for $T_A \ll T_S$ the probability distributions of the resulting amplitude and phase are

$$p(Z) = \frac{Z}{\sigma^2} \exp\left(-\frac{Z^2 + |\mathcal{V}|^2}{2\sigma^2}\right) I_0\left(\frac{Z|\mathcal{V}|}{\sigma^2}\right), \qquad Z > 0 \tag{6.63a}$$

$$p(\phi) = \frac{1}{2\pi} \exp\left(-\frac{|\mathcal{V}|^2}{2\sigma^2}\right) \left\{1 + \sqrt{\frac{\pi}{2}} \frac{|\mathcal{V}| \cos\phi}{\sigma} \exp\left(\frac{|\mathcal{V}|^2 \cos^2\phi}{2\sigma^2}\right) \times \left[1 + \operatorname{erf}\left(\frac{|\mathcal{V}| \cos\phi}{\sqrt{2}\sigma}\right)\right]\right\}, \tag{6.63b}$$

where I_0 is the modified Bessel function of zero order, erf is the error function, and σ is as given by Eq. (6.50). The amplitude distribution is identical to that for a sine wave in noise, and the derivation is given by Rice (1944, 1954), Vinokur (1965), and Papoulis (1965), of which the last two also derive the result for the phase. $p(Z)$ is sometimes referred to as the Rice distribution, and for $\mathcal{V} = 0$ it reduces to the Rayleigh distribution. Curves of $p(Z)$ and $p(\phi)$ are given in Fig. 6.9. Comparison of the curves for $|\mathcal{V}|/\sigma = 0$ and 1 indicates that the presence of a weak signal is more easily detected by examining the visibility phase than by examining the amplitude.

Approximation for $p(Z)$ and $p(\phi)$ for the cases where $|\mathcal{V}|/\sigma \ll 1$ and $|\mathcal{V}|/\sigma \gg 1$ are given in Section 9.3 under *Noise in VLBI Observations*. Expressions for the moments of Z and ϕ and their rms deviations are also given in that section. The rms phase deviation σ_ϕ is a particularly useful quantity, especially for astrometric and diagnostic work. The expression for σ_ϕ, valid for the case where $|\mathcal{V}|/\sigma \gg 1$, is $\sigma_\phi \simeq \sigma/|\mathcal{V}|$ [Eq. (9.53)]. This result also follows intuitively from an examination of Fig. 6.8. By substituting Eq. (6.50) into the expression for σ_ϕ, setting $|\mathcal{V}|$ equal to the flux density S of the source, which is appropriate if the source is unresolved, and using Eq. (6.46) to relate the flux density and antenna temperature, we obtain

$$\sigma_\phi = \frac{T_S}{\eta_Q T_A \sqrt{2\Delta\nu_{\mathrm{IF}}\tau_a}}. \tag{6.64}$$

This equation is valid for the conditions $T_S/\sqrt{2\Delta\nu_{\mathrm{IF}}\tau_a} \ll T_A \ll T_S$, which are the conditions most frequently encountered, and is useful for determining whether or not the noise in the phase measurements of an interferometer is due exclusively to receiver noise. Excess phase noise can be contributed by the atmosphere, by system instabilities, and in the case of VLBI by the frequency standards.

Relative Sensitivities of Different Interferometer Systems

Next we compare the sensitivity of several different interferometer systems, using as a measure of sensitivity the modulus of the signal divided by the rms noise, that is, $\mathcal{V}/\varepsilon$ in terms of the quantities at the correlator output in Fig. 6.8. Parameters such as averaging times and IF bandwidths are the same for all cases considered. To compare double- and single-sideband cases, it is convenient to introduce a factor

$$\alpha = \frac{\text{double-sideband system temp. of double-sideband system}}{\text{system temperature of single-sideband system}}. \tag{6.65}$$

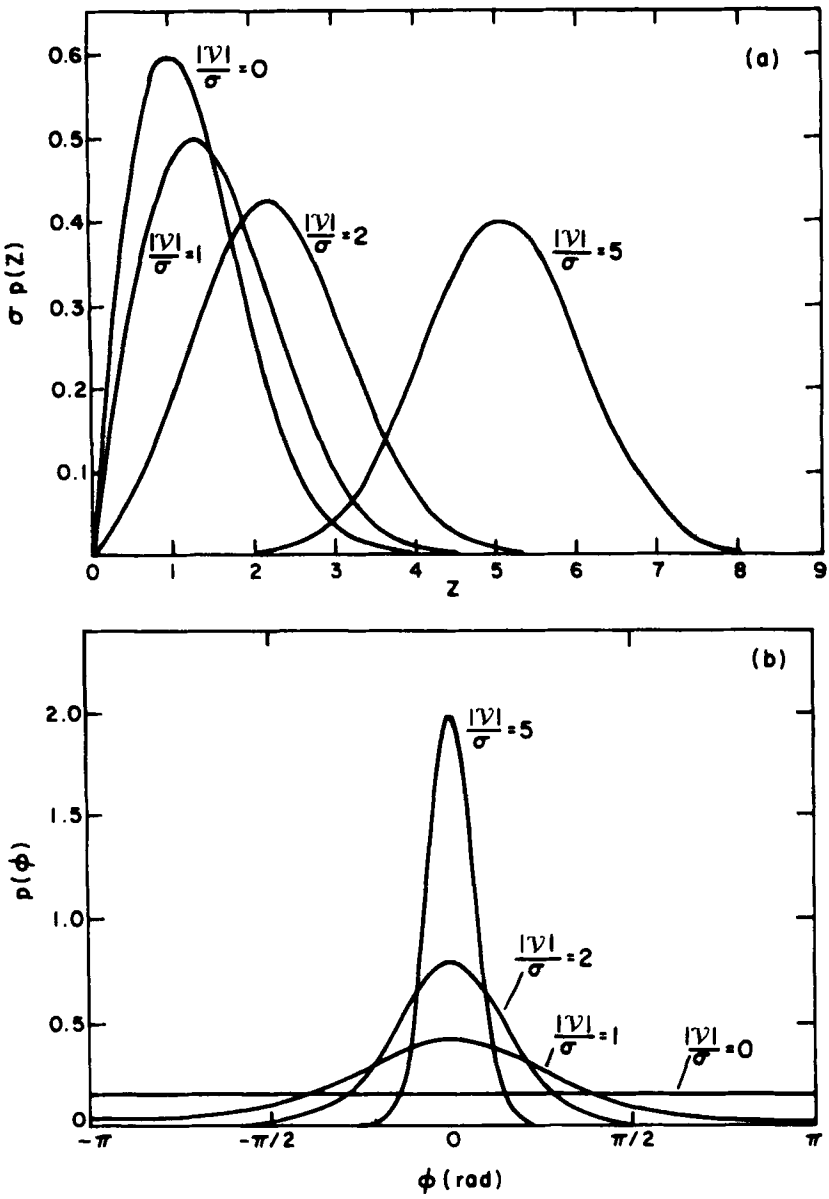


Figure 6.9 Probability distributions of (a) the amplitude, and (b) the phase, of the measured complex visibility as functions of the signal-to-noise ratio. $|\mathcal{V}|$ is the modulus of the signal component. After Moran (1976).

Recall that the system temperature of a receiver can be defined as the noise temperature of a thermal source at the input of a hypothetical noise-free (but otherwise identical) receiver that would produce the same noise level at the receiver output. [Equation (1.4) can be used for the equivalent noise temperature at the source if the Rayleigh–Jeans approximation does not apply.] For a double-sideband receiver the system temperature is described as double-sideband or single-sideband depending on whether the thermal noise source emits noise in both sidebands or only one, respectively. With these definitions, the single-sideband noise temperature is twice the double-sideband noise temperature.

For a single-sideband system the rms noise from one output of a correlator (either the real or imaginary output in the case of a complex correlator) is σ after averaging for a time τ_a, as given by Eq. (6.50). The corresponding noise power is σ^2. For a double-sideband system the rms output noise at a correlator output is $2\alpha\sigma$. In all cases the signal results from an unresolved source. For a single-sideband system we take the signal voltage from the correlator output to be $\mathcal{V}$, as in Fig. 6.8. For a double-sideband system with the input signal in one sideband only, the signal at the correlator output is $\mathcal{V}$, and for a double-sideband system with input in both sidebands, the correlator output is $2\mathcal{V}$.

Values of the relative sensitivity for various systems are discussed below and summarized in Table 6.1. Similar results are given by Rogers (1976).

1. *Single-sideband system with complex correlator.* The output signal is $\mathcal{V}$ and the rms noise from each correlator output is σ. As shown by Fig. 6.8 and Eq. (6.51), the ratio of the signal amplitude to rms noise is $\mathcal{V}/(\sqrt{2}\sigma)$. We shall take this as the standard with respect to which the relative sensitivities of other systems are defined.
2. *Single-sideband system and simple correlator with fringe fitting.* To measure both the real and imaginary parts of the complex visibility, the fringes are not stopped but appear as a sinusoid of amplitude $\mathcal{V}$ at the fringe frequency ν_f. The signal is accompanied by noise of rms amplitude σ. The amplitude and phase are measured by "fringe fitting," that is, performing a least-squares fit of a sinusoid to the correlator output. This procedure involves multiplying the correlator output waveform by $\cos(2\pi\nu_f t)$ and $\sin(2\pi\nu_f t)$ and integrating over the period τ_a. The results represent the real and imaginary parts, respectively, of the cross-correlation. We calculate the effects of fringe fitting on the signal and noise separately, and assume, with no loss of generality, that the fringes are in phase with the cosine component in the fringe fitting, in which case the sine component of the signal is zero. The correlator output has a bandwidth $\Delta\nu_c$ which is sufficient to pass the fringe-frequency waveform, and it is sampled at time intervals $\tau_s = 1/(2\nu_c)$ and digitized. Within the period τ_a there are $N = 2\Delta\nu_c\tau_a$ samples. Thus for the cosine component of the signal the amplitude is

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{V}\cos^2(2\pi i\nu_f\tau_s) = \frac{\mathcal{V}}{2} + \frac{\mathcal{V}}{2N}\sum_{i=1}^{N}\cos(4\pi i\nu_f\tau_s). \tag{6.66}$$

TABLE 6.1 Relative Signal-to-Noise Ratios for Several Types of Systems

System Type	Relative SNR
1. Single sideband with complex correlator	1
2. Single sideband with simple correlator	$\frac{1}{\sqrt{2}}$
3. Single sideband, simple correlator, fringe stopping, $\pi/2$ phase switching	$\frac{1}{\sqrt{2}}$
4. Double sideband, simple correlator,[a] fringe fitting, continuum signal	$\frac{1}{\sqrt{2\alpha}}$
5. Double sideband, simple correlator, fringe stopping, $\pi/2$ phase switching, continuum signal	$\frac{1}{\sqrt{2\alpha}}$
6a. Double sideband, fringe stopping, sideband separation [Eqs. (6.30) to (6.33)], signal in one sideband only	$\frac{1}{2\alpha}$
6b. As (6a) but for continuum signal and visibilities in both sidebands combined	$\frac{1}{\sqrt{2\alpha}}$
7a. VLBI, double sideband, complex correlator, one sideband removed by averaging of fast fringes	$\frac{1}{2\alpha}$
7b. As (7a) but for continuum signal, correlated separately for each sideband and results combined	$\frac{1}{\sqrt{2\alpha}}$
8. Single sideband, digital spectral correlator with simple correlator elements and correlation measured as a function of time offsets (see Section 8.7)	1

[a] For double sideband with complex correlator, see text pertaining to Fig. 6.5.

The second term on the right-hand side represents the end effects and is approximately zero if there are an integral number of half-cycles of the fringe frequency within the period τ_a. It also becomes relatively small as $\nu_f \tau_a$ increases, and we assume here that there are enough fringe cycles (say, ten or more) within time τ_a that end effects can be neglected. To determine the effect of fringe fitting on the noise, we represent the sampled noise by $n(i\tau_s)$, multiply by the cosine function, and determine the variance (mean squared value). Averaged over time τ_a, the result is

$$\frac{1}{N}\left[\sum_{i=1}^{N} n(i\tau_s)\cos(2\pi i \nu_f \tau_s)\right]^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{N} n(i\tau_s)\cos(2\pi i \nu_f \tau_s) n(k\tau_s)\cos(2\pi k \nu_f \tau_s). \quad (6.67)$$

We need to determine the expectation value of this expression, which we denote by angle brackets. Only terms for which $i = k$ contribute to the expectation. Thus the noise variance becomes

$$\left\langle \frac{1}{2N} \sum_{i=1}^{N} n^2(i\tau_s)[1 + \cos(4\pi i \nu_f \tau_s)] \right\rangle = \frac{\sigma^2}{2}. \tag{6.68}$$

This result shows that half of the noise power, σ^2, that is available at the correlator output appears in the cosine component of the fringe fitting. Similarly, the other half appears in the sine component. The combined rms noise of the two components is σ, and the ratio of signal to noise after fringe fitting is $\mathcal{V}/(2\sigma)$. The relative sensitivity is $1/\sqrt{2}$.

3. *Single-sideband system with simple correlator and $\pi/2$ phase switching of LO.* In this case the fringes have been stopped, and to determine the complex visibility, a phase change of $\pi/2$ is periodically inserted into one oscillator [e.g., $\theta_n \to \theta_n + \pi/2$ in Eq. (6.11) or (6.15)] so that the correlator is effectively time-shared between the real and imaginary parts of the cross-correlation function, which are averaged separately. The visibility phase can thereby be determined. The signal in the two phase conditions is $\mathcal{V}\cos(\phi_v)$ and $\mathcal{V}\sin(\phi_v)$, and the rms noise associated with each of these terms is $\sqrt{2}\,\sigma$ (the $\sqrt{2}$ factor enters because the noise in each output is averaged over time $\tau/2$ only). Thus the modulus of the signal is $\mathcal{V}$ and the rms noise from the two components is 2σ. The signal-to-noise ratio is $\mathcal{V}/(2\sigma)$ and the relative sensitivity is $1/\sqrt{2}$.
4. *Double-sideband system with simple correlator and fringe fitting.* We consider the case of a continuum source with signal in both sidebands, and assume that the instrumental delay is adjusted so that the signal appears entirely in the (real) output of a simple correlator, as a fringe-frequency cosine wave of amplitude $\mathcal{V}$. In terms of Eq. (6.18), the factor $\cos(2\pi\nu_0\,\Delta\tau_a + \phi_G)$ is unity. Then for the double-sideband system the signal amplitude is $2\mathcal{V}$ and the rms noise is $2\alpha\sigma$. The fringe-fitting procedure follows that of case 2, but in this case the signal amplitude is greater by a factor of two and is equal to $\mathcal{V}$. The rms noise is greater by a factor of 2α. Thus the signal-to-noise ratio is $\mathcal{V}/(2\alpha\sigma)$ and the relative sensitivity is $1/(\sqrt{2}\,\alpha)$.
5. *Double-sideband system with simple correlator and $\pi/2$ phase switching of LO.* Here the fringes have been stopped, and to determine the visibility phase, it is necessary to perform $\pi/2$ phase switching as in case 3 above. (For a double-sideband system the phase switching must be on the first local oscillator.) The amplitude of the signal is $2\mathcal{V}$ because the system is double sideband, and the rms noise level from the correlator output is increased to $2\sqrt{2}\,\alpha\sigma$ because the averaging time for each component is reduced to $\tau_a/2$ by the time sharing of the correlator between the two phase conditions. This rms level is associated with both the cosine and sine components of the signal, so the signal-to-noise ratio is $\mathcal{V}/(2\alpha\sigma)$. The relative sensitivity is $1/(\sqrt{2}\,\alpha)$.
6. *One sideband of a double-sideband system with $\pi/2$ phase switching of the LO and sideband separation after correlation.* A complex correlator is used and the procedure corresponding to Eqs. (6.30) to (6.33) is followed. We

consider the upper sideband, and ignore lower-sideband signal terms. The components r_1, r_2, r_3, r_4 have amplitudes $\mathcal{V}$ multiplied by the cosine or sine of Ψ_u. Thus from Eqs. (6.30) and (6.31), ignoring lower-sideband terms, the right-hand side of Eq. (6.32) becomes $\frac{1}{2}(2\mathcal{V}\cos\Psi_u + j2\mathcal{V}\sin\Psi_u)$, the modulus of which is $\mathcal{V}$. The rms noise associated with each term r_1, r_2, r_3, and r_4 is $2\sqrt{2}\alpha\sigma$ since the system is double sideband and, because of the LO switching, the effective averaging time is $\tau_a/2$. Thus the rms noise associated with the right-hand side of Eq. (6.32) is $2\sqrt{2}\alpha\sigma$, as in case 5. The signal-to-noise ratio is $\mathcal{V}/(2\sqrt{2}\alpha\sigma)$, and the relative sensitivity is $1/(2\alpha)$. This applies to a signal in one sideband such as a spectral line. For a continuum source the cross-correlation can be measured for each of the two sidebands, and if the results are then averaged the relative sensitivity becomes $1/(\sqrt{2}\alpha)$. The terms r_2 and r_4 are eliminated in averaging the right-hand sides of Eqs. (6.32) and (6.33), and the result is the same as for a simple correlator with LO phase switching described under case 5 above.

7. *VLBI observations with a double-sideband system and complex correlator.* In VLBI observations a double-sideband system is sometimes used and fringe rotation is inserted after playback of the recorded signal, as mentioned in Section 6.1. For one sideband the fringes are stopped, but for the other they are lost in the averaging at the correlator output because the fringe frequencies are high. Thus, for one playback, we have the signal of a single-sideband system and the noise of a double-sideband system in each of the real and imaginary outputs, that is, a signal-to-noise ratio of $\mathcal{V}/(2\sqrt{2}\alpha\sigma)$ and a relative sensitivity of $1/(2\alpha)$ for each individual sideband.

8. *Measurement of cross-correlation as a function of time delay.* Digital spectral correlators that measure cross-correlation as a function of time delay are described in Section 8.7. In a lag-type correlator, the cross-correlation is measured as a function of time offset, implemented by introducing instrumental delays. The Fourier transform of the cross-correlation as a function of relative time delay between the signals is the cross-correlation as a function of frequency, as required in spectral line measurements. As mentioned in Section 6.1 under *Simple and Complex Correlators*, it is only necessary to use simple correlators for this measurement. The range of time offsets of the two signals covers both positive and negative values, and the resulting measurements of cross-correlation contain both even and odd components. Fourier transformation then provides both the real and imaginary components of the cross-correlation as a function of frequency. The full sensitivity is obtained so long as the range of time offsets is comparable to the reciprocal signal bandwidth or greater; see *Fringe sideband rejection loss* in Table 9.6 of Chapter 9. Note that in Table 6.1 we have not included the quantization loss discussed in Section 8.3. A demonstration of the relative sensitivity using a simple correlator when the measurements are made as a function of time delay is given by Mickelson and Swenson (1991).

Of the cases included in Table 6.1, the single sideband with complex correlator is the one generally used where possible, because of the sensitivity and avoidance of the complications of double-sideband operation. Cases 2 and 3 in the table are included mainly for completeness of the discussion. For high frequencies at which low-noise amplifiers are not available (generally above ~100 GHz), the most sensitive type of receiver input is an SIS mixer. This has an inherently double-sideband response, and although a sideband can be removed by filtering or using a sideband-separating arrangement (Appendix 7.1), double-sideband operation may be preferred to avoid any loss in sensitivity, or in flexibility of tuning, that results from the greater complexity required to remove one sideband. For double-sideband operation the most important cases in the table are 6a and 6b. The case where the unwanted sideband is only partially rejected is discussed in Appendix 6.1.

System Temperature Parameter α

As already noted, double-sideband systems are mainly used at millimeter and submillimeter wavelengths, at which the receiver input stage is commonly an SIS mixer. Such a system can be converted to single-sideband operation by filtering out the unwanted sideband and terminating the corresponding input in a cold load. If the atmospheric losses are high and the receiver temperature is low, most of the system noise will come from the antenna, and terminating one sideband in a cold load will approximately halve the level of noise within the receiver. The system temperature of the single-sideband system will then be approximately equal to the *double*-sideband system temperature of the double-sideband system, and the value of α [defined in Eq. (6.65)] tends toward 1. On the other hand, if atmospheric and antenna losses are low and most of the system noise comes from the mixer and IF stages, then terminating one sideband input in a cold load rather than the cold sky makes little difference to the noise level in the receiver. The system temperature of the single-sideband system will be close to the *single*-sideband system temperature of the double-sideband system, which is twice the double-sideband value. The value of α then tends toward $1/2$. To recapitulate, if the atmospheric noise dominates the receiver noise, then α tends toward 1, but if the receiver noise dominates, then α tends toward $1/2$. Note, however, that α is not confined to the range $1/2 < \alpha < 1$. For example, if noise from the antenna is low but the termination of the image sideband in the single-sideband system is uncooled and injects a high noise level, then α can be $< 1/2$. If the front end is tuned close to an atmospheric absorption line in such a way that the additional sideband of the double-sideband system falls in a frequency range of enhanced atmospheric noise, then α can be > 1.

6.3 EFFECT OF BANDWIDTH

As seen in the preceding section, the sensitivity of a receiving system to a broadband cosmic signal increases with the system bandwidth. Here we are concerned with the effect of bandwidth on the angular range over which fringes are detected,

and on the fringe amplitude. These effects result from the variation of fringe frequency, in cycles per radian on the sky, with the received radio frequency. If the monochromatic response is integrated over the bandwidth, the fringes are reinforced for directions close to that for which the time delays from the source to the correlator inputs are equal, but for other directions the fringes vary in phase across the bandwidth. This effect, when measured in a plane containing the interferometer baseline, causes the fringe amplitude to decrease with angle in a manner similar to that caused by the antenna beams (Swenson and Mathur 1969), and is sometimes referred to as the *delay beam*. It can be used to confine the response of an interferometer to a limited area of the sky and thereby reduce the possibility of source confusion, which can occur when the fringe patterns of two or more sources are recorded simultaneously. Examples of such usage can be found in some early interferometers built for operation at frequencies below 100 MHz (Goldstein 1959, Douglas et al. 1973). The technique is less useful for instruments in which the antennas track in hour angle because the width of the delay beam becomes larger as the projected baseline is foreshortened.

Mapping in the Continuum Mode

The effect of bandwidth on the fringe amplitude was discussed in Section 2.2. Equation (2.3) gives an expression for the fringes observed for a point source with an east–west baseline of length D, and a rectangular signal passband of width $\Delta\nu$. The fringe amplitude is proportional to a factor

$$R_b' = \frac{\sin(\pi D l \Delta\nu/c)}{\pi D l \Delta\nu/c}. \tag{6.69}$$

Consider an array for which D is typical of the longest baselines. The synthesized beamwidth of the array, θ_b, is approximately equal to $\lambda_0/D = c/\nu_0 D$, where ν_0 is the observing frequency and λ_0 the corresponding wavelength. (Note that in this section ν_0 is the center frequency of the RF input band, not an IF band.) Thus Eq. (6.69) becomes

$$R_b' \simeq \frac{\sin(\pi \Delta\nu l/\nu_0\theta_b)}{\pi \Delta\nu l/\nu_0\theta_b}. \tag{6.70}$$

The parameter $\Delta\nu l_1/\nu_0\theta_b$ is equal to the fractional bandwidth multiplied by the angular distance of the source from the (l, m) origin measured in beamwidths. If this parameter is equal to unity, $R_b' = 0$ and the measured visibility is reduced to zero. To keep R_b' close to unity, we require $\Delta\nu l/\nu_0\theta_b \ll 1$. Thus, to avoid underestimation of the visibility at long baselines, there is a limit on the angular size of the map that is inversely proportional to the fractional bandwidth.

We now examine the same effect in more detail by considering the distortion in the synthesized map. First recall that the response of an array can be written as

$$\mathcal{V}(u, v)W(u, v) \rightleftharpoons I(l, m) ** b_0(l, m), \tag{6.71}$$

where $\rightleftharpoons$ represents Fourier transformation. The fringe visibility is multiplied by $W(u, v)$, the spatial sensitivity function of the array for a particular observation. The Fourier transform of the left-hand side of Eq. (6.71) gives the intensity distribution $I(l, m)$ convolved with the synthesized beam function $b_0(l, m)$. For simplicity we have omitted the primary antenna beam and minor effects related to use of the discrete Fourier transform. The synthesized beam is defined here as the Fourier transform of $W(u, v)$.

In operation in the continuum mode, the visibility data measured with bandwidth $\Delta\nu$ are treated as though they were measured with a monochromatic receiving system tuned to the center frequency ν_0. Thus for all frequencies within the bandwidth, the assigned values of u and v are those appropriate to frequency ν_0. At another frequency ν within the passband, the true spatial frequency coordinates u_ν and v_ν are related to the assigned values u and v by

$$(u, v) = \left(\frac{u_\nu \nu_0}{\nu}, \frac{v_\nu \nu_0}{\nu}\right). \tag{6.72}$$

The contribution to the measured visibility from a narrow band of frequencies centered on ν is

$$\mathcal{V}_\nu(u, v) = \mathcal{V}_\nu\left(\frac{u_\nu \nu_0}{\nu}, \frac{v_\nu \nu_0}{\nu}\right) \rightleftharpoons \left(\frac{\nu}{\nu_0}\right)^2 I\left(\frac{l\nu}{\nu_0}, \frac{m\nu}{\nu_0}\right), \tag{6.73}$$

where we have used the similarity theorem of Fourier transforms [see, e.g., Bracewell (2000)]. Thus the contribution to the measured intensity is the true intensity distribution scaled in (l, m) by a factor ν/ν_0 and in intensity by $(\nu/\nu_0)^2$. The derived intensity distribution is convolved with $b_0(l, m)$, the synthesized beam corresponding to frequency ν_0. The beam does not vary with frequency since the same spacial sensitivity function $W(u, v)$ is used to represent the whole frequency passband. The overall response is obtained by integrating over the passband with appropriate weighting and is

$$I_b(l, m) = \left[\frac{\int_0^\infty \left(\frac{\nu}{\nu_0}\right)^2 |H_{\mathrm{RF}}(\nu)|^2 I\left(\frac{l\nu}{\nu_0}, \frac{m\nu}{\nu_0}\right) d\nu}{\int_0^\infty |H_{\mathrm{RF}}(\nu)|^2 \, d\nu}\right] ** b_0(l, m). \tag{6.74}$$

Note that the integrals must be taken over the whole radio-frequency passband, denoted by the subscript RF, which includes both sidebands in the case of a double-sideband system. We assume that the passband function $H_{\mathrm{RF}}(\nu)$ is identical for all antennas. The values of l and m in the intensity function in Eq. (6.74) are multiplied by the factor ν/ν_0, which varies as we integrate over the passband, being equal to unity at the band center. Thus one can envisage the integrals in the square brackets in Eq. (6.74) as a process of averaging a large number of maps, each with a different scale factor. The scale factors are equal to ν/ν_0, and the range of values of ν is determined by the observing passband. The maps are

aligned at the origin, and thus the effect of the integration over frequency is to produce a radial smearing of the intensity distribution before it is convolved with the beam. The response to a point source at position (l, m) is radially elongated by a factor equal to $\sqrt{l^2 + m^2}\,\Delta\nu/\nu_0$. For distances from the origin at which the elongation is large compared to the synthesized beamwidth, features on the sky become attenuated by the smearing, so there is an effective limitation of the field of view. The measured intensity is the smeared distribution convolved with the synthesized beam.

Details of the behavior of the derived intensity distribution can be deduced from Eq. (6.74). For example, suppose that the beam contains a circularly symmetrical sidelobe at a large distance from the beam axis, and that in a map the response to a distant source causes the sidelobe to fall near the origin. Is the sidelobe broadened near the origin? Since the distant source is elongated, the sidelobe will be smeared in a direction parallel to that of a line joining the source and the origin, as shown in Fig. 6.10. It will be broadened near the origin, but not at a point 90° around the sidelobe as measured from the source.

To estimate the magnitude of the suppression of distant sources, it is useful to calculate R_b, the peak response to a point source at a distance r_1 from the origin of the (l, m) plane, as a fraction of the response to the same source at the origin. Because the effect we are considering is a radial smearing, we need only consider the intensity along a radial line through the (l, m) origin as shown in Fig. 6.11a. We use idealized parameters; the bandpass is represented by a rectangular function of width $\Delta\nu$ and the synthesized beam by a circularly symmetrical Gaussian function of standard deviation $\sigma_b = \theta_b/\sqrt{8 \ln 2}$, where θ_b is the half-power beamwidth. For simplicity the factor $(\nu/\nu_0)^2$ in the integral in the numerator of Eq. (6.74) is omitted, which is a reasonable approximation since the fractional bandwidth often does not exceed 5%. The convolution becomes a one-dimensional (radial) process, as shown in Fig. 6.11b. The radially elon-

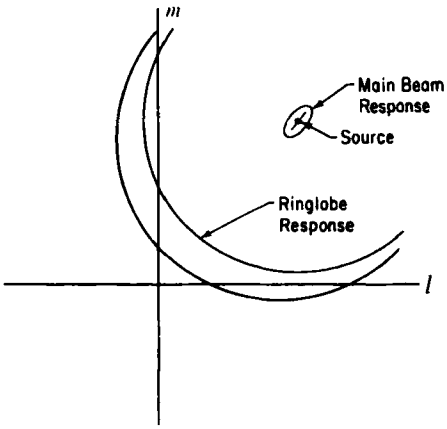


Figure 6.10 Radial smearing resulting from the bandwidth effect for a point source at (l_1, m_1). The effects on the responses of the main beam and a ringlobe (i.e., a sidelobe of the form in Fig. 5.15) are shown.

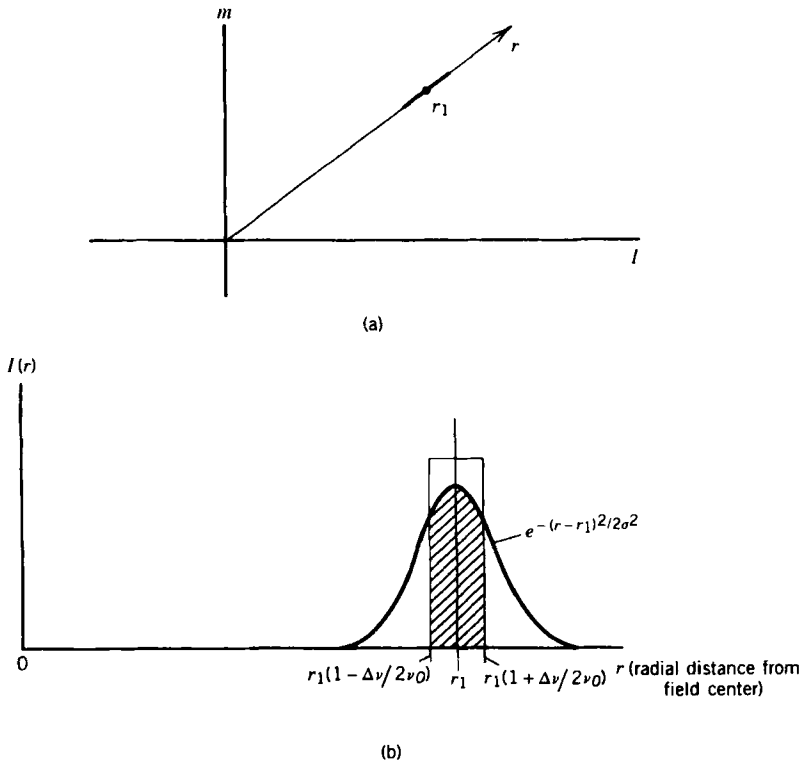


Figure 6.11 Response of an array with a broadband receiving system to a point source at distance r_1 from the origin of the (l, m) plane. (a) The point source (delta function) at r_1 becomes radially broadened into a rectangular function of unit area indicated by the heavy line. (b) Cross section of the intensity distribution in the r direction. The synthesized beam is represented by the Gaussian function. The peak intensity of the response to the source is proportional to the shaded area.

gated source is represented by a rectangular function from $r_1(1 - \Delta\nu/2\nu_0)$ to $r_1(1 + \Delta\nu/2\nu_0)$, normalized to unit area. The beam is represented by the function $e^{-r^2/2\sigma_b^2}$ which is normalized to unity on the beam axis. When the beam is centered on the source, as shown in Fig. 6.11, R_b is given by

$$R_b = \frac{\nu_0}{r_1\Delta\nu}\int_{r_1(1-\Delta\nu/2\nu_0)}^{r_1(1+\Delta\nu/2\nu_0)} e^{-(r-r_1)^2/2\sigma_b^2}\,dr = \sqrt{2\pi}\,\frac{\sigma_b\nu_0}{r_1\Delta\nu}\,\mathrm{erf}\left(\frac{r_1\Delta\nu}{2\sqrt{2}\,\sigma_b\nu_0}\right)$$
$$= 1.0645\frac{\theta_b\nu_0}{r_1\Delta\nu}\,\mathrm{erf}\left(0.8326\frac{r_1\Delta\nu}{\theta_b\nu_0}\right). \tag{6.75}$$

A curve of R_b as a function of the parameter $r_1\Delta\nu/\theta_b\nu_0$, which is the distance of the source from the origin measured in beamwidths, multiplied by the fractional

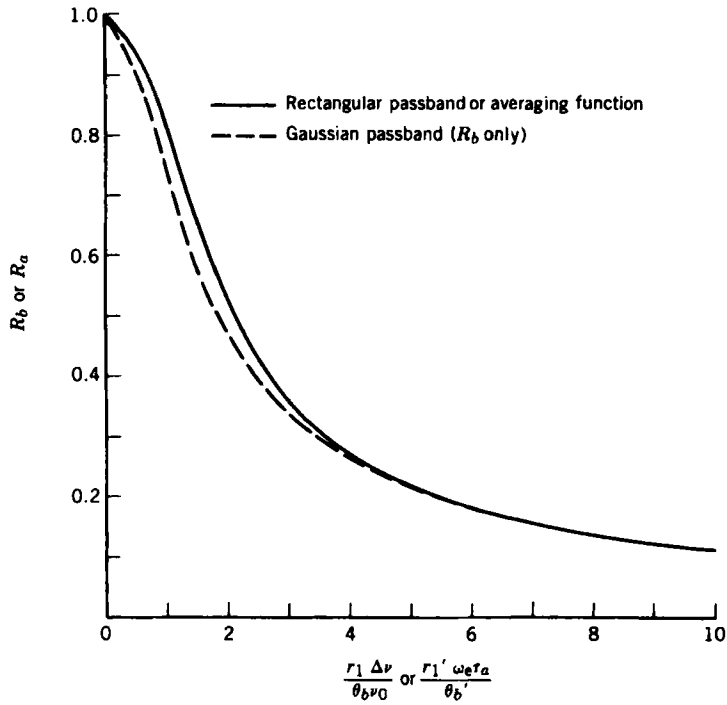


Figure 6.12 Relative amplitude of the peak response to a point source as a function of the distance from the field center and either the fractional bandwidth or the averaging time.

bandwidth, is shown in Fig. 6.12. Values of 0.2 and 0.5 for this parameter reduce the response by 0.9% and 5.5%, respectively.

If the receiving bandpass is represented by a Gaussian function of equivalent width $\Delta\nu$ (i.e., standard deviation $= \Delta\nu/2.5066$), the reduction factor becomes

$$R_b = \frac{1}{\sqrt{1 + (0.939 r_1 \Delta\nu/\theta_b \nu_0)^2}}. \tag{6.76}$$

A curve of this function is also included in Fig. 6.12. Comparison of the two curves indicates the dependence on the passband shape.

Wide-Field Mapping with a Multichannel System

Broadband maps can also be obtained by observing with a multichannel system (i.e., a spectral line system as described in Section 8.7). In this case the passband is divided into a number of channels by using either a bank of narrowband filters or a digital spectral correlator. The visibility is measured independently for each channel, so the values of u and v can be scaled correctly and an independent map obtained for each channel. This scaling causes the spatial sensitivity function to vary over the band, and at frequency ν the synthesized beam

is $(\nu/\nu_0)^2 b_0(l\nu/\nu_0, m\nu/\nu_0)$, where $b_0(l, m)$ is the monochromatic beam at frequency ν_0. The maps can be combined by summation, and if given equal weights, the result for N channels is represented by

$$I(l, m) ** \left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\nu_1}{\nu_0}\right)^2 b_0\left(\frac{l\nu_i}{\nu_0}, \frac{m\nu_i}{\nu_0}\right)\right]. \tag{6.77}$$

In this case there is no smearing of the intensity distribution, but the beam suffers a radial smearing that has the desirable effect of suppressing distant sidelobes. Therefore, this mode of observation is well suited for mapping wide fields. The improvement in the beam results from the increase in the number of (u, v) points measured, an effect that is also used in multifrequency synthesis discussed in Section 11.7

6.4 EFFECT OF VISIBILITY AVERAGING

Visibility Averaging Time

In most synthesis arrays the output of each correlator is averaged for consecutive time periods, τ_a, and thus consists of real or complex values spaced at intervals τ_a in time. It is advantageous to make τ_a long enough to keep the data rate from the correlator readout conveniently small. A limit on τ_a results from a consideration of the sampling theorem discussed in Section 5.2, and is briefly explained as follows. In discrete Fourier transformation of the visibility to intensity, the data points are spaced at intervals Δu and Δv, as shown in Fig. 5.3. If the size of the field to be mapped is θ_f in the l and m directions, then $\Delta u = \Delta v = 1/\theta_f$. In time τ_a, the motion of a baseline vector within the (u, v) plane should not be allowed to exceed Δu; otherwise the visibility data will not represent independent measurements, and information will be lost. Consider the case where the longest baseline is east–west in orientation and the source under observation is at a high declination, which results in the fastest motion of the baseline vector. If the baseline length is D_λ wavelengths, the vector in the (u, v) plane traces out an approximately circular locus, the tip of which moves at a speed of $\omega_e D_\lambda$ wavelengths per unit time, where ω_e is the angular velocity of rotation of the earth. Thus we require that $\tau_a \omega_e D_\lambda < 1/\theta_f$, which results, in practice, in $\tau_a \approx C/(\omega_e D_\lambda \theta_f)$, where C is a factor likely to be in the range 0.1–0.5. Note that $D_\lambda \theta_f$ is approximately the number of synthesized beamwidths across the field, and thus τ_a must be somewhat smaller than the time taken for the earth to rotate through one radian, divided by this number. Although shorter baselines could be averaged for longer times, in most synthesis arrays all correlator outputs are read at the same time, at a rate appropriate for the longest baselines. Another consideration is that sporadic interference and instrumental malfunctions can be edited out of the data with minimal information loss if τ_a is not too long. For large arrays τ_a is generally in the range of tens of milliseconds to tens of seconds. Determining the visibility at the $(\Delta u, \Delta v)$ grid points from the sampled data on the (u, v) loci

is discussed in Section 10.2 under *Mapping by Discrete Fourier Transformation*, and this process may also influence the choice of τ_a.

Effect of Time Averaging

We now examine in more detail the effect of the averaging on the synthesized intensity distribution. In reducing the data, all visibility values within each interval τ_a are treated as though they applied to the time at the center of the averaging period. Thus, the measurements at the beginning of each averaging period, for example, enter into the visibility data with assigned values of u and v that apply to times $\tau_a/2$ later than the true values. In effect, the resulting map consists of the average of a large number of maps, each with a different timing offset distributed progressively throughout the range $-\tau_a/2$ to $\tau_a/2$. These timing offsets apply only to the assignment of (u, v) values and do not resemble a clock error that affects the whole receiving system.

Consider an unresolved source, represented by a delta function. To simplify the situation, we consider observations with east–west baselines, and examine the effects in the (u', v') plane and the corresponding (l', m') sky plane (see Section 4.2). The spacing loci are circular arcs generated by vectors rotating at angular velocity ω_e, as shown in Fig. 6.13a. Consider first the case of an east–west linear array; then, of the antenna spacing components (X, Y, Z) defined in Fig. 4.1, only Y is nonzero. The circular arcs of the spacing loci are centered on the (u', v') origin as in Fig. 6.13b, and a timing offset δt is the equivalent to a rotation of the (u', v') axes through an angle $\omega_e\,\delta t$. The visibility of the source is the sum of two sets of sinusoidal corrugations, one real and one imaginary:

$$\delta(l_1', m_1') \rightleftharpoons \cos 2\pi(u'l_1' + v'm_1') - j\sin 2\pi(u'l_1' + v'm_1'). \tag{6.78}$$

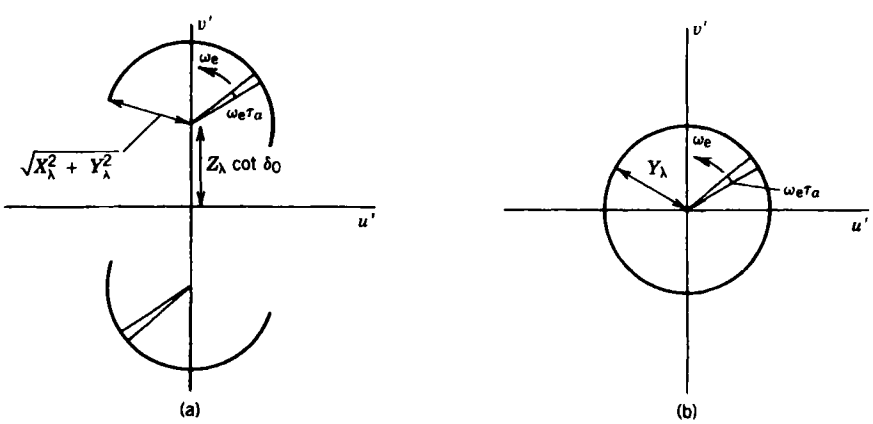


Figure 6.13 Spacing loci in the (u', v') plane, (a) for the general case and (b) for an east–west baseline. The angle $\omega_e\tau_a$ over which the averaging takes place is enlarged for clarity: for example, with an averaging time of 30 sec the angle would be 7.5 arcmin.

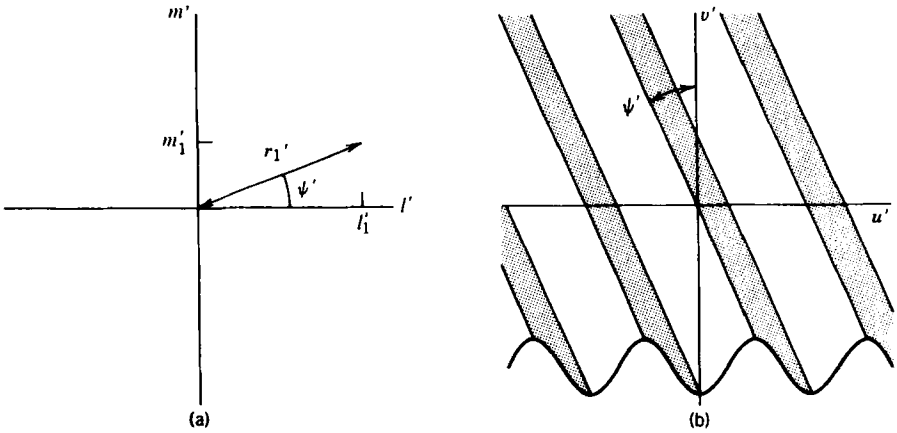


Figure 6.14 (a) Point source at (l'_1, m'_1) and (b) the real part of the corresponding visibility function. The ridges of the sinusoidal corrugations that represent the visibility in the (u', v') plane are orthogonal to the radius vector r'_1 at the position of the source in the (l', m') plane.

The angle of the corrugations is related to the position angle $\psi' = \tan^{-1}(m'_1/l'_1)$ of the point source, as shown in Fig. 6.14. A change in ψ' causes an equivalent rotation of the corrugations, and vice versa. For an east–west array, time offsets therefore correspond to proportional rotations of the intensity in the (l', m') plane. It follows that the effect of the time averaging is to produce a circumferential smearing similar to that resulting from the receiving bandwidth but orthogonal to it. If we express positions in the (l', m') plane in terms of the radial coordinates (r', ψ') shown in Fig. 6.14a, the map obtained from the averaged data can be expressed in terms of the sky brightness $I(r', \psi')$ by

$$I_a(r', \psi') = \left[\frac{1}{\omega_e \tau_a} \int_{-\omega_e \tau_a/2}^{\omega_e \tau_a/2} I(r', \psi')\, d\psi'\right] ** b_0(r', \psi'), \tag{6.79}$$

where b_0 is the synthesized beam.

The fractional decrease in the peak response to the point source is most easily considered in the (l', m') plane. With an east–west baseline the contours of the synthesized beam are approximately circular in the (l', m') plane, as long as the observing time is approximately 12 h, which results in spacing loci in the form of complete circles in the (u', v') plane. If we assume that the synthesized beam can be represented by a Gaussian function, as in the calculations for the bandwidth effect, the curve for the rectangular bandwidth in Fig. 6.12 can also be used for the averaging effect. In one case the spreading function is radial and of width $r_1 \Delta\nu/\nu_0$, and in the other it is circumferential and of width $r'_1 \omega_e \tau_a$. Thus, for the averaging effect, we can replace $r_1 \Delta\nu/\theta_b \nu_0$ in Eq. (6.75) and Fig. 6.12 (solid curve) by $r'_1 \omega_e \tau_a/\theta'_b$, noting that $r'_1 = \sqrt{l_1^2 + m_1^2 \sin^2 \delta_0}$ and θ'_b, the synthesized

beamwidth in the (l', m') plane, is equal to the east–west beamwidth in the (l, m) plane. Hence, for the decrease in the response to a point source resulting from averaging, we can write

$$R_a = 1.0645 \frac{\theta_b'}{r_1' \omega_e \tau_a} \operatorname{erf}\left(0.8326 \frac{r_1' \omega_e \tau_a}{\theta_b'}\right). \tag{6.80}$$

Generally, one chooses τ_a so that R_a is only slightly less than unity at any point in the map, in which case we can approximate the error function by the integral of the first two terms in the power series for a Gaussian function:

$$R_a \simeq 1 - \frac{1}{3}\left(\frac{0.8326 \omega_e \tau_a}{\theta_b'}\right)^2 (l_1^2 + m_1^2 \sin^2 \delta_0). \tag{6.81}$$

This is a useful formula for checking that τ_a is not too large.

Two aspects of the behavior predicted by Eq. (6.81) should be mentioned. First, if the source is near the m' axis and at a low declination, the averaging has very little effect. This is because the ridges of the sinusoidal corrugations of the visibility function then run approximately parallel to the u' axis, and in the transformation $u' = u \operatorname{cosec} \delta_0$ the period of the variations in the v direction is expanded by a large factor. In comparison, the arc through which any baseline vector moves in time τ_a is small, and hence the averaging has only a small effect on the visibility amplitude. Second, for a source on the l' axis, R_a is independent of δ_0. In this case the ridges of the corrugations run parallel to the v axis, and the expansion of the scale in the v direction has no effect on the sinusoidal period.

For arrays that contain baselines other than east–west, the centers of the corresponding loci in the (u', v') plane are offset from the origin, as in Fig. 6.13a, and a time offset is no longer equivalent to a simple rotation of axes. However, this may not increase the smearing of the visibility, so the effect is likely to be no worse than for an east–west array with baselines of similar lengths.

APPENDIX 6.1 PARTIAL REJECTION OF A SIDEBAND

In a single-sideband system using a mixer as the input stage, the unwanted (image) sideband may be rejected by one of several schemes. These include use of a waveguide filter, a Martin–Puplett interferometer [Martin and Puplett (1969), Payne (1989)], a tuned backshort, or a sideband-separating configuration of two mixers (as in Appendix 7.1). Practical considerations, particularly at millimeter wavelengths, can limit the rejection of the image sideband. Let the response to the image sideband, in terms of the power gain of the receiver, be ρ times the response to the wanted (signal) sideband, where $0 < \rho < 1$.

In the case of spectral line observation, where the wanted line occurs only in the signal sideband, the effect of the noise introduced by the image sideband is

to increase the rms noise at the correlator output by a factor $(1 + \rho)$. Thus the sensitivity is reduced by a factor $(1 + \rho)^{-1}$.

In the case of continuum observation, the image sideband also introduces a component of signal at the correlator. Assume that the visibility is the same in both sidebands, the fringes are stopped, and $\pi/2$ phase switching of the first local oscillator allows measurement of the complex visibility. A complex correlator is used, and for simplicity we consider that the instrumental phase is adjusted so that the line AB in Fig. 6.5b is coincident with the real axis. We can represent the complex correlator output with zero phase shift of the local oscillator as

$$C_0 = G_{mn}(\mathcal{V} + \rho\mathcal{V}^*), \tag{A6.1}$$

and with the $\pi/2$ phase switch as

$$C_{\pi/2} = G_{mn}(j\mathcal{V} - j\rho\mathcal{V}^*). \tag{A6.2}$$

Here G_{mn} is the gain in the signal sideband, so ρG_{mn} is the gain in the image sideband. Note that in the expression for $C_{\pi/2}$ the j factors have opposite signs for the two sidebands, because the $\pi/2$ phase shift causes the corresponding vectors in the complex plane to rotate through $\pi/2$ in opposite directions, as in Fig. A6.1. The optimum estimate of the visibility is then found to be

$$\mathcal{V} = \frac{1}{2G_{mn}}\left[\frac{1}{1+\rho^2}(C_0 - jC_{\pi/2}) + \frac{\rho}{1+\rho^2}(C_0 + jC_{\pi/2})^*\right]. \tag{A6.3}$$

The first term within square brackets represents the response of the signal sideband and the second the image. The total noise power delivered to the correlator input is proportional to $(1 + \rho)$, so the rms noise associated with the first term in

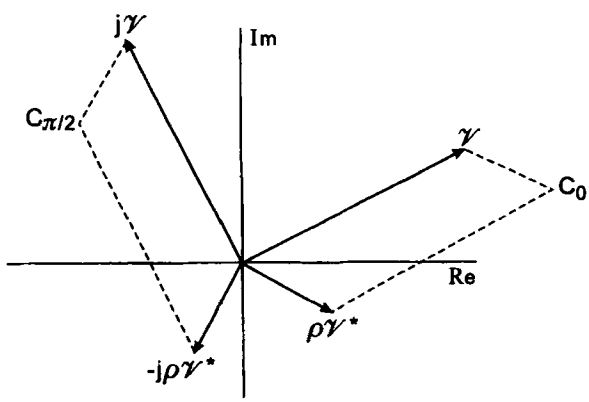


Figure A6.1 Vectors in the complex plane representing the parameters in Eqs. (A6.1) and (A6.2). The constant gain factor G_{mn} is omitted. If ρ is known and C_0 and $C_{\pi/2}$ are measured, Eq. (A6.3) gives the optimum estimate of $\mathcal{V}$.

the square brackets is proportional to $(1 + \rho)/(1 + \rho^2)$, and for the second term the equivalent expression is $\rho(1+\rho)/(1+\rho^2)$. Thus the rms noise in the estimate of $\mathcal{V}$ from (A6.3) is proportional to $(1 + \rho)/\sqrt{(1 + \rho^2)}$. The sensitivity is proportional to $\sqrt{(1 + \rho^2)}/(1 + \rho)$. For $\rho \approx 1/10$ or less, the ρ^2 term is very small and the sensitivity degradation factor is approximately $(1 + \rho)^{-1}$ (Thompson and D'Addario 2000).

REFERENCES

Anantharamaiah, K. R., R. D. Ekers, V. Radhakrishnan, T. J. Cornwell, and W. M. Goss, Noise in Images of Very Bright Sources, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Pub. Astron. Soc. Pacific Conf. Ser., **6**, 431–442, 1989.

Blum, E. J., Sensibilité des Radiotélescopes et Récepteurs à Corrélation, *Ann. d'Astrophys.*, **22**, 140–163, 1959.

Bracewell, R. N., *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 2000 (earlier eds. 1965, 1978).

Colvin, R. S., *A Study of Radio Astronomy Receivers*, Ph.D. thesis, Stanford Univ., Stanford, CA, 1961.

Douglas, J. N., F. N. Bash, F. D. Ghigo, G. F. Moseley, and G. W. Torrence, First Results from the Texas Interferometer: Positions of 605 Discrete Sources, *Astron. J.*, **78**, 1–17, 1973.

Goldstein, S. J., Jr., The Angular Size of Short-Lived Solar Radio Disturbances, *Astron. J.*, **130**, 393–399, 1959.

Lawson, J. L. and G. E. Uhlenbeck, *Threshold Signals*, Radiation Laboratory Series, Vol. 24, McGraw-Hill, New York, 1950, p. 68.

Lo, W. F., P. E. Dewdney, T. L. Landecker, D. Routledge, and J. F. Vaneldik, A Cross-Correlation Receiver for Radio Astronomy Employing Quadrature Channel Generation by Computed Hilbert Transform, *Radio Sci.*, **19**, 1413–421, 1984.

Martin, D. H. and E. Puplett, Polarized Interferometric Spectrometry for the Millimeter and Submillimeter Spectrum, *Infrared Phys.*, **10**, 105–109, 1969.

Mickelson, R. L. and Swenson, G. W. Jr., A Comparison of Two Correlation Schemes, *IEEE Trans. Instrum. Meas.*, **IM-40**, 816–819, 1991.

Middleton, D., *An Introduction to Statistical Communication Theory*, McGraw-Hill, New York, 1960, p. 343.

Moran, J. M., Very Long Baseline Interferometric Observations and Data Reduction, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 228–260.

Papoulis, A., *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 1965.

Payne, J. M., Millimeter and Submillimeter Wavelength Radio Astronomy, *Proc. IEEE*, **77**, 993–1017, 1989.

Read, R. B., Two-Element Interferometer for Accurate Position Determinations at 960 Mc, *IRE Trans. Antennas Propag.*, **AP-9**, 31–35, 1961.

Rice, S. O., Mathematical Analysis of Random Noise, *Bell Syst. Tech. J.*, **23**, 282–332, 1944; **24**, 46–156, 1945; repr. in *Noise and Stochastic Processes*, N. Wax, Ed., Dover, New York, 1954.

Rogers, A. E. E., Theory of Two-Element Interferometers, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, 1976, pp. 139–157 (see Table 1).

Swenson, G. W., Jr. and N. C. Mathur, On the Space-Frequency Equivalence of a Correlator Interferometer, *Radio Sci.*, **4**, 69–71, 1969.

Thompson, A. R. and L. R. D'Addario, *Relative Sensitivity of Double- and Single-Sideband Systems for both Total Power and Interferometry*, ALMA Memo. 304, National Radio Astronomy Observatory, Socorro, NM, 2000.

Tiuri, M. E., Radio Astronomy Receivers, *IEEE Trans. Antennas Propag.*, **AP-12**, 930–938, 1964.

Tiuri, M. E. and A. V. Räisänen, Radio-Telescope Receivers, in *Radio Astronomy*, J. D. Kraus, 2nd ed., Cygnus-Quasar Books, Powell, OH, 1986, Ch. 7.

Tucker, J. R. and M. J. Feldman, Quantum Detection at Millimeter Wavelengths, *Rev. Mod. Phys.*, **57**, 1055–1113, 1985.

Vander Vorst, A. S. and R. S. Colvin, The Use of Degenerate Parametric Amplifiers in Interferometry, *IEEE Trans. Antennas Propag.*, **AP-14**, 667–668, 1966.

Vinokur, M., Optimisation dans la Recherche d'une Sinusoide de Période Connue en Présence de Bruit, *Ann. d'Astrophys.*, **28**, 412–445, 1965.

Wright, M. C. H., B. G. Clark, C. H. Moore, and J. Coe, Hydrogen-Line Aperture Synthesis at the National Radio Astronomy Observatory: Techniques and Data Reduction, *Radio Sci.*, **8**, 763–773, 1973.

Wozencraft, J. M. and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965, p. 205.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

7 Design of the Analog Receiving System

The basic functions of the receiving system have been outlined in earlier chapters. Here we consider certain aspects of the system design in more detail. These concern mainly the equipment between the antennas and the correlators and, in particular, those characteristics of it that are critical to the accuracy and sensitivity of the visibility measurements. They include system noise temperature, phase stability, frequency responses, spurious signals, and automatic level control. The analysis leads to specification of tolerances on system parameters that are consistent with the goals of sensitivity and accuracy. Analog systems only are included, and digital sampling, delaying, and correlating of signals are the subject of Chapter 8.

7.1 PRINCIPAL SUBSYSTEMS OF THE RECEIVING ELECTRONICS

We give only a brief description of the main features of a receiving system. Optimum techniques and components for implementation of the electronic hardware vary continuously as the state of the art advances, and descriptions in the literature provide examples of the practical techniques current at various times; see, for example, Read (1961), Elsmore, Kenderdine, and Ryle (1966), Baars et al. (1973), Bracewell et al. (1973), Wright et al. (1973), Welch et al. (1977, 1996), Thompson et al. (1980), Batty et al. (1982), Erickson, Mahoney, and Erb (1982), Napier, Thompson, and Ekers (1983), Sinclair et al. (1992), Young et al. (1992), and Napier et al. (1994). The earlier papers in this list are mainly of interest from the viewpoint of the historical development of the technology.

Figure 7.1 shows a simplified schematic diagram of the receiving system of a large array of the linked-element type. For engineering convenience it is useful to divide the overall system into various subsystems that are outlined below.

Low-Noise Input Stages

In radio astronomy receivers, minimizing the noise temperature usually involves cryogenic cooling of the amplifier or mixer stages from the input up to a point at which noise from succeeding stages is unimportant. The low-noise input stages are often packaged with a cooling system, and sometimes also a feed horn, in

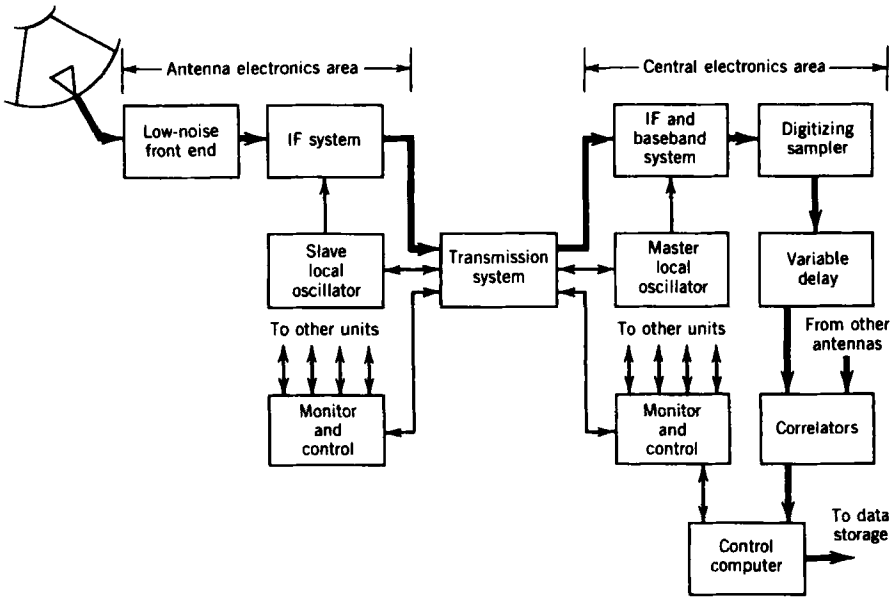


Figure 7.1 Simplified schematic diagram of the receiving system of a typical synthesis array. All the blocks indicate subsystems described in the text, except for the monitor and control blocks, which constitute the digital communication system through which the computer monitors critical voltages, sets local oscillator frequencies, etc. Except for the master local oscillator and the computer, one of each block is required per antenna. The heavy line shows the path of the received signal. In some systems the baseband and sampler units are located at the antennas, and the signals are transmitted in digital form.

a single package variously referred to as a *receiver* or a *front end*. The active components are usually transistor amplifiers or, for millimeter wavelengths, SIS (superconductor–insulator–superconductor) mixers followed by transistor amplifiers. For descriptions see, for example, Reid et al. (1973), Weinreb et al. (1977a), Weinreb, Fenstermacher, and Harris (1982), Casse, Woestenburg, and Visser (1982), Phillips and Woody (1982), Tiuri and Räisänen (1986), Payne (1989), Phillips (1994), Payne et al. (1994), Pospieszalski et al. (1997), and Webber et al. (1998).

In discussing the level of noise associated with a receiver, we begin by considering the case where the Rayleigh–Jeans approximation suffices. This is the domain in which $h\nu/kT \ll 1$, where h is Planck's constant and T is the temperature of the thermal noise source involved. As noted in the discussion following Eq. (1.1), this condition can be written as ν (GHz) $\ll 20T$ (K). It is convenient to specify noise power in terms of the temperature of a resistive load matched to the receiver input. In the Rayleigh–Jeans approximation, noise power available at the terminals of a resistor at temperature T is $kT\Delta\nu$, where k is Boltzmann's constant and $\Delta\nu$ is the bandwidth within which the noise is measured (Nyquist 1928). One kelvin of temperature represents a power spectral density of $(1/k)$ W Hz^{-1}.

The receiver temperature T_R is a measure of the internally generated noise power within the system and is equal to the temperature of a matched resistor at the input of a hypothetical noise-free (but otherwise identical) receiver that would produce the same noise power at the output. The system temperature, T_S, is a measure of the total noise level and includes, in addition to T_R, the noise power from the antenna and any lossy components between the antenna and the receiver:

$$T_S = T_A' + (L - 1)T_L + LT_R, \tag{7.1}$$

where T_A' is the antenna temperature resulting from the atmosphere and other unwanted sources of noise, L is the power loss factor of the transmission line from the antenna to the receiver [defined as (power in)/(power out)], and T_L is the temperature of the line. In defining the noise temperature of the receiver we should note that in practice a receiver is always used with the input attached to some source impedance which is itself a source of noise. The noise at the receiver output thus consists of two components, the noise from the source at the input, which is the antenna and transmission line in Eq. (7.1), and the noise generated within the receiver.

Noise Temperature Measurement

The noise temperature of a receiver is often measured by the Y-factor method. The thermal noise sources used in this measurement are usually impedance-matched resistive loads connected to the receiver input by waveguide or coaxial line. The receiver input is connected sequentially to two loads at temperatures T_{hot} and T_{cold}. The measured ratio of the receiver output powers in these two conditions is the factor Y:

$$Y = \frac{T_R + T_{\text{hot}}}{T_R + T_{\text{cold}}}, \tag{7.2}$$

and thus

$$T_R = \frac{T_{\text{hot}} - YT_{\text{cold}}}{Y - 1}. \tag{7.3}$$

Commonly used values are $T_{\text{hot}} = 290$ K (ambient temperature) and $T_{\text{cold}} = 77$ K (liquid nitrogen temperature).

The receiver temperature can be expressed in terms of the noise temperatures of successive stages through which the signal flows [see, e.g., Kraus (1966)]:

$$T_R = T_{R1} + T_{R2}G_1^{-1} + T_{R3}(G_1G_2)^{-1} + \cdots. \tag{7.4}$$

Here T_{Ri} is the noise temperature of the ith receiver stage and G_i is its power gain. If the first stage is a mixer instead of an amplifier, G_1 may be less than unity, and the second stage noise temperature then becomes very important.

For cryogenically cooled receivers for millimeter and shorter wavelengths, the Rayleigh–Jeans approximation can introduce significant errors. The power spectral density (power per unit bandwidth) of the noise is no longer linearly proportional to the temperature of the radiator or source. The ratio h/k is equal to 0.048 K per GHz, so if, for example, $T = 4$ K (liquid helium temperature), then $h\nu/kT = 1$ for $\nu = 83$ GHz. Thus quantum effects become important as frequency is increased and temperature decreased. Under these conditions the noise power per unit bandwidth divided by k provides an effective noise temperature that can be used in noise calculations, instead of the physical temperature. Two formulas are in use that give the effective temperature for a thermal source when quantum effects become important. One is the Planck formula and the other the Callen and Welton formula (Callen and Welton 1951). The effective noise temperatures for a waveguide carrying a single mode and terminated in a thermal load, or for a transmission line terminated in a resistive load, given by the two formulas are as follows:

$$T_{\text{Planck}} = T\left[\frac{\frac{h\nu}{kT}}{e^{h\nu/kT} - 1}\right] \tag{7.5}$$

$$T_{\text{C\&W}} = T\left[\frac{\frac{h\nu}{kT}}{e^{h\nu/kT} - 1}\right] + \frac{h\nu}{2k}, \tag{7.6}$$

where T is the physical temperature. From Eqs. (7.5) and (7.6), we obtain

$$T_{\text{C\&W}} = T_{\text{Planck}} + \frac{h\nu}{2k}. \tag{7.7}$$

The Callen and Welton formula is equal to the Planck formula with an additional term, $h\nu/2k$, which represents an additional half photon. This half photon is the noise level from a body at absolute zero temperature and is referred to as the zero-point fluctuation noise. Figure 7.2 shows the relationships between physical temperature and noise temperature corresponding to the Rayleigh–Jeans, Planck, and Callen and Welton formulas, for a frequency of 230 GHz. Note that for the case of $h\nu/kT \ll 1$ we can put $\exp(h\nu/kt) - 1 \simeq (h\nu/kt) + \frac{1}{2}(h\nu/kt)^2$, in which case the Callen and Welton formula reduces to the Rayleigh–Jeans formula, but the result from the Planck formula is lower by $h\nu/2k$.

When using Eq. (7.3) to derive the noise temperature of a receiver, the values of T_{hot} and T_{cold} should be the noise temperatures derived from the Planck or Callen and Welton formulas, not the physical temperatures of the loads (except in the Rayleigh–Jeans domain). Thus for the Planck formula we can write

$$T_{R(\text{Planck})} = \frac{T_{\text{hot(Planck)}} - Y T_{\text{cold(Planck)}}}{Y - 1}, \tag{7.8}$$

and a similar equation for the Callen and Welton formula. From Eqs. (7.4), (7.5), and (7.6) we obtain

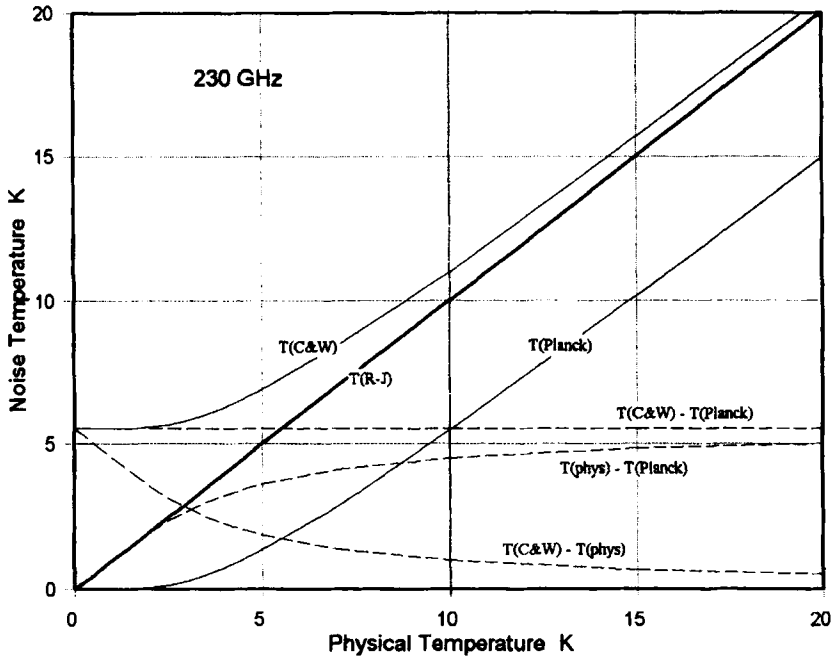


Figure 7.2 Noise temperature versus physical temperature for blackbody radiators at 230 GHz, according to the Rayleigh–Jeans, Planck, and Callen and Welton formulas. Also shown (broken lines) are the differences between the three radiation curves. The Rayleigh–Jeans curve converges with the Callen and Welton curve at high temperature, while the Planck curve is always $h\nu/2k$ below the Callen and Welton curve. From Kerr et al. (1997).

$$T_{R(\text{Planck})} = T_{R(\text{C\&W})} + \frac{h\nu}{2k}. \tag{7.9}$$

In using any measurement of receiver noise temperature it is important to know whether, in deriving it, the Planck formula, the Callen and Welton formula, or the physical temperature of the loads (i.e., the Rayleigh–Jeans approximation) was used. If the noise temperatures of the individual components are derived from the physical temperatures using the Callen and Welton formula, the temperature sum will be greater by $h\nu/2k$ than if the Planck formula were used; see Eq. (7.7). However, if the Callen and Welton formula is used to derive the receiver noise temperature, the result will be less by $h/2k$ than if the Planck formula were used; see Eq. (7.9). Thus the system temperature, which is the sum of the input temperature and the receiver temperature, will be the same whichever of the two formulas is used. However, to avoid confusion, it is important to use one formula or the other consistently throughout the derivation of the noise temperatures.

Differing opinions have been expressed on the nature of the zero-point fluctuation noise, and whether it should be considered as originating in the load connected to the receiver or in the receiver input stages; see, for example, Tucker and Feldman (1985), Zorin (1985), and Wengler and Woody (1987). At frequencies

at which quantum effects become most important, the usual type of input stage in radio astronomy receivers is the SIS (superconductor–insulator–superconductor) mixer, for which the quantum theory of operation is given by Tucker (1979). For a summary of some conclusions from various authors relevant to noise temperature considerations, see Kerr, Feldman, and Pan (1997) and Kerr (1999).

To recapitulate—the radiation level predicted by the Callen and Welton formula is equal to the Planck radiation level plus the zero-point fluctuation component $h\nu/2$. The latter component is attributable to the power from a blackbody or matched resistive load at absolute zero temperature. An amplifier noise temperature derived using the Callen and Welton formula to interpret the measured Y factor is lower than that derived using the Planck formula by $h\nu/2k$. However, an antenna temperature obtained using the Callen and Welton formula is higher by $h\nu/2k$ than the corresponding Planck formula value. The system temperature, which is the sum of the noise temperature and the antenna temperature is the same in either case. Since it is the system temperature that determines the sensitivity of a radio telescope, these details may seem unimportant. However, in procuring an amplifier or mixer for a receiver input stage it is important to know how the noise temperature is specified.

In addition to the noise generated in the electronics, the noise in a receiving system contains components that enter from the antenna. These components arise from cosmic sources, the cosmic background radiation, the earth's atmosphere, the ground, and other objects in the sidelobes of the antenna. The opacity of the atmosphere, from which the atmospheric contribution to the system noise arises, is discussed in Chapter 13.

Local Oscillator

As explained in the previous chapter, local oscillator signals are required at the antennas and often at other points along the signal paths to the correlators. The corresponding oscillator frequencies for different antennas must be maintained in phase synchronism to preserve the coherence of the signals. The phases of the oscillators at corresponding points on different antennas need not be identical, but the differences should be stable enough to permit calibration. Maintaining synchronism at different antennas requires transmitting one or more reference frequencies from a central master oscillator to the required points, where they may be used to phase-lock other oscillators. The frequencies required at the mixers can then be synthesized.

Special phase shifts are required at certain mixers to implement fringe rotation (fringe stopping), as described in Section 6.1 under *Delay Tracking and Fringe Rotation*, and phase switching, described in Section 7.5. Often these can best be implemented by digital synthesis techniques, such as the use of an integrated circuit device known as a number-controlled oscillator. This can provide a signal at a frequency of, say, a few megahertz that contains the required frequency offsets and phase changes. The phase changes can be transferred to the local oscillator frequency by using the synthesized signal as a reference frequency in a phase-locked loop.

IF and Signal Transmission Subsystems

After amplification in the low-noise front-end stages, the signals pass through various IF amplifiers and a transmission system before reaching the correlators. Transmission between the antennas and a central location can be effected by means of coaxial or parallel-wire lines, waveguide, optical fibers, or direct radiation by microwave radio link. Cables are often used for small distances, but for long distances the cable attenuation may require the use of too many line amplifiers, and optical fiber, for which the transmission loss is much lower, is generally preferred. Low-loss TE_{01}-mode waveguide (Weinreb et al. 1977b; Archer, Caloccia, and Serna 1980) was used in the construction of the VLA, which preceeded the development of optical fiber by a few years. Cable or optical fiber can be buried at depths of 1–2 m to reduce temperature variations. Bandwidths of signals transmitted by cables are usually limited to some tens or hundreds of megahertz by attenuation, and radio links are similarly limited by available frequency allocations. For very wide bandwidths optical fibers offer the greatest possibilities.

After arriving at the correlator location, the received signals are usually converted to a final intermediate frequency where bandwidth selection filters and compensating time delays are inserted. Phase errors, resulting from temperature effects in filters, and delay-setting errors can be minimized by using the lowest possible intermediate frequency at this point. Accordingly, the final IF amplifiers often have a *baseband* response defined by a lowpass filter. The response at the low-frequency end falls off at a frequency that is a few percent of the upper cutoff frequency. In instruments that use a digital correlator for spectral line observations, a series of filters with bandwidths varying in steps of a factor of two are commonly provided, just preceding the digital samplers. These bandwidths are chosen to match the characteristics of a digital correlator as described in Section 8.7. In some cases an image-rejection mixer (see Appendix 7.1) is used for the conversion to baseband, but the suppression of the unwanted sideband is then generally no greater than 20–30 dB.

Optical Fiber Transmission

The introduction of optical fiber systems provided a very great advance in transmission capability for broadband signals over long distances. Signals are modulated onto optical carriers, commonly in the wavelength range 1300–1550 nm, and transmitted along glass fiber. The fiber attenuation is a minimum of approximately 0.2 dB km^{-1} near 1550 nm, and is about 0.4 dB km^{-1} at 1300 nm. These values are much lower than can be obtained in radio frequency transmission lines. In the fiber, a glass core is surrounded by a glass cladding of lower refractive index, so light waves launched into the core at a small enough angle with respect to the axis of the fiber can propagate by total internal reflection. If the inner-core diameter is approximately 50 μm, a number of different modes can be supported. These modes travel with slightly different velocities, which results in a limitation in performance of this multimode fiber. If the core is reduced to approximately 10 μm in diameter, only the HE_{11} mode propagates. Single-mode fiber of this

type is required for the longest distances and/or the highest frequencies and bandwidths. At 1550 nm an interval of 1 nm in wavelength corresponds to a bandwidth of approximately 125 GHz. The low attenuation and the bandwidth capacity facilitate the use of wide bandwidths and long baselines in linked-element arrays. Signals can be transmitted in analog form or digitized (as described in Chapter 8) and transmitted as pulse trains. Since fiber transmission involves the characteristics of the lasers that generate the optical carriers and the detectors that recover the modulation, as well as the characteristics of the fiber, the design of fiber transmission systems is rather more complicated than that of systems using cable. Here we briefly review a few basic features of fiber transmission. For further information, see, for example, Agrawal (1992) or Borella et al. (1997).

In practice the bandwidth and distance of the transmission are limited by the noise in the laser that generates the optical signal at the transmitting end of the fiber and the noise in the diode demodulator and the amplifier at the receiving end. To avoid degradation of the sensitivity in analog transmission, the power spectral density of the signal (measured in $\mathrm{W\,Hz^{-1}}$) must be greater than the power spectral density of the noise generated in the transmission system by $\sim$20 dB for most radio astronomy applications. However, the total signal power is limited by the need to avoid nonlinearity of the response of the modulator or demodulator. The result is a limit on the bandwidth of the signal, since for signals with a flat spectrum the power is proportional to the bandwidth. In practice, the limit for a single transmitter and receiver pair is a bandwidth of 10–20 GHz for transmission distances of some tens of kilometers. Optical amplifiers, which most commonly operate at wavelengths near 1550 nm, can be used to increase the range of transmission.

In the modulation process, the *power* of the carrier is varied in proportion to the *voltage* of the signal. Because of this, the effect of small unwanted components in fiber transmission systems are greatly reduced. Consider, for example, a small component of the optical signal resulting from a reflection within the fiber. If the optical power of the reflected component is x dB less than that of the main component, then after demodulation at the photodetector the signal power contributed by the reflected component is $2x$ dB less than that from the main optical component. This also applies to small unwanted effects resulting from finite isolation of couplers, isolators, and other elements. Variations in the frequency response resulting from standing waves in microwave transmission lines are significantly less in optical fiber than in cable.

A feature that must be taken into account in applications of optical fiber is the dispersion in velocity, $\mathcal{D}$, usually specified in $\mathrm{ps(nm \cdot km)^{-1}}$. The difference in the time of propagation for two optical wavelengths that differ by $\Delta\lambda$ traveling a distance ℓ in the fiber is $\mathcal{D}\,\Delta\lambda\ell$. Figure 7.3 shows the dispersion for two types of fiber. Curve 1 is for a type of fiber widely used in early applications, and curve 2 represents a design in which the zero-dispersion wavelength is shifted to coincide approximately with the minimum-attenuation wavelength of 1550 nm. This optimization of the performance at 1550 nm is achieved by designing the fiber so that the dispersion of the cylindrical waveguide formed by the core of the fiber cancels the intrinsic dispersion of the glass at that wavelength.

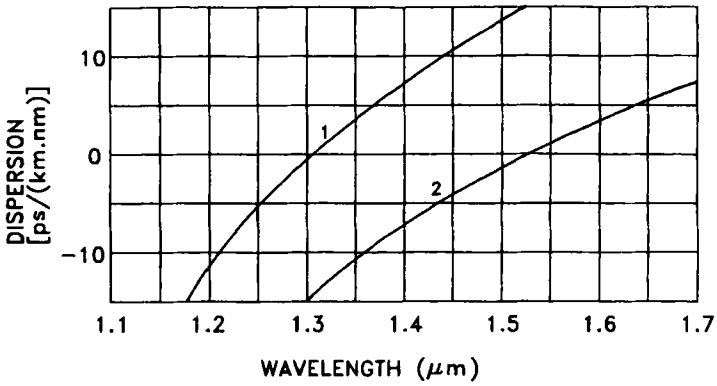


Figure 7.3 Dispersion $\mathcal{D}$ in single-mode optical fiber of two different designs, as a function of the optical wavelength.

Consider a spectral component, at frequency ν_m, of a broadband signal that is modulated onto an optical carrier. Amplitude modulation of the signal results in sidebands spaced $\pm\nu_m$ in frequency on each side of the carrier. Because of the velocity dispersion, the two sidebands and the carrier each propagate down the fiber with slightly different velocities, and thus exhibit relative offsets in time at the receiving end. Such time offsets result in attenuation of the amplitude of the high-frequency components of analog signals, and broadening of the pulses used to represent digital data. Thus, for both analog and digital transmission, dispersion as well as noise can limit the bandwidth × distance product. An analysis of the effect of dispersion on analog signals is given in Appendix 7.2.

Delay and Correlator Subsystems

The compensating delays and correlators can be implemented by either analog or digital techniques. An analog delay system may consist of a series of switchable delay units with a binary sequence of values in which the delay of the nth unit is $2^{n-1}\tau_0$, where τ_0 is the delay of the smallest unit. Such an arrangement, with N units, provides a range of delay from zero to $(2^N - 1)\tau_0$ in steps of τ_0. For delays up to about $1\,\mu$s, lengths of coaxial cable or optical fiber can be used. For longer delays, cables become unwieldy and acoustic-wave devices can provide larger increments. Systems with analog delays usually have analog correlators. The design of analog multiplying circuits has been discussed by Allen and Frater (1970). An example of a broadband analog correlator is described by Padin (1994). In spectral-line correlator systems of the analog type the final IF amplifier contains a bank of filters, the center frequencies of which are spaced at intervals equal to the filter bandwidth. Each filter defines a signal channel, and a separate correlator is required for each channel of every antenna pair.

The development of digital circuitry capable of operating at high clock frequencies has led to the practice of digitizing the final IF signal, so that the delay

and correlators can be implemented digitally. These digital systems are discussed in Chapter 8. Their advantage is that greater precision can be achieved in the visibility measurements, since with analog delays it is very difficult to keep the bandpass response from varying as different units are switched into and out of the signal channels. It is also difficult to maintain accurate calibration of the delay of long analog units unless they are kept at a constant temperature.

7.2 LOCAL OSCILLATOR AND GENERAL CONSIDERATIONS OF PHASE STABILITY

Round-Trip Phase Measuring Schemes

Synchronizing of the oscillators at the antennas can be accomplished by phase locking them to a reference frequency that is transmitted out from a central master oscillator. Buried cables or fibers offer the advantage of the greatest stability of the transmission path. At a depth of 1–2 m the diurnal temperature variation is almost entirely eliminated, but the annual variation is typically attenuated by a factor of 2–10 only. For a discussion of temperature variation in soil as a function of depth, see the *Handbook of Geophysics and Space Environments* (USAF 1965). As an example, a 10-km-long buried cable with a temperature coefficient of length of $10^{-5}\ \mathrm{K}^{-1}$ might suffer a diurnal temperature variation of 0.1 K, resulting in a change of 1 cm in electrical length. This change would cause a variation of 12° in the phase of a 1-GHz signal traversing the cable. An equal variation would occur in a 50-m length of cable running from the ground to the receiver enclosure on an antenna and subjected to a diurnal temperature variation of 20 K. Rotating joints and flexible cables can also contribute to phase variations.

Path length variations can be determined by monitoring the phase of a signal of known frequency that traverses the path. It is necessary for the signal to travel in two directions, that is, out from the master oscillator and back again, since the master provides the reference against which the phase must be measured. This technique is described as *round-trip phase measurement*. Correction for the measured phase changes can be implemented in hardware by using a phase shifter driven by the measurement system, or in software by inserting corrections in the data from the correlator, either in real time or during the later stages of data analysis. It is also possible to generate a signal in which the phase changes are greatly reduced by combining signals that travel in opposite directions in the transmission line. As an illustration of the last procedure, consider a signal applied to the near end of a loss-free transmission line that results in a voltage $V_0 \cos(2\pi \nu t)$ at the far end. At a distance ℓ, measured back from the far end, the outgoing signal is $V_1 = V_0 \cos 2\pi \nu (t + \ell/v)$, where v is the phase velocity along the line. Suppose that the signal is reflected from the far end without change in phase. At the same point, distant ℓ from the far end, the returned signal is $V_2 = V_0 \cos 2\pi \nu (t - \ell/v)$, and the total signal voltage is

$$V_1 + V_2 = 2V_0 \cos(2\pi \nu t) \cos\left(\frac{2\pi \nu \ell}{v}\right). \tag{7.10}$$

The first cosine function in Eq. (7.10) represents the radio frequency signal, the phase of which (modulo π) is independent of ℓ and of line length variations. The second cosine function is a standing-wave amplitude term. Such a system cannot easily be implemented in practice because of attenuation and unwanted reflections, and thus more complicated schemes have evolved. In what follows we consider cable transmission, although the basic principles are applicable to other systems. Some general considerations, including the use of microwave links, are given by Thompson et al. (1968).

Swarup and Yang System

Several different round-trip schemes have been devised as instruments have developed, and one of the earliest of these was by Swarup and Yang (1961). A system based on this scheme is shown in Fig. 7.4. Part of the outgoing signal is reflected from a known reflection point at an antenna, and variation in the path length to the reflector is monitored by measuring the relative phase of the reflected component at the detector. The phase of the reflected signal is compared with that of a reference signal. The phase of the latter is variable by means of a movable probe that samples the outgoing signal. Since many other reflections may occur in the transmission line, it is necessary to identify the desired component. To do this a modulated reflector, for example a diode loosely coupled to the line, is used.

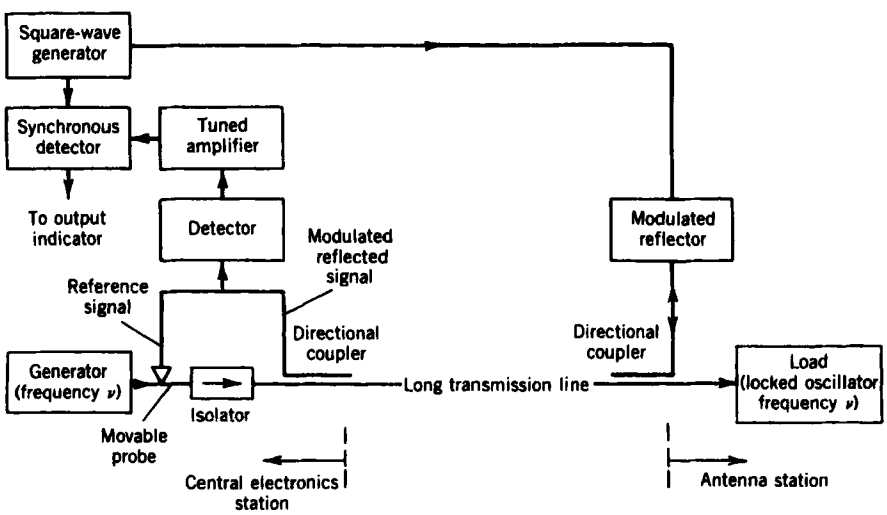


Figure 7.4 System for measuring variations in the electrical path length in a transmission line, based on the technique of Swarup and Yang (1961). The output of the synchronous detector is a sinusoidal function of the difference between the phases of the reference (outgoing) and reflected components at the detector. A null output is obtained when these signal phases are in quadrature, and the position of the probe for a null is thus a measure of the phase of the reflected signal. Because of the isolator in the line, the probe samples only the outgoing component of the signal.

This is switched between conducting and nonconducting states by a square wave voltage, and a synchronous detector is used to separate the modulated component of the reflected signal.

An increase $\Delta\ell$ in the length of the transmission line is detected as a corresponding movement of $2\Delta\ell$ in the probe position for the null. It results in an increase of $2\pi\Delta\ell\nu_1/v$ in the phase of the frequency ν_1 at the antenna, where v is the phase velocity in the line. The corresponding changes in local oscillator phases and IF phases transmitted over the same path can be calculated and applied as a correction to the visibility phases. Alternatively, the correction can be applied directly to the signals through a phase shifter or a mechanical line extender. In the original application by Swarup and Yang, the transmission line was part of a branching feeder network to an array of antennas.

Frequency-Offset Round-Trip System

A second scheme, shown in Fig. 7.5, is one in which the round-trip phase is measured directly. The signals traveling in opposite directions are at frequencies ν_1 and ν_2 that differ by only a small amount, but enough to enable them to be separated easily. This type of system is widely used, and we examine its performance in some detail. Note that although directional couplers or circulators allow

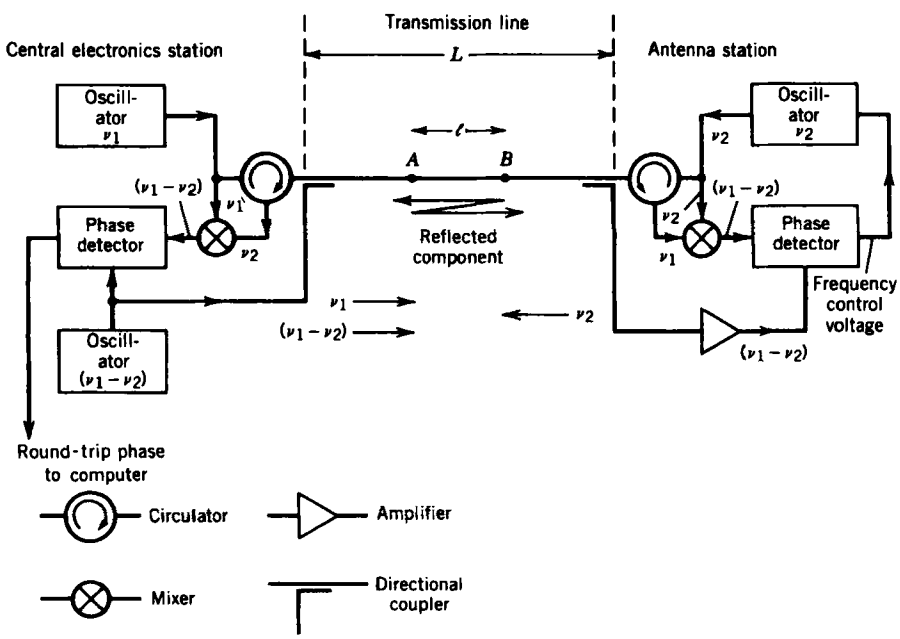


Figure 7.5 Phase-lock scheme for the oscillator ν_2 at the antenna. Frequencies ν_1 and $\nu_1 - \nu_2$ are transmitted to the antenna station where they provide the phase reference to lock the oscillator. ν_1 and ν_2 are almost equal, so $\nu_1 - \nu_2$ is small. A signal at frequency ν_2 is returned to the central station for the round-trip phase measurement.

the signals at the same frequency but going in opposite directions in the line to be separated, the signal from the unwanted direction is suppressed by only 20–30 dB relative to the wanted one. An unwanted component at a level of −30 dB can cause a phase error of 1.8°. However, the frequency offset enables the signals to be separated with much higher isolation.

An oscillator at frequency ν_2 at an antenna is phase locked to the difference frequency of signals at ν_1 and $\nu_1 - \nu_2$, which travel to the antenna via a transmission line. The difference frequency $(\nu_1 - \nu_2)$ is small compared with ν_1 and ν_2. The frequency ν_2 is returned to the master oscillator location for the round-trip phase comparison.

At the antenna, the phases of the signals at frequencies ν_1 and $\nu_1 - \nu_2$ relative to their phases at the central location are $2\pi\nu_1 L/v$ and $2\pi(\nu_1 - \nu_2)L/v$, where L is the length of the cable. The phase of the ν_2 oscillator at the antenna is constrained by a phase-locked loop to equal the difference of these phases, that is, $2\pi\nu_2 L/v$. The phase change in the ν_2 signal in traveling back to the central location is $2\pi\nu_2 L/v$, and thus the measured round-trip phase (modulo 2π) is $4\pi\nu_2 L/v$. Now suppose that the length of the line changes by a small fraction, β. The phase of the oscillator ν_2 at the antenna relative to the master oscillator changes to $2\pi\nu_2 L(1+\beta)/v$. The required correction to the ν_2 oscillator is just half the change in the measured round-trip phase. The problem that arises is that several effects, including reflections and velocity dispersion in the transmission line, can cause an error in the round-trip phase. Such an error results in a phase offset of the oscillator at the antenna, which is not serious if it remains constant. However, in practice it is likely to vary with ambient temperature. The largest error usually results from reflections, and control of this error places an upper limit on the frequency difference $\nu_1 - \nu_2$. We now examine this limit.

Consider what happens if reflections occur at points A and B separated by a distance ℓ along the line as in Fig. 7.5. The complex voltage reflection coefficients at these points are ρ_A and ρ_B, and their values will be assumed to be the same at frequencies ν_1 and ν_2. Signals ν_1 and ν_2, after traversing the cable, include components that have been reflected once at A and once at B. The coefficients ρ_A and ρ_B are sufficiently small that components suffering more than one reflection at each point can be neglected. For the frequency ν_1 arriving at the antenna, the amplitude (voltage) of the reflected component relative to the unreflected one is

$$\Lambda = |\rho_A||\rho_B|10^{-\ell\alpha/10}, \tag{7.11}$$

where α is the (power) attenuation coefficient of the cable in decibels per unit length. Note that the attenuation in voltage is equal to the square root of the attenuation in power. The phase of the reflected component relative to the unreflected one is, modulo 2π,

$$\theta_1 = 4\pi\ell\nu_1 v^{-1} + \phi_A + \phi_B, \tag{7.12}$$

where ϕ_A and ϕ_B are the phase angles of ρ_A and ρ_B (that is, $\rho_A = |\rho_A|e^{j\phi_A}$, etc.), and v is the phase velocity in the line. Figure 7.6 shows a phasor representation of the reflected and unreflected components and their phase θ_1. The reflected

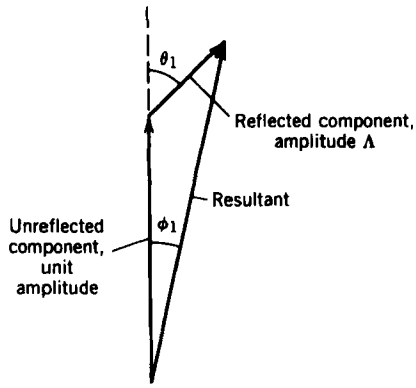


Figure 7.6 Phasor diagram of components at frequency ν_1 transmitted by the cable.

component causes the resultant phase to be deflected through an angle ϕ_1 given by

$$\phi_1 \simeq \tan\phi_1 = \frac{\Lambda \sin\theta_1}{1 + \Lambda\cos\theta_1}. \tag{7.13}$$

Similarly, the phase of the frequency ν_2 is deflected through an angle ϕ_2, given by equations equivalent to Eqs. (7.12) and (7.13) with subscript 1 replaced by 2.

With the reflection effects represented by ϕ_1 and ϕ_2, the round-trip phase for a line of length L is

$$4\pi\nu_2 L v^{-1} + \phi_1 + \phi_2. \tag{7.14}$$

If the line length increases uniformly to $L(1+\beta)$, the angles ϕ_1 and ϕ_2 vary in a nonlinear manner with ℓ and become $\phi_1 + \delta\phi_1$ and $\phi_2 + \delta\phi_2$, respectively. The round-trip phase then becomes

$$4\pi\nu_2 L v^{-1}(1+\beta) + \phi_1 + \delta\phi_1 + \phi_2 + \delta\phi_2. \tag{7.15}$$

(The effect of the reflection on the phase of the signal at frequency $\nu_1 - \nu_2$ has been omitted since $\nu_1 - \nu_2$ is much smaller than ν_1 or ν_2, and reflections for the relatively low frequency may be very small. Also, the rate of change of phase of $\nu_1 - \nu_2$ with line length is correspondingly small.) The applied correction for the increase in line length is half the measured change in round-trip phase:

$$2\pi\nu_2\beta L v^{-1} + \tfrac{1}{2}(\delta\phi_1 + \delta\phi_2). \tag{7.16}$$

However, the exact correction would be equal to the change in the phase of ν_2 at the antenna, which is

$$2\pi\nu_2\beta L v^{-1} + \delta\phi_2. \tag{7.17}$$

Consequently, the phase correction is in error by

$$\tfrac{1}{2}(\delta\phi_1 + \delta\phi_2) - \delta\phi_2 = \tfrac{1}{2}(\delta\phi_1 - \delta\phi_2). \tag{7.18}$$

If ν_1 and ν_2 were equal, the phase error would be zero. It is possible therefore to specify a maximum allowable frequency difference in terms of the maximum tolerable error.

The difference between the phase angles ϕ_1 and ϕ_2 is obtained from Eq. (7.13) as follows:

$$\begin{aligned}\phi_1 - \phi_2 &= \frac{\partial\phi_1}{\partial\nu_1}(\nu_1 - \nu_2)\\ &= \frac{4\pi\ell v^{-1}\Lambda\cos\theta_1(1 + \Lambda\cos\theta_1) + 4\pi\ell v^{-1}\Lambda^2\sin^2\theta_1}{(1 + \Lambda\cos\theta_1)^2}(\nu_1 - \nu_2).\end{aligned} \tag{7.19}$$

The reflected amplitude Λ must be much less than unity if phase errors are to be tolerable, so terms in Λ^2 can be omitted from the numerator in Eq. (7.19), and the denominator is approximately unity. Thus,

$$\phi_1 - \phi_2 \simeq 4\pi\ell v^{-1}\Lambda(\nu_1 - \nu_2)\cos\theta_1. \tag{7.20}$$

The variation of $\phi_1 - \phi_2$ with line length is given by

$$\begin{aligned}\delta\phi_1 - \delta\phi_2 &= \beta\ell\frac{\partial}{\partial\ell}(\phi_1 - \phi_2)\\ &= 4\pi v^{-1}\Lambda[\cos\theta_1 - 0.1\ell\alpha(\ln 10)\cos\theta_1 - 4\pi v^{-1}\ell\nu_1\sin\theta_1]\\ &\quad\times(\nu_1 - \nu_2)\beta\ell.\end{aligned} \tag{7.21}$$

The maximum values of the terms in square brackets in Eq. (7.21) are dominated by the third term, which is of the order of the number of wavelengths in the line. If the two smaller terms are neglected, we obtain the magnitude of the phase error as follows:

$$\tfrac{1}{2}(\delta\phi_1 - \delta\phi_2) \simeq 8\pi^2 v^{-2}|\rho_A||\rho_B|\beta\ell^2 10^{-\alpha\ell/10}\nu_1(\nu_1 - \nu_2)\sin\theta_1. \tag{7.22}$$

The factor $\ell^2 10^{-\alpha\ell/10}$ has a maximum value at

$$\ell = 20(\alpha\ln 10)^{-1}. \tag{7.23}$$

This maximum occurs because for small values of ℓ the change in the angle θ with frequency or cable expansion is small, and for large values of ℓ the reflected component is greatly attenuated. The maximum value is equal to

$$[\ell^2 10^{-\alpha\ell/10}]_{\max} = 10.21\alpha^{-2}. \tag{7.24}$$

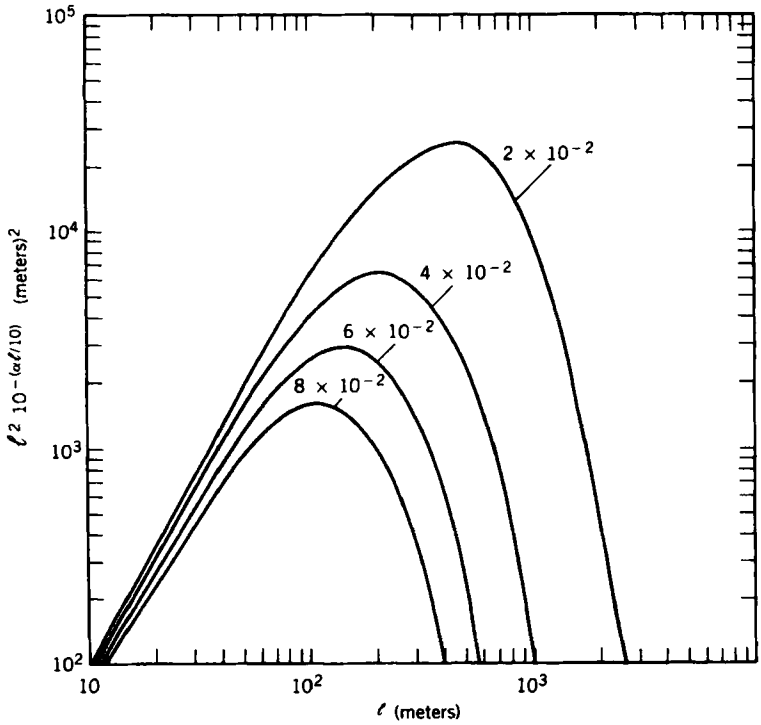


Figure 7.7 The function $\ell^2 10^{-\alpha\ell/10}$ plotted against ℓ for four values of the transmission-line attenuation, α dB m^{-1}. This function is a factor in the round-trip phase error given by Eq. (7.22).

Curves of $\ell^2 10^{-\alpha\ell/10}$ are plotted in Fig. 7.7 for various values of α that correspond to good-quality cables. It is evident that reducing the attenuation in a cable increases the error in the round-trip phase correction in Eq. (7.22).

The type of reflections that may be encountered depends on the type of transmission line and how it is used. For example, consider a buried coaxial cable that runs along a set of stations used for a movable antenna. The principal cause of reflections in such a cable is the connectors that are inserted at the antenna stations. Unless the antenna is at the closest station, there are one or more interconnecting loops, where unused stations are bypassed, between the antenna and the master oscillator. If there are n connectors in the cable, there are $N = n(n-1)/2$ pairs between which reflections can occur. Also, if the phasors of the corresponding reflected components combine randomly, the overall rms error in the phase correction is, from Eq. (7.22),

$$\delta\phi_{\text{rms}} = \sqrt{32}\pi^2 v^{-2}|\rho|^2 \beta \nu_1(\nu_1 - \nu_2) F(\alpha, \ell), \tag{7.25}$$

where

$$F(\alpha, \ell) = \sqrt{\sum_{i=1}^{n} \sum_{k<i} \ell_{ik}^4 10^{-2\alpha\ell_{ik}/10}}, \tag{7.26}$$

the rms value has been used for $\sin \theta_1$, and the reflection coefficients are all approximated by an average magnitude $|\rho|$.

As an example, suppose that an interferometer is designed for observations near 100 GHz, and that it incorporates 10 antenna stations in a linear configuration at approximately equal increments in distance up to 1 km from the master oscillator. The interconnecting oscillator cable carries a reference signal at $\nu_1 = 2$ GHz, and for this cable $|\rho| = 0.1$, $\alpha = 0.06$ dB m^{-1}, $v = 2.4 \times 10^8$ ms^{-1}, and the temperature coefficient of electrical length is 10^{-5} K^{-1}. From Eq. (7.26) we find that $F(\alpha, \ell) = 1.1 \times 10^4$. For a temperature variation of 0.1 K in the cable, $\beta = 10^{-6}$. If phase errors at 100 GHz are required to be less than 1°, $\delta\phi_{\text{rms}}$ must not exceed 0.02°, and from Eq. (7.25) ν_1 and ν_2 must not differ by more than 1.6 MHz.

Automatically Correcting System

An interesting variation on the round-trip scheme, shown in Fig. 7.8, was suggested by J. Granlund (NRAO 1967). It is particularly suitable for providing a stable reference frequency at a number of points along a linear array of antennas. Frequencies ν_1 and ν_2 are generated by stable oscillators and are injected at opposite ends of the transmission line. The frequency difference $\nu_1 - \nu_2$ is again very small. At an intermediate station the two signals are extracted by directional

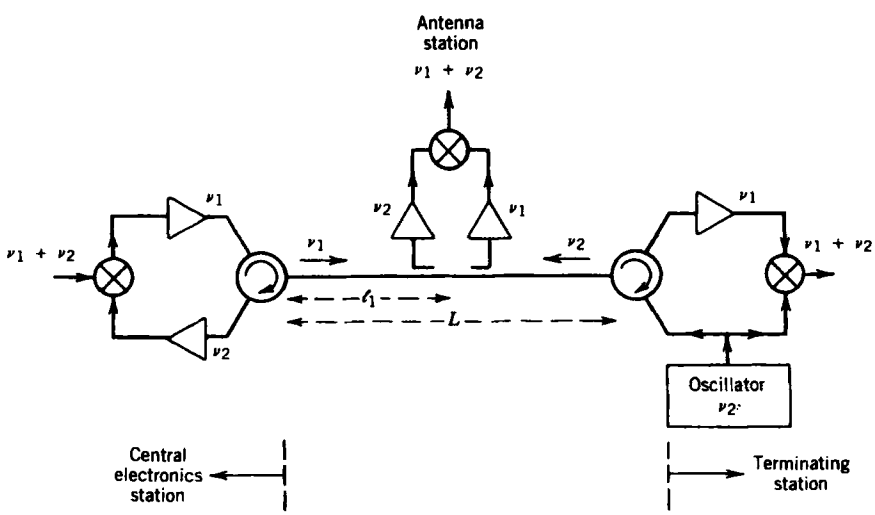


Figure 7.8 Scheme proposed by J. Granlund (NRAO 1967) for establishing a reference signal at frequency $\nu_1 + \nu_2$ at various stations along a transmission line. One such antenna station is shown.

couplers and multiplied to form the sum frequency. The phase of this sum at the antenna station in Fig. 7.8 is

$$2\pi\nu_1\ell_1 v^{-1} + 2\pi\nu_2(L-\ell_1)v^{-1} = 2\pi\nu_1 L v^{-1} - 2\pi(\nu_1-\nu_2)(L-\ell_1)v^{-1}. \tag{7.27}$$

For two points at positions ℓ_1 and ℓ_2 on the line, the difference in the sum-frequency phases is

$$\Delta\phi = 2\pi(\nu_1-\nu_2)(\ell_1-\ell_2)v^{-1}. \tag{7.28}$$

This difference would be zero if ν_1 and ν_2 were equal, but it is necessary to maintain a finite frequency difference because the directivity of the couplers alone is seldom sufficient to separate the two signals adequately. The effect of the line length variation is not measured explicitly in this case, but the correction occurs automatically, except for the small term in Eq. (7.28). Reflections in the cable can produce errors, as described for the previous scheme, and may be the limiting consideration for the frequency offset. A practical implementation of the scheme of Fig. 7.8 is described by Little (1969).

Fiberoptic Transmission of LO Signals

Optical fiber can replace cables and transmission lines in most of the LO (local oscillator) schemes discussed above. Some features of optical fiber transmission that should be taken into account are outlined below.

- Different optical wavelengths can be used in the two directions of a round-trip system to help separate the signals. At the antenna, the frequency of the laser signal from the master LO can be offset by a few tens of megahertz by using a special modulating device, and injected into the line in the return direction. Alternatively, a different laser can be used for the return signal. It is important to take into account the effects of the fiber dispersion and temperature-induced changes in the laser wavelengths, particularly in the case where two different lasers are used. However, if the laser wavelengths are chosen to be very close to the zero-dispersion wavelength of the fiber, the resulting errors can be minimized.
- As mentioned in Section 7.1, the performance of optical components such as isolators and directional couplers is much better than that of corresponding microwave components. With careful design, it is possible to use such components to separate signals at the same laser wavelength traveling in opposite directions in a fiber. Round-trip phase systems have been made in which a radio frequency signal is transmitted on an optical carrier, and at the receiving end a half-silvered mirror is used to return a component of the signal back along the fiber for a round-trip measurement. It may be necessary to use an optical isolator at the transmitting end to ensure that any of the re-

turned signal that reaches the laser is very small. Reflection of a laser signal back into the output can disturb the operation of the laser.

- In general, when a multi-fiber cable is flexed, the effective lengths of the individual fibers vary smoothly and remain matched to a much greater degree than is the case for bundled coaxial cables. As a result, it may be possible to use two separate fibers for the two different directions in a round-trip scheme, depending on the accuracy required.
- Twisting of a straight fiber that is held under constant tension has been found to cause less change in the electrical length than bending of a fiber. Twisting, however, can result in small changes in the amplitude of the transmitted signal, resulting from the residual sensitivity of the optical receiver to the angle of the linear polarization of the light.
- It is possible to stabilize the length of the path through a fiber by use of round-trip phase measurement at the optical wavelength. In practice this requires the use of an automatic correction loop in which a length adjustment device is controlled by the round-trip phase, since length variations comparable to the optical wavelength can occur on timescales of much less than one second.
- A local oscillator frequency can be transmitted as the frequency difference of two optical laser signals which travel in the same fiber. The radio frequency is generated by combining the optical signals in a photo-optic diode. Radio power of several microwatts can be obtained, which is sufficient to provide local oscillator power for an SIS mixer. This scheme is particularly attractive for receivers at millimeter and submillimeter wavelengths (Payne et al. 1998).
- For standard optical fiber, the temperature coefficient of length is approximately $7 \times 10^{-6}\ \mathrm{K}^{-1}$. High-stability fiber, developed by Sumitomo for special applications, has a temperature coefficient that is about an order of magnitude less.

Phase-Locked Loops and Reference Frequencies

Some practical points in the implementation of local oscillator systems should be briefly mentioned. In two of the schemes described above, an oscillator at the antenna is controlled by a phase-locked loop. Details of the design of phase-locked loops are given, for example, by Gardner (1979), and here we mention only the choice of the natural frequency of the loop. Unless the natural frequency is about an order of magnitude less than the frequency at the inputs of the phase detector, the loop response may be fast enough to introduce undesirable phase modulation at the phase detector frequency. In the system in Fig. 7.5, the frequency of the input signals to the phase detector is the offset frequency $\nu_1 - \nu_2$, an upper limit on which has been placed by consideration of the reflections in the line. Also, the bandwidth of the noise to which the loop responds is proportional to the natural frequency. These considerations place an upper limit on the natural frequency of the loop, which in turn limits the choice of the oscillator to be locked. An oscil-

lator with inherently poor phase stability (when unlocked) requires a loop with a higher natural frequency than does a more stable oscillator. Crystal-controlled oscillators are highly stable and require loop natural frequencies of only a few hertz. They are especially suitable for long transmission lines because the noise bandwidth of the loop is correspondingly small. With crystal-controlled oscillators at the antennas, it is possible to send out the reference frequency in bursts, rather than continuously. Signals traveling in opposite directions can then be separated by time multiplexing and no frequency offset is required. However, the change in impedance of the circuits at the ends of the cable when the direction of the signal is reversed could become a limiting factor in the accuracy of the round-trip phase measurement. Systems of this type have been designed for several large arrays (Thompson et al. 1980; Davies, Anderson, and Morison 1980).

In addition to the establishment of a phase-locked oscillator at each antenna at a reference frequency (equal to ν in Fig. 7.4, ν_2 in Fig. 7.5, and $\nu_1 + \nu_2$ in Fig. 7.8), it is necessary to generate the multiples or submultiples of this frequency that are required for frequency conversions of the received signal. In frequency multiplication, phase variations increase in proportion to the frequency. Within the multiplier chain from the frequency standard to the first LO frequency, the choice of frequency that is transmitted from the central location to the antenna is generally not critical. However, if significant noise is added in the transmission process, it may be better to transmit a high frequency to minimize multiplication of phase errors resulting from the added noise.

Minimization of phase variations in the frequency-multiplication circuit is largely a matter of reducing temperature-related effects, and in this regard the scheme depicted in Fig. 7.9 is worthy of mention. It may be useful to generate a "comb" spectrum consisting of many harmonics that can be used, for example, for tuning in discrete frequency intervals. This can be done by applying the fundamental frequency to a varactor diode, but the voltage at which the varactor goes into conduction varies with temperature, so the phase of the waveform at which it starts to conduct during each cycle varies. This causes variation in the phases of the harmonics that are generated. In the circuit in Fig. 7.9, the effect of this variation is eliminated. The input fundamental waveform at frequency ν is not applied

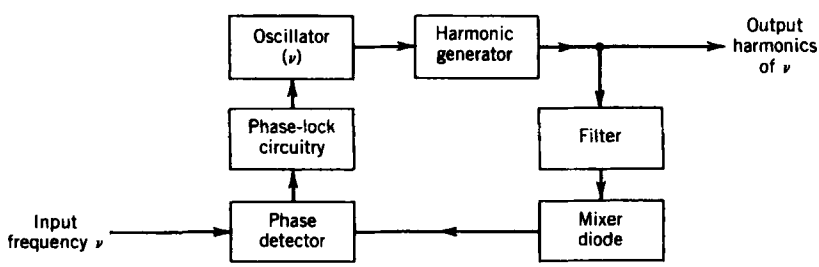


Figure 7.9 Scheme for generating a comb spectrum of harmonics of a frequency ν, in which phase changes in the harmonic generator are eliminated by enclosing it within a phase-locked loop. The filter passes two harmonics that combine in the mixer diode to generate a signal at frequency ν.

directly to the harmonic generator but is used to lock an oscillator at frequency ν. This oscillator drives the harmonic generator. The waveform at the oscillator frequency that is compared with the input frequency is taken after the varactor by selecting two adjacent harmonics and combining them in a mixer diode. The phase-locked loop holds constant the phase of this output waveform relative to the input frequency ν, and adjusts the phase of the oscillator to compensate for a change in time of switch-on of the varactor.

In the case of a connected-element array, low-frequency components of the phase noise of the master oscillator cause similar effects in the local oscillator phase at each antenna, and therefore their contributions to the relative phase of the signals at the correlator input tend to cancel. However, the frequency components of the phase noise suffer phase changes as a result of the time delay in the path of the reference signal from the master oscillator to each antenna, and also as a result of the time delay of the IF signal from the corresponding mixer to the correlator input (including the variable delay that compensates for the geometric delay). Thus the cancelation is important only for frequency components of the phase noise that are low enough that differences in these phase changes, from one antenna to another, are small. The bandwidths of phase-locked loops in the local oscillator signals can also limit the frequency range over which phase noise in the master oscillator is canceled. In practice, cancellation of phase noise from the master oscillator is likely to be effective up to a frequency in the range of some tens of hertz to a few hundred kilohertz, depending upon the parameters of the particular system.

Phase Stability of Filters

Tuned filters used for selecting local oscillator frequencies are also a source of temperature-related phase variations. The phase response ϕ of a filter changes by approximately $n\pi/2$ across the 3-dB bandwidth $\Delta\nu$, where n is the number of sections (poles). Thus, the rate of change of phase with frequency, measured at the center frequency ν_0, is

$$\left.\frac{\partial\phi}{\partial\nu}\right|_{\nu_0} = \frac{n\pi k_1}{2\Delta\nu}, \tag{7.29}$$

where k_1 is a constant of order unity that depends on the design of the filter. The center frequency varies with physical temperature T by

$$\frac{\partial\nu_0}{\partial T} = k_2\nu_0, \tag{7.30}$$

where k_2 is a constant related to the coefficients of expansion and variation of the dielectric constant of the filter. Thus the rate of variation of phase with temperature is given by

$$\frac{\partial\phi}{\partial T} = \left.\frac{\partial\phi}{\partial\nu}\right|_{\nu_0}\frac{\partial\nu_0}{\partial T} = nk_1k_2\left(\frac{\pi}{2}\right)\left(\frac{\nu_0}{\Delta\nu}\right). \tag{7.31}$$

The factor $\nu_0/\Delta\nu$ is the Q factor of the filter. The combined constant k_1k_2 can be determined empirically and is typically of order 10^{-5} K^{-1} for tubular bandpass filters with center frequencies in the range 1 MHz to 1 GHz. Thus, for example, if one allows a 1-K temperature variation for such a filter and places an upper limit of 0.1° on its contribution to the phase variation, the fractional bandwidth must not be less than $n/100$, or 5.4% for a six-pole filter. Filters of narrow fractional bandwidth should be used with caution. To pick out a particular frequency from a series of closely spaced harmonics it may be preferable to use a phase-locked oscillator rather than a filter.

Effect of Phase Errors

Rapidly varying phase errors, such as those resulting from noise in local oscillator circuits, cause a loss in signal amplitude and, hence, in sensitivity. They may also cause errors in the visibility phase, but the effect is small, since fast variations in the visibility phase are substantially reduced by the visibility averaging. To determine the loss in sensitivity, the signals from two antennas can be represented by $V_m e^{\phi_m(t)}$ and $V_n e^{\phi_n(t)}$ at the correlator inputs, where the ϕ terms are the phase errors for antennas m and n. The correlator output is

$$r = \langle V_m e^{\phi_m(t)} V_n^* e^{\phi_n(t)} \rangle, \tag{7.32}$$

where the angle brackets represent the expectation. Then if $\Delta\phi = [\phi_m(t) - \phi_n(t)]$ is the phase error, we have

$$r = V_1 V_2^* [\langle \cos \Delta\phi \rangle + j \langle \sin \Delta\phi \rangle]. \tag{7.33}$$

If the probability distribution of $\Delta\phi$ is an even function with zero mean, which is frequently the case, the time average of the sine term has an expectation of zero. Then, by using the first two terms of the series expression for a cosine, we obtain a result in terms of the rms phase error, $\Delta\phi_{\rm rms}$:

$$r \simeq \left[1 - \tfrac{1}{2}\Delta\phi_{\rm rms}^2\right]. \tag{7.34}$$

The cosine approximation is accurate to 1% for values of $\Delta\phi_{\rm rms}$ less than $\sim 37°$. A reduction in sensitivity of 1% occurs for $\Delta\phi_{\rm rms} = 8.1°$.

7.3 FREQUENCY RESPONSES OF THE SIGNAL CHANNELS

Optimum Response

The signals in a synthesis array usually pass through a number of amplifiers, filters, mixers, and transmission lines from the outputs of the antennas to the inputs of the correlators. The characteristics of these components are impressed on the signals, and therefore we should consider their effect on the sensitivity and accu-

racy of the visibility measurements. These characteristics can be specified largely in terms of the overall frequency response of the receiving channel. The important considerations are the optimum frequency response and the tolerances on the deviations of the channels of different antennas from this optimum response. The following discussion is based on an analysis by Thompson and D'Addario (1982).

We assume that the astronomical signal and the receiver noise both have flat spectra over the width of an IF band or spectral channel. Then the spectrum of the signal delivered to the correlators from a given antenna is determined by the frequency response of the associated receiving equipment. If $H(\nu) = |H(\nu)|e^{j\phi(\nu)}$ is the voltage–frequency response function, the output from the correlator for antennas m and n, resulting from cosmic signals, is proportional to

$$\begin{aligned}\tfrac{1}{2}\int_{-\infty}^{\infty} H_m(\nu)H_n^*(\nu)\,d\nu &= \mathcal{R}e\left[\int_0^{\infty} H_m(\nu)H_n^*(\nu)\,d\nu\right] \\ &= \mathcal{R}e\left[\int_0^{\infty} |H_m(\nu)||H_n(\nu)|e^{j(\phi_m-\phi_n)}\,d\nu\right],\end{aligned} \tag{7.35}$$

where we have used the relation in Eq. (A3.6), $H_m H_n^*$ being hermitian, and the subscripts denote the antennas. We are concerned here with the dependence of the signal-to-noise ratio of an observation on the frequency responses of the signal channels. In practice, the frequency responses are nonzero only within a limited frequency band of width $\Delta\nu$. From Eq. (6.42) we can define a factor $\mathcal{D}$ equal to the signal-to-noise ratio relative to that with identical rectangular responses of width $\Delta\nu$:

$$\mathcal{D} = \frac{\mathcal{R}e\left[\int_0^{\infty} H_m(\nu)H_n^*(\nu)\,d\nu\right]}{\sqrt{\Delta\nu\int_0^{\infty}|H_m(\nu)|^2|H_n(\nu)|^2\,d\nu}}. \tag{7.36}$$

This equation has a maximum value if $|H_m(\nu)|$ and $|H_n(\nu)|$ are constant across the band $\Delta\nu$, that is, if the amplitude response is a rectangular function. If, in addition, $\phi(\nu)$ is identical for both antennas, $\mathcal{D}$ is equal to unity. Thus, a rectangular passband yields the greatest sensitivity within a limited bandwidth. Note that the same integral of $H_m H_n^*$ applies to both the real and imaginary parts of a complex correlator, and hence it also applies to the modulus of the visibility.

Of the other ways in which the receiving passband modifies the response of a synthesis array, the most important is the smearing of detail in the synthesized response, which limits the field of view that can be usefully mapped. This effect has been described in Section 6.3. For a given sensitivity, a rectangular passband results in the least smearing, since it is the most compact in the frequency dimension.

An exact rectangular passband, of course, is only an ideal concept. In practice, the steepness of the sides of the passband must be determined by the particular design and the number of poles in the response. The response can be made to approximate a rectangular shape more closely as the number of poles increases, with a proportionate increase in $\partial\phi/\partial T$ as shown by Eq. (7.31). To examine the tolerable deviations of the actual passband responses, two effects must be consid-

ered: (1) the decrease in the signal-to-noise ratio and (2) the introduction of errors in determining gain factors for individual antennas, as will be described.

Tolerances on Variation of the Frequency Response: Degradation of Sensitivity

We first consider the effects on the sensitivity. Equation (7.36) provides a degradation factor $\mathcal{D}$, which is the signal-to-noise ratio with frequency responses $H_m(\nu)$ and $H_n(\nu)$, expressed as a fraction of that which would be obtained with rectangular passbands of width $\Delta(\nu)$. In constructing a receiving system, the usual goal is to keep the passband flat with steep edges, but, in practice, effects such as differential attenuation and reflections in cables introduce slopes and ripples in the frequency response that are not identical from one antenna to another. To examine these effects, $\mathcal{D}$ can be calculated for an initially rectangular passband with various distortions imposed. The distortions considered are the following:

1. Amplitude slope across the passband, with the logarithm of the amplitude varying linearly with frequency.
2. Sinusoidal amplitude ripple; this could result from a reflection in a transmission line.
3. Displacement of the center frequency of the passband.
4. Variation in phase response as a function of frequency.
5. Delay-setting error, which introduces a component of phase linear with frequency.

Expressions for the frequency response involving the above effects are given in the first column of Table 7.1. The second column of the table gives the signal-to-noise degradation factor $\mathcal{D}$, and subscripts m and n indicate parameter values for particular antennas. The expressions in Table 7.1 have been used to derive the maximum tolerable passband distortion for each of the effects, allowing a loss in sensitivity of no more than 2.5% ($\mathcal{D} = 0.975$). The resulting limits on the passband distortion are shown in Table 7.2.

Tolerances on Variation of the Frequency Response: Gain Errors

A second effect that sets limits on the deviations of the frequency responses results from errors that can be introduced in the calibration procedure. If we omit the noise terms, the output of the correlator for an antenna pair can be expressed as

$$r_{mn} = G_{mn}\mathcal{V}_{mn}, \tag{7.37}$$

where $\mathcal{V}_{mn}$ is the source-dependent complex visibility from which the intensity map can be computed, and G_{mn} is a gain factor related to the frequency responses of the signal channels. We suppose that these responses incorporate the characteristics of the antennas and electronics in such a way that G_{mn}

TABLE 7.1 Deviation of the Frequency Characteristic from an Ideal Rectangular Response, and Corresponding Expressions for $\mathcal{D}$ and G_{mn}

Frequency Response	Signal-to-Noise Degradation, $\mathcal{D}$	Antenna-Pair Gain, G_{mn}
Amplitude slope[a] $\lvert H(\nu)\rvert^2 = H_0^2 e^{\sigma(\nu-\nu_0)}\Pi\left(\frac{\nu-\nu_0}{\Delta\nu}\right)$	$\sqrt{\frac{4\left[e^{(\sigma_m+\sigma_n)\Delta\nu/2}-1\right]}{\Delta\nu(\sigma_m+\sigma_n)\left[e^{(\sigma_m+\sigma_n)\Delta\nu/2}+1\right]}}$	$\frac{2G_0}{\Delta\nu(\sigma_m+\sigma_n)}\left[e^{(\sigma_m+\sigma_n)\Delta\nu/4} - e^{-(\sigma_m+\sigma_n)\Delta\nu/4}\right]$
Sinusoidal ripple $H(\nu) = H_0\left[1+\gamma e^{j2\pi(\nu-\nu_0)\tau}\right]\Pi\left(\frac{\nu-\nu_0}{\Delta\nu}\right)$	$\left[\frac{1+2\mathrm{Re}\{\gamma_m\gamma_n^*\}+\lvert\gamma_m\gamma_n\rvert^2}{1+\lvert\gamma_m\rvert^2+\lvert\gamma_n\rvert^2+2\mathrm{Re}\{\gamma_m\gamma_n^*\}+\lvert\gamma_m\gamma_n\rvert^2}\right]^{1/2}$ (see footnote *b*)	$G_0\left[1+\frac{2}{\pi}(\gamma_m+\gamma_n^*)+\gamma_m\gamma_n^*\right]$ (see footnote *c*)
Center-frequency displacement[d] $H(\nu) = H_0\Pi\left(\frac{\nu-\delta\nu-\nu_0}{\Delta\nu}\right)\times e^{jN\pi(\nu-\delta\nu-\nu_0)/\Delta\nu}$	$\sqrt{1-\frac{\delta\nu_m-\delta\nu_n}{\Delta\nu}}$	$G_0\left[1-\frac{\delta\nu_m-\delta\nu_n}{\Delta\nu}\right]e^{jN\pi(\delta\nu_m-\delta\nu_n)/\Delta\nu}$
Phase variation $H(\nu) = H_0\Pi\left(\frac{\nu-\nu_0}{\Delta\nu}\right)e^{j\phi(\nu)}$	$1-\frac{1}{2}\langle\phi_{mn}^2\rangle$ $\phi_{mn}(\nu) = \phi_m(\nu)-\phi_n(\nu)-\langle\phi_m(\nu)-\phi_n(\nu)\rangle$ (see footnote *e*)	$G_0\left[1-\frac{1}{2}\langle\phi_{mn}^2\rangle\right]$
Delay-setting error $H(\nu) = H_0 e^{j2\pi\nu\tau}\Pi\left(\frac{\nu-\nu_0}{\Delta\nu}\right)$	$\frac{\sin[\pi\Delta\nu(\tau_m-\tau_n)]}{\pi\Delta\nu(\tau_m-\tau_n)}$	$G_0\left[\frac{\sin\pi\Delta\nu(\tau_m-\tau_n)}{\pi\Delta\nu(\tau_m-\tau_n)}\right]e^{j\pi\Delta\nu(\tau_m-\tau_n)}$ (see footnote *f*)

[a] The unit rectangle function $\Pi(x)$ is equal to 1 for $\lvert x\rvert \le \frac{1}{2}$ and zero for $\lvert x\rvert > \frac{1}{2}$. Parameters are as follows: H_0 and G_0, gain constants; σ, slope parameter; ν_0, passband center frequency; γ, relative amplitude of sinusoidal component; $\delta\nu$, frequency offset; τ, delay error.

[b] For integral value of $\Delta\nu\tau$ (integral number of cycles across passband).

[c] $\Delta\nu\tau = \frac{1}{2}$ (half cycle of sinusoidal ripple across passband).

[d] Linear phase response with difference $N\pi$ between passband edges.

[e] The brackets $\langle\ \rangle$ indicate a mean over the passband.

[f] Phase term corresponds to baseband response (center frequency $= \Delta\nu/2$).

TABLE 7.2 Examples of Frequency Response Tolerances

	Criterion	
Type of Variation	2.5% Degradation in Signal-to-Noise Ratio	1% Maximum Gain Error
Amplitude slope	3.5 dB edge-to-edge	2.7 dB edge-to-edge
Sinusoidal ripple	2.9 dB peak-to-peak	2.0 dB peak-to-peak
Center-frequency displacement	$0.05\Delta\nu$	$0.007\Delta\nu$
Phase variation	$\phi_{mn} = 12.8°$ rms	$\phi_{mn} = 9.1°$ rms
Delay-setting error	$0.12/\Delta\nu$	$0.05/\Delta\nu$

is proportional to the correlator output for antenna pair (m, n) when a point source of unit flux density at the field center is observed. In practice, the G_{mn} values may be determined from observations of calibration sources for which the visibilities are known. The measured antenna-pair gains can be used to correct the correlator output data directly, but there are advantages if, instead, they are used to determine (voltage) gain factors $g = |g|e^{i\phi}$ for the individual antennas such that

$$G_{mn} = g_m g_n^*. \tag{7.38}$$

Since, in a large array, there are many more correlated antenna pairs than antennas [up to $n_a(n_a - 1)/2$ pairs for n_a antennas], not all the calibration data need be used. This adds important flexibility to the calibration procedure; for example, a source resolved at the longest spacings of an array can be used to determine the antenna gains from measurements made only at the shorter spacings. The same principle leads to adaptive calibration described in Section 11.4.

In general, the factoring in Eq. (7.38) requires that the frequency responses be identical for all antennas, or differ only by constant multiplicative factors. If this requirement is fulfilled, we can assign gain factors

$$g = \sqrt{\int_0^\infty |H(\nu)|^2 \, d\nu}. \tag{7.39}$$

In practice, the frequency responses differ, and an approximate solution to Eq. (7.38) can be obtained by choosing the g values to minimize

$$\sum |G_{mn} - g_m g_n^*|^2, \tag{7.40}$$

where the summation is taken over all antenna pairs (m, n) for which G_{mn} can be measured by observation of a calibration source. In calibrating subsequent observations of unknown sources, $g_m g_n^*$ is used in place of G_{mn} in Eq. (7.37) for all antenna pairs, whether they are directly calibrated or not. To avoid introducing

errors with this scheme, the residuals

$$\varepsilon_{mn} = G_{mn} - g_m g_n^* \tag{7.41}$$

must be small, which requires that the frequency responses be sufficiently similar. Thus, we are concerned here with the deviations of the frequency responses from one another rather than from an ideal response.

By using model responses for groups of antennas, calculating the pair gains, the best-fit antenna gains, and the residuals, tolerances on the bandpass distortion can be assigned. Pair gains for the various distortions discussed earlier are given in the third column of Table 7.1. Table 7.2 shows examples of tolerances. The results depend to some extent on the distribution of distortions in the model responses, which for the results shown were chosen with the intention of maximizing the residuals. The criteria of 2.5% loss in sensitivity and 1% maximum gain error shown in Table 7.2 were used during the early operation of the VLA (Thompson and D'Addario 1982), and are not necessarily generally applicable. More stringent criteria may be appropriate depending on the sensitivity and dynamic range to be achieved. The acceptable level of gain error for any instrument can be determined by making calculations of the response to source models with simulated errors of various levels introduced into the model visibility data. Bagri and Thompson (1991) give a discussion of the sources and effects of gain errors in the VLA.

Delay-Setting Tolerances

Inaccuracies in adjustment of the compensating time delays in an array can result from either of two effects. There are errors in calculating the correct setting, which result from errors in calibration of antenna positions or of the delay devices. These errors can be reduced by using a calibration source that is close in position to the source being mapped. Tolerances on such errors are determined by the effects summarized in Table 7.2. There are also errors that result from the discretely adjustable nature of the delays. In analog systems, delay elements providing a binary sequence of values are switched in and out of the signal path. In digital systems the delay can be adjusted in steps governed by a train of timing pulses, as described in Section 8.5. In either case there is a minimum delay increment τ_0. If the delay for each antenna is readjusted whenever the magnitude of the error is equal to $\tau_0/2$, the probability distribution of the error is uniform from $-\tau_0/2$ to $\tau_0/2$. The rms delay error for a single antenna is then $\tau_0/(2\sqrt{3})$. For any pair of antennas in an array, the errors for two antennas can generally be assumed to vary independently. Thus the differential delay error for any pair has a probability distribution equal to the convolution of the distributions for the individual antennas. This distribution is a triangular function with maximum errors $\pm\tau_0$ and an rms value of $\tau_0/\sqrt{6}$. For single-sideband receiving systems the delay errors introduce an error in the phase of the visibility, the rms value of which is equal to $2\pi\,\nu_{\rm rms}$ times the rms delay error, where $\nu_{\rm rms}$ is the rms frequency of the IF band in which the delay is inserted.

One method of eliminating the delay-step phase error is to make τ_0 equal to the reciprocal of the mean frequency at which the delay is inserted. Adjustments of the delay then involve phase changes that are integral numbers of complete rotations of the phase. This technique requires that the IF bandwidth be small compared with the center frequency, so that τ_0 is not a large fraction of the reciprocal bandwidth. The technique has therefore been most useful in some of the earlier arrays that had narrow receiving bandwidths. For instruments with wider bandwidths the phase errors can be made tolerable by using small enough values for both τ_0 and the intermediate frequency at which the delay is introduced. If a baseband IF response is used, that is, one in which the passband is defined by a lowpass filter, $\nu_{\rm rms}$ is equal to $\Delta\nu/\sqrt{3}$. This scheme is well suited for use with a digital delay system. It is used, for example, in the VLA, for which $\tau_0 = 1/(32\Delta\nu)$, and taking the rms delay error as $\tau_0/\sqrt{6}$, the resulting phase error is $\pi/(48\sqrt{2})$ rad $= 2.65°$. From Eq. (7.34) the resulting loss in sensitivity is 0.11%.

In addition to causing a loss in sensitivity, delay-induced phase errors contribute to errors in the phase of the measured visibility. In this case it is the values after time averaging, not the instantaneous values, that are critical. The effective averaging time is of the order of the time taken for the baseline vector to cross a cell in the simple case of cell averaging discussed in Section 5.2. In a synthesis array the compensating delay for each antenna is adjusted to equalize the delay relative to some reference point as the source moves across the sky. If the antenna spacings are large, the delay may change by several increments during most cell crossings, and the resulting phase errors are reduced by the data averaging. However, for any pair of antennas, the rate of change of the geometric delay, which is proportional to u, goes through zero when the baseline vector crosses the v axis (see Section 4.3). The rate of change of the instrumental delays at that point depends on the location of the antennas relative to the chosen delay reference point, but may be small for some antennas in an array. Depending on the details of the array, it may be expected that for some (possibly small) fraction of the visibility data the phase errors will not be significantly reduced by the averaging.

Implementation of Bandpass Tolerances

The tolerances summarized in Table 7.2 apply to the overall system from the antennas to the correlator inputs. In practice, the frequency response is determined mainly by filters in the late stages, immediately preceding the correlators or digital samplers. Specifications on such filters should provide for the required matching of responses and should include consideration of the temperature effects discussed in Section 7.2 under *Phase Stability of Filters*. The frequency selectivity of elements in the earlier stages can then be held to the minimum required for rejection of interfering signals, thus minimizing the effect on the overall response. It is also possible to implement the filtering digitally after the sampling, instead of in the analog IF stages. The digital sampling is then performed on the full IF bandwidth. Digital filtering is briefly discussed in Section 8.7, and has the

advantage that the resulting passband does not depend on the tuning of individual filters and is relatively insensitive to temperature variations.

7.4 POLARIZATION MISMATCH ERRORS

The response of two antennas to an unpolarized source is greatest when the antennas are identically polarized. Small variations in the polarization characteristics of one antenna relative to another occur as a result of mechanical tolerances. These variations lead to errors in the assignment of antenna gains in a manner similar to the variations in frequency responses. To examine this effect, we calculate the response of two arbitrarily polarized antennas to a randomly polarized source, which is given by the term for the Stokes parameter I_v in Eq. (4.29). Definitions of symbols are in terms of the polarization ellipse (see Fig. 4.8 and related text). The position angle of the major axis is ψ, the axial ratio is $\tan\chi$, and subscripts m and n indicate two antennas of an array. As an example, we consider antennas with nominally identical circular polarization for which we can write $\chi_m = \pi/4 + \Delta\chi_m$ and $\chi_n = \pi/4 + \Delta\chi_n$, where the Δ terms represent the deviations of the corresponding parameter from the ideal value. The required response is

$$\begin{aligned} G_{mn} = G_0[&\cos(\psi_m - \psi_n)\cos(\Delta\chi_m - \Delta\chi_n) \\ &+ j\sin(\psi_m - \psi_n)\cos(\Delta\chi_m + \Delta\chi_n)]. \end{aligned} \tag{7.42}$$

Now $\psi_m - \psi_n$ and the Δ terms represent construction tolerances and are all small. Thus we can expand the trigonometric functions and retain only the first- and second-order terms. Equation (7.42) then becomes

$$G_{mn} = G_0\left\{1 - \tfrac{1}{2}[(\psi_m - \psi_n)^2 + (\Delta\chi_m - \Delta\chi_n)^2] + j(\psi_m - \psi_n)\right\}. \tag{7.43}$$

An analysis similar to the procedure for frequency responses in Section 7.3 can be made by assigning polarization characteristics to a model group of antennas and determining pair gains, best-fit antenna gains, and gain residuals. For simplicity, it is assumed that the spread of values is of similar magnitude for the parameters χ and ψ. A 1% maximum gain residual then results from a spread of $\pm 3.6°$ in χ and ψ. A value of $\Delta\chi = 3.6°$ corresponds to an axial ratio of 1.13 for the polarization ellipse, and it is not difficult to obtain feeds for which the deviation from circularity is within this value near the beam center. A similar analysis for linearly polarized antennas gives tolerances of the same order (Thompson 1984).

7.5 PHASE SWITCHING

Reduction of Response to Spurious Signals

The technique of phase switching for a two-element interferometer has been described in Chapter 1, where it was explained as an early method of obtaining ana-

log multiplication of signals. The principle is as indicated in Fig. 1.8. However, in later instruments the power-law detector is replaced by a correlator. Although more direct methods of signal multiplication are now used, phase switching is still useful to eliminate small offsets in correlator outputs that can result from imperfections in circuit operation or from spurious signals. The latter are difficult to eliminate entirely in any complicated receiving system, since combinations of harmonics of oscillator frequencies that fall within the observing frequency band or any intermediate frequency band may infiltrate the electronics. Such signals, at levels too low to detect by common test procedures, can be strong enough to produce unwanted components in the output. For an array of n_a antennas, a receiving bandwidth $\Delta\nu$, and an observing duration τ, signals at the limit of detectability are at a power level of order $(n_a\sqrt{\Delta\nu\tau})^{-1}$ relative to the noise; for example, 75 dB below the noise for $n_a = 27$, $\Delta\nu = 50$ MHz, and $\tau = 8$ h. Similar effects can be produced by cross coupling of small amounts of noise from one IF system to another.

Since spurious signals produce components of the visibility that change only slowly with time, they show up as spurious detail near the origin of the map. If they enter the signal channel at a point that comes after the phase switch, so that they produce a component with no switch-frequency variation at the synchronous detector, they can generally be reduced by several orders of magnitude by phase switching.

Implementation of Phase Switching

Consider the problem of phase switching a multielement array in which the products of the signals from all possible pairs of antennas are formed. Phase switching can be represented by multiplication of the received signals by periodic functions that alternate in time between values of +1 and −1. For the mth and nth antennas let these functions be $f_m(t)$ and $f_n(t)$. Synchronous detection of the correlator output for these two antennas requires a reference waveform $f_m(t)f_n(t)$, and any nonvarying, unswitched components from the multiplier are reduced by a factor

$$\frac{1}{\tau}\int_0^\tau f_m(t)f_n(t)\,dt, \tag{7.44}$$

after averaging for a time τ. For the periodic waveforms we are concerned with, this factor will be zero if τ is a multiple of the minimum period of orthogonality $\tau_{\rm or}$ for $f_m(t)$ and $f_n(t)$. In fact, the unwanted output components may not be exactly constant, because the tracking of the compensating delays introduces slow changes in the phases with which the spurious signals are combined. However, the unwanted outputs will be strongly reduced by the synchronous detection as long as their variation is small over the period $\tau_{\rm or}$. If the orthogonality of the phase-switching functions depends on the relative timing of transitions, the timing should be adjusted so that the functions are orthogonal at the correlator inputs. Thus, it may be necessary to adjust the timing of the switching waveforms at the antennas to compensate for the varying instrumental delay inserted as a source moves across the sky.

Implementation of phase switching on an array of n_a antennas calls for n_a mutually orthogonal, two-state waveforms. Sets of square waves whose frequencies are proportional to integral powers of two (Rademacher functions) are orthogonal with τ_{or} equal to the period of the lowest nonzero frequency. In phase switching, τ_{or} is equal to the data averaging time, which is typically a few seconds but for special cases may be as low as 10 ms. The shortest interval between switching transitions τ_{sw} is equal to the half period of the fastest square wave. Technically, it is convenient if τ_{or}/τ_{sw} does not greatly exceed about two orders of magnitude. If one antenna remains unswitched, then $\tau_{or}/\tau_{sw} = 2^{n_a-1}$. Square waves of the *same* frequency are orthogonal if their phases differ by a quarter of a cycle in time. When this condition for orthogonality is also included, $\tau_{or}/\tau_{sw} = 2^{n+1}$, where n is the smallest integer greater than or equal to $(n_a - 3)/2$. This reduces the value of τ_{or}/τ_{sw}, but the orthogonality then depends on the relative timing of the transitions at the correlator, which is not the case for square waves of different frequencies. In either case τ_{or}/τ_{sw} is inconveniently large for a large array and, for example, for $n_a = 27$ it is of order 10^8 in the first case and 10^4 in the second.

It is useful to note that a condition for a pair of square waves of different frequency to be orthogonal, for arbitrary time shifts, is that they do not contain Fourier components of the same frequency. A property of square waves is that all even numbered Fourier components (i.e. even harmonics of the fundamental frequency) have zero coefficients, but odd numbered components have nonzero coefficients. Thus, although sinusoids with frequencies proportional to 1, 2, 3, . . . are mutually orthogonal, square waves with such frequencies, in general, are not. For example, square waves of frequencies 1, 2, and 4 have no common Fourier components, and are mutually orthogonal, but 1, 3, and 5 have common components and are not mutually orthogonal. D'Addario (2001) shows by generalization of this analysis that the lowest frequency sets of N mutually orthogonal square waves consist of those with frequencies proportional to 2^n for $n = 0, 1, \ldots, (N-1)$, that is, the Rademacher sets discussed above. Since the different square waves of a Rademacher set contain no common Fourier components, their orthogonality is not affected by the relative time shifts. Note, also, that strict orthogonality is not essential for phase switching. Unwanted responses can be reduced by a factor of 10^{-4} or less by using square waves with k cycles per averaging period for values of k that are prime numbers greater than 100.

The beneficial effects of phase switching can also be obtained by sinusoidal phase modulation of the signals, that is, by introducing a set of orthogonal sinusoids as frequency offsets at the antennas. The wanted outputs then appear at the correlator output shifted in frequency from the response to unwanted components that do not suffer the frequency offsets. Unless fringe rotation is performed in the correlator, removing the frequency shift from the wanted components is more complicated than the equivalent operation in a phase switching system. In the case of VLBI, the frequency offsets at the antennas that result naturally from the sidereal motion of a source are generally sufficiently large that phase switching is not necessary.

An alternative set of two-state orthogonal functions that can be used to implement phase switching are the Walsh functions, which are rectangular waveforms in which the time interval between transitions between +1 and −1 is a varying but integral multiple of a basic interval, as in Fig. 7.10. For a description of Walsh functions (Walsh

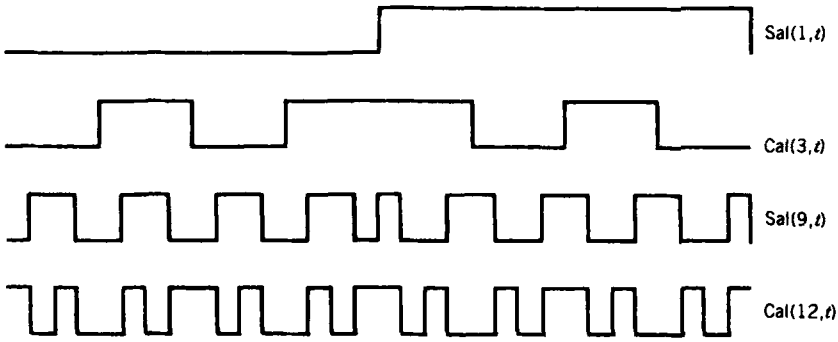


Figure 7.10 Four examples of Walsh functions, each of which repeats after the one cycle of the time base interval plotted above. Within this interval the sal functions are odd, and the cal functions are even. The value of each function alternates between 1 and −1. The first number in parentheses in the name of each function is the sequency, which is equal to half the number of zero crossings in the time base interval. Time t is measured as a fraction of the time base.

1923) see, for example, Harmuth (1969, 1972) or Beauchamp (1975). Various systems of designating and ordering Walsh functions are in use. In one system (Harmuth 1972) those with even symmetry are designated as cal(k, t) and those with odd symmetry as sal(k, t). Here t is time expressed as a fraction of the time base T, which is the interval at which the waveform repeats, and k is the *sequency*, which is equal to half the number of zero crossings within the time base. Walsh functions with different sequencies are orthogonal, and cal and sal functions of the same sequency are othogonal but differ only by a time offset. The orthogonality requires that the time bases of the individual Walsh functions be aligned in time, so time offsets are not permitted. Walsh functions with sequencies that are integral powers of two are square waves. If one antenna is unswitched, and if only the cal or only the sal functions are used, the highest sequency is $(n_a - 1)$. Then $\tau_{or}/\tau_{sw} = 2n$, where n is the smallest power-of-two integer greater than or equal to $(n_a - 1)$. If both cal and sal functions are used, then n is the smallest power-of-two integer greater than or equal to $(n_a - 1)/2$. For example, for $n_a = 64$, τ_{or}/τ_{sw} is 128 in the first case and 64 in the second.

Another designation for Walsh functions, wal(n, t), includes both cal and sal functions, cal (n, t) = wal $(2n, t)$ and sal (n, t) = wal $(2n - 1, t)$. One method of generating Walsh functions makes use of Hadamard matrices, of which the one of lowest order is

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{7.45}$$

Higher-order matrices can be obtained by replacing each element of H_2 by the matrix H_2 multiplied by the element replaced [which is equivalent to forming an outer product: see Eq. (4.51)]. If this is performed twice, for example, we obtain

$$H_8 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \begin{matrix} \text{cal}(0,t), & \text{pal}(0,t) \\ \text{sal}(4,t), & \text{pal}(4,t) \\ \text{sal}(2,t), & \text{pal}(2,t) \\ \text{cal}(2,t), & \text{pal}(6,t) \\ \text{sal}(1,t), & \text{pal}(1,t) \\ \text{cal}(3,t), & \text{pal}(5,t) \\ \text{cal}(1,t), & \text{pal}(3,t) \\ \text{sal}(3,t), & \text{pal}(7,t). \end{matrix} \tag{7.46}$$

The rows of the matrices correspond to the Walsh functions indicated, the signs being reversed for odd sequencies in this particular generation process. The waveform required at the phase detector is the product of the phase-switching functions at the two antennas involved. The product of two such Walsh functions is a Walsh function, the sequency of which is greater than, or equal to, the difference between the sequencies of the two original functions.

Walsh functions can also be generated as products of Rademacher functions. Rademacher functions are designated $R(n, t)$, where n is an integer and the half period of the square wave is $T/2^n$; that is, there are 2^{n-1} complete cycles within the time base, T. The function $\text{R}(0, t)$ has a constant value of unity. In the examples in Fig. 7.10, $\text{sal}(1, t)$ is a Rademacher function, and $\text{cal}(3, t)$ and $\text{sal}(9, t)$ are each products of $\text{sal}(1, t)$ and one other Rademacher function. When considering Walsh functions as products of Rademacher functions it is convenient to use the Paley designation, $\text{pal}(n, t)$ (Paley 1932). The integer n is called the *natural order* of the Walsh function. A Walsh function $\text{pal}(n, t)$, which is the product of Rademacher functions $\text{R}(i, t)$, $\text{R}(j, t), \ldots, \text{R}(m, t)$, has a natural-order number $n = 2^{i-1} + 2^{j-1} +, \ldots, +2^{m-1}$. The product of two Walsh functions is another Walsh function, of which the natural-order number is given by modulo-2 addition (that is, no-carry addition) of the binary natural-order numbers of the component Walsh functions.

Table 7.3 shows the relationship between the natural-order numbers for a series of Walsh functions, and the Rademacher functions of which they are composed. The product of two Walsh functions can be expressed as the product of the component Rademacher functions, for example,

$$\begin{aligned} \text{pal}(7,t) \times \text{pal}(10,t) &= [\text{R}(1,t) \times \text{R}(2,t) \times \text{R}(3,t)] \times [\text{R}(2,t) \times \text{R}(4,t)] \\ &= \text{R}(1,t) \times \text{R}(2,t) \times \text{R}(2,t) \times \text{R}(3,t) \times \text{R}(4,t) \\ &= \text{R}(1,t) \times \text{R}(3,t) \times \text{R}(4,t) \\ &= \text{pal}(13,t), \end{aligned} \tag{7.47}$$

where we have used the fact that the product of a Walsh or Rademacher function with itself is equal to unity. The natural orders of the two Walsh functions, 7 and 10, in binary form are 0111 and 1010. The modulo-2 addition of these binary numbers is 1101, which is equal to 13, the natural order of the Walsh function product.

TABLE 7.3 Rademacher Components of Some Walsh Functions

Natural Order Designation	Rademacher Components					Sequency Designation
	$R(0,t)$	$R(1,t)$	$R(2,t)$	$R(3,t)$	$R(4,t)$	
pal(0, t)	1					cal(0, t)
pal(1, t)		1				sal(1, t)
pal(2, t)			1			sal(2, t)
pal(3, t)		1	1			cal(1, t)
pal(4, t)				1		sal(4, t)
pal(5, t)		1		1		cal(3, t)
pal(6, t)			1	1		cal(2, t)
pal(7, t)		1	1	1		sal(3, t)
pal(8, t)					1	sal(8, t)
pal(9, t)		1			1	cal(7, t)
pal(10, t)			1		1	cal(6, t)
pal(11, t)		1	1		1	sal(7, t)
pal(12, t)				1	1	cal(4, t)
pal(13, t)		1		1	1	sal(5, t)
pal(14, t)			1	1	1	sal(6, t)
pal(15, t)		1	1	1	1	cal(5, t)

The examination of Walsh functions as products of Rademacher functions leads to a useful insight into the efficiency of Walsh function phase switching in eliminating unwanted components (Emerson 1983). Let $\mathcal{U}(t)$ be an unwanted response within the receiving system, for example, resulting from crosstalk in IF signals or from an error in the sampling level of a digitizer. $\mathcal{U}(t)$ arises after the initial phase switching, so when synchronous detection with the phase-switching waveform is performed at a later stage, $\mathcal{U}(t)$ becomes $\mathcal{U}(t)\mathrm{pal}(n,t)$, and this product is significantly reduced in the subsequent averaging. Suppose that $\mathrm{pal}(n,t)$ is the product of m Rademacher functions, $\mathrm{R}(i,t), \mathrm{R}(j,t), \ldots, \mathrm{R}(\ell,t)$. We can consider multiplying $\mathcal{U}(t)$ by $\mathrm{pal}(n,t)$ as equivalent to multiplying by each of the Rademacher components in turn. Also, we assume that the period of the Rademacher functions is small compared with the timescale of variations of $\mathcal{U}(t)$. Then, after the first multiplication and averaging, the mean residual spurious voltage is

$$\mathcal{U}_1(t) = \frac{\left[\mathcal{U}(t) + \delta t \frac{d\mathcal{U}}{dt}\right] - \mathcal{U}(t)}{2} = \frac{\delta t}{2}\frac{d\mathcal{U}}{dt} = \frac{T}{2^{i+1}}\frac{d\mathcal{U}}{dt}, \tag{7.48}$$

where δt is equal to the half period of the Rademacher function, $T/2^i$. $\mathcal{U}_1$ is calculated for one cycle of $\mathrm{R}(i,t)$, but within the assumption that $\mathcal{U}(t)$ is slowly varying, $\mathcal{U}_1$ can be taken as equal to the average over the Walsh time base T. Multiplication by the second Rademacher function is obtained by replacing $\mathcal{U}$ in Eq. (7.48) by $\mathcal{U}_1$, which yields

$$\mathcal{U}_2(t) = \frac{T}{2^{j+1}} \frac{d\mathcal{U}_1}{dt} = \frac{T^2}{2^{i+j+2}} \frac{d^2\mathcal{U}}{dt^2}. \tag{7.49}$$

For the m Rademacher components, we obtain

$$\mathcal{U}_m(t) = \frac{T^m}{2^{(i+j+\cdots+\ell+m)}} \frac{d^m\mathcal{U}}{dt^m}, \tag{7.50}$$

so only the higher derivatives of $\mathcal{U}$ remain.

Walsh functions pal(n, t) for which n is an integral power of two are the least effective in eliminating unwanted responses, since they are each just a single Rademacher function. As shown by examination of Table 7.3, those for which $n = 2^k - 1$, where k is an integer, contain the largest number of Rademacher components. In arrays with a small number of antennas, for which a large number of different switching functions is not required, it is possible to select Walsh functions that are the most effective in reducing unwanted components. Similarly, Walsh functions can be more effective than square waves in some applications to single antennas, such as beam switching between a source and a reference position on the sky (Emerson 1983). An early application of Walsh functions to transmission lines is discussed by Fowle (1904).

Another set of possible phase-switching functions are m-sequences, considered by Keto (2000) for cases where both 90° and 180° phase changes are required.

Interaction of Phase Switching with Fringe Rotation and Delay Adjustment

The effectiveness of phase switching in reducing the response to spurious signals depends on the point in the signal channel at which these unwanted signals are introduced. The three following cases illustrate the most important possibilities.

1. The unwanted signal enters the antennas or some point in the signal channels that is ahead of the phase switching, the fringe rotation, and the compensating delays. The unwanted signal then suffers phase switching like the wanted signals and is not suppressed in the synchronous detection (although it may be reduced by the fringe rotation if the fringe frequency is high, as in the case of VLBI). Externally generated interference behaves in this manner, and its effect is discussed in Chapter 15.
2. The unwanted signal enters after the phase switching but before the fringe rotation and delay compensation. The fringe-rotation phase shifts, designed to reduce to zero the fringe frequencies of the desired signals at the correlator output, act on the spurious signal and cause it to appear at the correlator output as a component at the natural fringe frequency for a point source at the phase reference position. This component then undergoes synchronous detection with a Walsh function. If the natural fringe frequency transiently matches the frequency of a Fourier component of this Walsh function, a spurious response can occur.

3. The spurious signal enters after the phase switching and the fringe rotation but before the delay compensation. The signal then suffers phase shifts resulting from the changing of the compensating delay. The resulting component at the correlator output has a frequency equal to the natural fringe frequency that would occur if the observing frequency were equal to the intermediate frequency at which the compensating delays are introduced. Thus the oscillations are one to three orders of magnitude lower in frequency than the natural fringe frequency, and it is consequently easier to avoid coincidence with the frequency of a component of the Walsh function.

From these considerations, it is usually advantageous to perform both the phase switching and the fringe rotation as early in the signal channel as possible. Figure 7.11 shows, as an example, the phase-switching scheme used in the VLA from

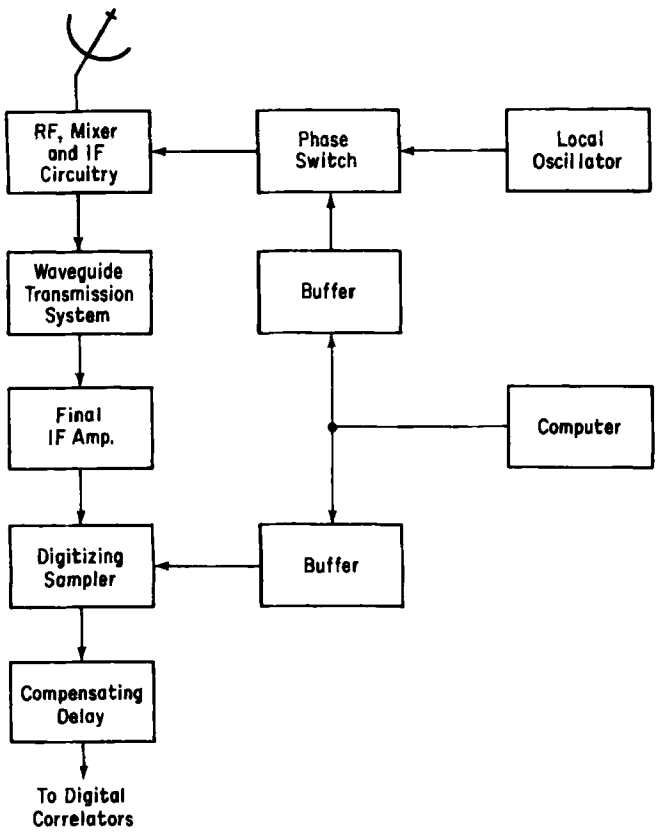


Figure 7.11 Simplified schematic diagram of the receiving channel for one antenna of the VLA. Walsh functions generated by the computer are periodically fed to digital buffers, from which they are clocked out to the phase switch and to sign-reversal circuitry at the sampler. From Granlund, Thompson, and Clark (1978); ©1978 IEEE.

a description by Granlund, Thompson, and Clark (1978). The phase switching at the antenna is performed on a local oscillator, rather than on the full signal band, so that a broadband phase switch is not needed. The signals are digitized at the output of the final IF amplifier and thereafter are delayed and multiplied digitally. Since digital circuits do not suffer from unwanted drifts and offsets as analog circuits may do, there is no need to include them between the phase switching and the synchronous detection. Thus, the latter can be performed by reversing the sign bit in the digitized signal data, and need be applied only to n_a signal channels rather than $n_a(n_a - 1)/2$ correlator outputs.

7.6 AUTOMATIC LEVEL CONTROL AND GAIN CALIBRATION

In most synthesis arrays automatic level control (ALC) circuits are used to hold constant the level of the total signal, that is, the cosmic signal plus the system noise, at certain critical points. A fraction of the total signal level is detected, and the resulting voltage is compared with a preset value, to generate a control signal that is fed back to some variable-gain element of the signal chain. Points at which the signal level is critical include modulators for transmission of IF signals on optical or microwave carriers and inputs to analog correlators or digital samplers. For a discussion of level tolerances in samplers, see Section 8.4.

The effect of an ALC loop is to hold constant the quantity $|g|^2(T_S + T_A)\Delta\nu$, where g is the voltage gain from the antenna output to the point of control, T_S is the system temperature, and T_A is the component of antenna temperature due to the source under observation. Thus $|g|^2$ is made to vary inversely as $(T_S + T_A)$, which can change substantially with the antenna pointing angle as a result of ground radiation in the sidelobes and atmospheric attenuation. To measure such gain changes, a signal from a broadband noise source can be injected at the input of the receiving electronics. This noise source is switched on and off, usually at a frequency of a few hertz to a few hundred hertz, and the resulting component is sampled and monitored using a synchronous detector. When the noise source is on, it adds a calibrating component T_C to the overall system temperature, which should not be more than a few percent of T_S to avoid degradation of sensitivity. The amplitude of the switched component is a direct measure of the system gain, and for $T_S \gg T_A$ the ratio of the signal levels with the noise source on and off is equal to $1 + T_C/T_S$, which provides a continuous measure of T_S. This scheme does not measure changes in antenna gain resulting from mechanical deformation, which must be calibrated separately by periodic observation of a radio source.

APPENDIX 7.1 SIDEBAND-SEPARATING MIXER

The principle of the sideband-separating mixer, or image-rejection mixer, is shown in Fig. A7.1. The terms $\cos(2\pi\nu_u t)$ and $\cos(2\pi\nu_\ell t)$ represent frequency components of the input waveform at the upper- and lower-sideband frequencies, respectively. The input is applied to two mixers, for which the local oscillator

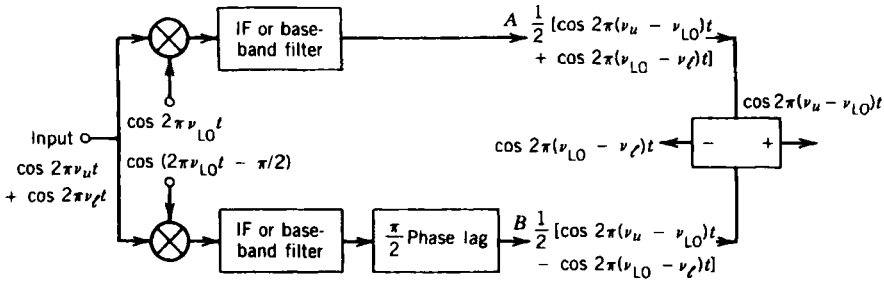


Figure A7.1 Schematic illustration of the principle of the sideband-separating (image-rejection) mixer. The upper-sideband response is obtained from the sum of the outputs A and B, and the lower-sideband response from the difference of these outputs.

waveforms at frequency ν_{LO} are in phase quadrature. The mixers generate products of the signal and local oscillator waveforms, and the filters pass only the terms of frequency equal to the difference of ν_{LO} and ν_u or ν_ℓ. The output from the lower mixer also passes through a $\pi/2$ phase lag network. From the resulting terms at points A and B one can see that by applying the waveforms at these points to a summing network, the upper-sideband response is obtained. Similarly, by using a differencing network, the lower-sideband response is obtained. In either case the accuracy of the suppression of the response to the unwanted sideband depends on the accuracy of the quadrature phase relationships, the matching of the frequency responses of the mixers and filters, and the insertion loss of the phase lag network. In practice, for conversion from a few gigahertz to baseband, suppression to a level of −20 dB is routinely achievable. With careful design, suppression to a level approaching −30 dB can be obtained (Archer, Granlund, and Mauzy 1981). For conversion from millimeter wavelengths, suppression of the unwanted sideband may be less effective.

APPENDIX 7.2 DISPERSION IN OPTICAL FIBER

For a frequency component, ν_m, of a signal modulated onto an optical carrier, $A\sin(2\pi\nu_{opt} + \phi)$, and transmitted down a fiber, the resulting signal at the output of the fiber can be represented by

$$\begin{aligned} &A^2[1 + m\cos(2\pi\nu_m t)]\sin^2(2\pi\nu_{opt}t + \phi) \\ &\quad = A^2\sin^2(2\pi\nu_{opt}t + \phi) \\ &\qquad + \frac{A^2 m}{2}\sin[2\pi(\nu_{opt} + \nu_m)(t - \Delta t) + \phi]\sin(2\pi\nu_{opt}t + \phi) \\ &\qquad + \frac{A^2 m}{2}\sin[2\pi(\nu_{opt} - \nu_m)(t + \Delta t) + \phi]\sin(2\pi\nu_{opt}t + \phi), \end{aligned} \tag{A7.1}$$

where m is the modulation index. This equation resembles the usual representation for amplitude modulation in communications, except that here the carrier *power* varies linearly with the modulation. Thus, on the left-hand side, the square of the carrier expression is used. For the terms of frequency $\nu_{\text{opt}} \pm \nu_m$, the time has been offset by $\pm\Delta t$ to represent the effects of the variation of propagation velocity with frequency. Δt can take both positive and negative values depending on the sign of the dispersion $\mathcal{D}$ shown in Fig. 7.3. Each term in Eq. (A7.1) is proportional to optical power and, thus, also to the modulation amplitude. By applying the identity for the product of two sines to each term on the right-hand side of Eq. (A7.1), and ignoring DC and optical frequency terms, we obtain for the amplitude at the output of the optical receiver,

$$\frac{A^2 m}{4}\left\{\cos[2\pi\nu_m(t+\Delta t) - 2\pi\nu_{\text{opt}}\Delta t] + \cos[2\pi\nu_m(t-\Delta t) - 2\pi\nu_{\text{opt}}\Delta t]\right\}$$
$$= \frac{A^2 m}{2}\left\{\cos[2\pi(\nu_m t - \nu_{\text{opt}}\Delta t)]\cos(2\pi\nu_m\Delta t)\right\}. \tag{A7.2}$$

The free-space wavelength corresponding to frequency ν_{opt} is λ_{opt}, and the wavelength difference between frequencies ν_{opt} and $\nu_{\text{opt}} + \nu_m$ is $\lambda_{\text{opt}}^2\nu_m/c$. If $\mathcal{D}$ is the dispersion and ℓ is the length of the fiber, $\Delta t = \mathcal{D}\ell\lambda_{\text{opt}}^2\nu_m/c$, and $\nu_{\text{opt}}\Delta t = \mathcal{D}\ell\lambda_{\text{opt}}\nu_m$. Thus the recovered modulation can be written as

$$\frac{A^2 m}{2}\left\{\cos[2\pi\nu_m(t - \mathcal{D}\ell\lambda_{\text{opt}})]\cos(2\pi\nu_m^2\mathcal{D}\ell\lambda_{\text{opt}}^2/c)\right\}. \tag{A7.3}$$

The phase change induced by Δt at the carrier frequency ν_{opt} appears in the phase of the modulation frequency in the first cosine function in Eq. (A7.2). At frequency ν_m this phase term is equivalent to a time delay $\mathcal{D}\ell\lambda_{\text{opt}}$, as seen in Eq. (A7.3). This delay is much larger than Δt, and represents the difference between the phase and group the velocities in the fiber. The second cosine modifies the amplitude of the modulation component ν_m. For example, with dispersion $\mathcal{D} = 2\ \text{ps}(\text{nm}\cdot\text{km})^{-1}$ (note that this is equal to $2\times10^{-6}\ \text{s m}^{-2}$), $\ell = 50$ km, $\lambda_{\text{opt}} = 1550$ nm, and $\nu_m = 10$ GHz, we obtain $\Delta t = 8$ ps, $\mathcal{D}\ell\lambda_{\text{opt}} = 155$ ns, and the response at frequency ν_m is reduced by 1.1 dB relative to the low-frequency end of the modulation spectrum. Note that we have assumed above that the frequency spread of the laser results entirely from the modulation spectrum, which is justifiable for a high-quality laser with an external modulator. Modulation of a diode laser by varying the voltage across it can result in unwanted frequency modulation, further spreading the optical spectrum.

REFERENCES

Agrawal, G. P. *Fiber-Optic Communication Systems*, Wiley, New York, 1992.

Allen, L. R. and R. H. Frater, Wideband Multiplier Correlator, *Proc. IEEE*, **117**, 1603–1608, 1970.

Archer, J. W., E. M. Caloccia, and R. Serna, An Evaluation of the Performance of the VLA Circular Waveguide System, *IEEE Trans. Microwave Theory Tech.*, **MTT-28**, 786–791, 1980.

Archer, J. W., J. Granlund, and R. E. Mauzy, A Broadband UHF Mixer Exhibiting High Image Rejection over a Multidecade Baseband Frequency Range, *IEEE J. Solid-State Circuits*, **SC-16**, 385–392, 1981.

Baars, J. W. M., J. F. van der Brugge, J. L. Casse, J. P. Hamaker, L. H. Sondaar, J. J. Visser, and K. J. Wellington, The Synthesis Radio Telescope at Westerbork, *Proc. IEEE*, **61**, 1258–1266, 1973.

Bagri, D. S. and A. R. Thompson, Hardware Considerations for High Dynamic Range Imaging, *Radio Interferometry: Theory, Techniques and Applications*, T. J. Cornwell and R. A. Perley, Eds., Pub. Astron. Soc. Pacific Conf. Ser., **19**, 47–54, 1991.

Batty, M. J., D. L. Jauncey, P. T. Rayner, and S. Gulkis, Tidbinbilla Two-Element Interferometer, *Astron. J.*, **87**, 938–944, 1982.

Beauchamp, K. G., *Walsh Functions and Their Applications*, Academic Press, London, 1975.

Borella, M. S., J. P. Jue, D. Banergee, B. Ramamurthy, and B. Mukherjee, Optical Components for WDM Lightwave Networks, *Proc. IEEE*, **85**, 1274–1307, 1997.

Bracewell, R. N., R. S. Colvin, L. R. D'Addario, C. J. Grebenkemper, K. M. Price, and A. R. Thompson, The Stanford Five-Element Radio Telescope, *Proc. IEEE*, **61**, 1249–1257, 1973.

Callen, H. B., and Welton, T. A., Irreversibility and Generalized Noise, *Phys. Rev.*, **83**, 34–40, 1951.

Casse, J. L., E. E. M. Woestenburg, and J. J. Visser, Multifrequency Cryogenically Cooled Front-End Receivers for the Westerbork Synthesis Radio Telescope, *IEEE Trans. Microwave Theory Tech.*, **MTT-30**, 201–209, 1982.

D'Addario, L. R., *Orthogonal Functions for Phase Switching and a Correction to ALMA Memo. 287*, ALMA Memo. 385, National Radio Astronomy Observatory, Socorro, NM, 2001.

Davies, J. G., B. Anderson, and I. Morison, The Jodrell Bank Radio-Linked Interferometer Network, *Nature*, **288**, 64–66, 1980.

Elsmore, B., S. Kenderdine, and M. Ryle, Operation of the Cambridge One-Mile Diameter Radio Telescope, *Mon. Not. R. Astron. Soc.*, **134**, 87–95, 1966.

Emerson, D. T., *The Optimum Choice of Walsh Functions to Minimize Drift and Cross-Talk*, Working Report 127 IRAM, Grenoble, July 18, 1983.

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

Fowle, F. F., The Transposition of Electrical Conductors, *Trans. Am. Inst. Elect. Eng.*, **23**, 659–689, 1904.

Gardner, F. M., *Phaselock Techniques*, 2nd ed., Wiley, New York, (1979).

Granlund, J., A. R. Thompson, and B. G. Clark, An Application of Walsh Functions in Radio Astronomy Instrumentation, *IEEE Trans. Electromagn. Compat.*, **EMC-20**, 451–453, 1978.

Harmuth, H. F., Applications of Walsh Functions in Communications, *IEEE Spectrum*, **6**, No. 11, 82–91, 1969.

Harmuth, H. F., *Transmission of Information by Orthogonal Functions*, 2nd ed., Springer-Verlag, Berlin, 1972.

Kerr, A. R., Suggestions for Revised Defintions of Noise Quantities, Including Quantum Effects, *IEEE Trans. MTT*, 1999.

Kerr, A. R., M. J. Feldman, and S.-K. Pan, Receiver Noise Temperature, the Quantum Noise Limit, and the Role of Zero-Point Fluctuations, *Proc. 8th Int. Symp. Space Terahertz Technology*, March 25–27, pp. 101–111, 1997; also *MMA Memo. 161*, NRAO, Charlottesville, VA, 1997.

Keto, E., Three-Phase Switching with m-Sequences for Sideband Separation in Radio Interferometry, *Pub. Aston. Soc. Pacific*, **112**, 711–715, 2000.

Kraus, J. D., *Radio Astronomy*, McGraw-Hill, New York, 1966, pp. 261–263; 2nd ed., Cygnus-Quasar Books, Powell, OH, 1986.

Little, A. G., A Phase-Measuring Scheme for a Large Radiotelescope, *IEEE Trans. Antennas Propag.*, **AP-17**, 547–550, 1969.

Napier, P. J., D. S. Bagri, B. G. Clark, A. E. E. Rogers, J. D. Romney, A. R. Thompson, and R. C. Walker, The Very Long Baseline Array, *Proc. IEEE*, **82**, 658–672, 1994.

Napier, P. J., A. R. Thompson, and R. D. Ekers, The Very Large Array: Design and Performance of a Modern Synthesis Radio Telescope, *Proc. IEEE*, **71**, 1295–1320, 1983.

NRAO, *A Proposal for a Very Large Array Radio Telescope*, Vol. II, National Radio Astronomy Observatory, Green Bank, WV, 1967, Ch. 14.

Nyquist, H., Thermal Agitation of Electric Charge in Conductors, *Phys. Rev.*, **32**, 110–113, 1928.

Padin, S., A Wideband Analog Continuum Correlator for Radio Astronomy, *IEEE Trans. Instrum. Meas.*, **IM-43**, 782–784, 1994.

Paley, R. E., A Remarlable Set of Orthogonal Functions, *Proc. London Math. Soc.*, **34**, 241–279, 1932.

Payne, J. M., Millimeter and Submillimeter Wavelength Radio Astronomy, *Proc. IEEE*, **77**, 993–1071, 1989.

Payne, J. M., L. D'Addario, D. T. Emerson, A. R. Kerr, and B. Shillue, Photonic Local Oscillator for the Millimeter Array, in *Advanced Technology MMW, Radio, and Terahertz Telescopes*, T. G. Phillips, Ed., Proc. SPIE, **3357**, 143–151, 1998.

Payne, J. M., J. W. Lamb, J. G. Cochran, and N. J. Bailey, A New Generation of SIS Recievers for Millimeter-Wave Radio Astronomy, *Proc. IEEE*, **82**, 811–823, 1994.

Phillips, T. G. Millimeter and Submillimeterwave Receivers, in *Astronomy with Millimeter and Submillimeter Wave Interferometry*, M. Ishiguro and W. J. Welch, Eds., Astron. Soc. Pacific Conf. Ser., **59**, 68–77, 1994.

Phillips, T. G. and D. P. Woody, Millimeter- and Submillimeter-Wave Receivers, *Ann. Rev. Astron. Astrophys.*, **20**, 285–321, 1982.

Pospieszalski, M. W., W. J. Lakatosh, E. Wollack, Loi D. Nguyen, M. Le, M. Lui, T. Liu, Milimeter-Wave Waveguide-Bandwidth Cryogenically-Coolable InP HEMPT Amplifiers, *1997 IEEE MTT-S International Microwave Symposium Digest*, Denver, CO, 1285–1288, 1997.

Read, R. B., Two-Element Interferometer for Accurate Position Determinations at 960 Mc, *IRE Trans. Antennas Propag.*, **AP-9**, 31–35, 1961.

Reid, M. S., R. C. Clauss, D. A. Bathker, and C. T. Stelzried, Low Noise Microwave Receiving Systems in a Worldwide Network of Large Antennas, *Proc. IEEE*, **61**, 1330–1335, 1973.

Sinclair, M. W., G. R. Graves, R. G. Gough, and G. G. Moorey, The Receiver System, *Proc. IREE Aust.*, **12**, 147–160, 1992.

Swarup, G. and K. S. Yang, Phase Adjustment of Large Antennas, *IRE Trans. Antennas Propag.*, **AP-9**, 75–81, 1961.

Thompson, A. R., *Tolerances on Polarization Mismatch*, VLB Array Memo. 346, National Radio Astron. Obs. Charlottesville, VA, March 1984.

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

Thompson, A. R., and L. R. D'Addario, Frequency Response of a Synthesis Array: Performance Limitations and Design Tolerances, *Radio Sci.*, **17**, 357–369, 1982.

Thompson, M. C., L. E. Wood, D. Smith, and W. B. Grant, Phase Stabilization of Widely Separated Oscillators, *IEEE Trans. Antennas Propag.*, **AP-16**, 683–688, 1968.

Tiuri, M. E. and A. V. Räisänen, Radio-Telescope Receivers, in *Radio Astronomy*, 2nd ed., J. D. Kraus, Cygnus-Quasar Books, Powell, OH, 1986, Ch. 7.

Tucker, J. R. Quantum Limited Detection in Tunnel Junction Mixers, *IEEE J. Quantum Elect.*, **QE-15**, 1234–1258, 1979.

Tucker, J. R. and Feldman, M. J., Quantum Detection at Millimeter Wavelengths, *Rev. Mod. Phys.*, **57**, 1055–1113, 1985.

USAF, *Handbook of Geophysics and Space Environments*, S. L. Valley, Ed., U.S. Air Force Cambridge Research Laboratories, Bedford, MA, 1965, pp. 3-20–3-22.

Walsh, J. L., A Closed Set of Orthogonal Functions, *Ann. J. Math*, **55**, 5–24, 1923.

Webber, J. C., A. R. Kerr, S.-K. Pan, and M. W. Pospieszalski, Receivers for the Millimeter Array, in *Advanced Technology MMW, Radio, and Terahertz Telescopes*, T. G. Phillips, Ed., Proc. SPIE, **3357**, 122–131, 1998.

Weinreb, S., M. Balister, S. Maas, and P. J. Napier, Multiband Low-Noise Receivers for a Very Large Array, *IEEE Trans. Microwave Theory Tech.*, **MTT-25**, 243–248, 1977a.

Weinreb, S., D. L. Fenstermacher, and R. W. Harris, Ultra-Low-Noise 1.2- to 1.7-GHz Cooled GaSaFET Amplifiers, *IEEE Trans. Microwave Theory Tech.*, **MTT-30**, 849–853, 1982.

Weinreb, S., R. Predmore, M. Ogai, and A. Parrish, Waveguide System for a Very Large Antenna Array, *Microwave J.*, **20**, March, 49–52, 1977b.

Welch, W. J., J. R. Forster, J. Dreher, W. Hoffman, D. D. Thornton, and M. C. H. Wright, An Interferometer for Millimeter Wavelengths, *Astron. Astrophys.*, **59**, 379–385, 1977.

Welch, W. J. and 36 coauthors, The Berkeley-Illinois-Maryland-Association Millimeter Array, *Pub. Astron. Soc. Pacific*, **108**, 93–103, 1996.

Wengler, M. J. and Woody, D. P., Quantum Noise in Heterodyne Detection, *IEEE J. Quantum Electron.*, **QE-23**, 613–622, 1987.

Wright, M. C. H., B. G. Clark, C. H. Moore, and J. Coe, Hydrogen-Line Aperture Synthesis at the National Radio Astronomy Observatory: Techniques and Data Reduction, *Radio Sci.*, **8**, 763–773, 1973.

Young, A. C., M. G. McCulloch, S. T. Ables, M. J. Anderson, and T. M. Percival, The Local Oscillator System, *Proc. IREE Aust.*, **12**, 161–172, 1992.

Zorin, A. B., Quantum Noise in SIS Mixers, *IEEE Trans. Magn.*, **MAG-21**, 939–942, 1985.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

8 Digital Signal Processing

The use of digital rather than analog instrumentation offers important practical advantages in the implementation of compensating time delays and the measurement of cross-correlation of signals. In digital delay circuits the accuracy of the delay depends on the accuracy of the timing pulses in the system, and long delays accurate to tens or hundreds of picoseconds are more easily achieved digitally than by using analog delay lines. Furthermore, there is no distortion of the signal by the digital units other than the calculable effects of quantization. On the other hand, with an analog system it is difficult to keep the shape of the frequency response within tolerances as delay elements are switched into and out of the signal channels. Correlators* with wide dynamic range are readily implemented digitally, including those with multichannel output, as required for spectral line observations. Analog implementation of multichannel correlators requires filter banks to divide the signal passband into many narrow channels. Such filters, when subject to temperature variations, can be a source of phase instability. Finally, except at the highest bit rates (frequencies), digital circuits require less adjustment than analog ones and are better suited to replication in large numbers for large arrays.

Digitization of the signal waveforms requires sampling of the voltages at periodic intervals and quantizing the sampled values so that each can be represented by a finite number of bits. The number of bits per sample is usually small, and may be as low as one. As a result, the digital data rate does not become unmanageably high. However, the coarse quantization that is necessary results in a loss in sensitivity, since modification of the signal levels to the quantized values effectively results in the addition of a component of "quantization noise." In most cases this loss is outweighed by the other advantages. In designing digital correlators there are compromises to be made between sensitivity and complexity, and the number of quantization levels to use is an important consideration in this chapter. Digital processing of radio astronomical signals was first used in the construction of *autocorrelators*, which measure the autocorrelation of a signal from a single antenna as a function of time offset. Then by Fourier transformation the power spectrum is obtained, for use in studies of spectral lines and other applications. The first such system in radio astronomy was constructed by Weinreb (1963). Another early digital autocorrelator was used by Goldstein (1962) to detect radar echoes from Venus.

*For explanation of the usage of the term correlator, see Section 3.3 under *Correlator*.

8.1 BIVARIATE GAUSSIAN PROBABILITY DISTRIBUTION

Before proceeding further it is appropriate to introduce the bivariate Gaussian probability function [see, e.g., Abramowitz and Stegun (1964), p. 936], which is central to what follows. If x and y are joint Gaussian random variables with zero mean and variance σ^2, the probability that one variable is between x and $x + dx$ and, simultaneously, the other is between y and $y + dy$ is $p(x, y)\,dx\,dy$, where

$$p(x, y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left[\frac{-(x^2 + y^2 - 2\rho xy)}{2\sigma^2(1-\rho^2)}\right]. \tag{8.1}$$

The form of this function is shown in Fig. 8.1. Here ρ is the cross-correlation coefficient equal to $\langle xy\rangle/\sqrt{\langle x^2\rangle\langle y^2\rangle}$, where $\langle\ \rangle$ denotes the expectation, which, with the usual assumption of ergodicity, is approximated by the average over many samples. Note that $-1 \le \rho \le 1$. For $\rho \ll 1$, the exponential can be expanded, giving

$$p(x, y) \simeq \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right)\right]\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-y^2}{2\sigma^2}\right)\right]\left(1 + \frac{\rho xy}{\sigma^2}\right). \tag{8.2}$$

For $\rho = 0$, the expression is simply the product of two Gaussian functions. Equation (8.1) can also be written

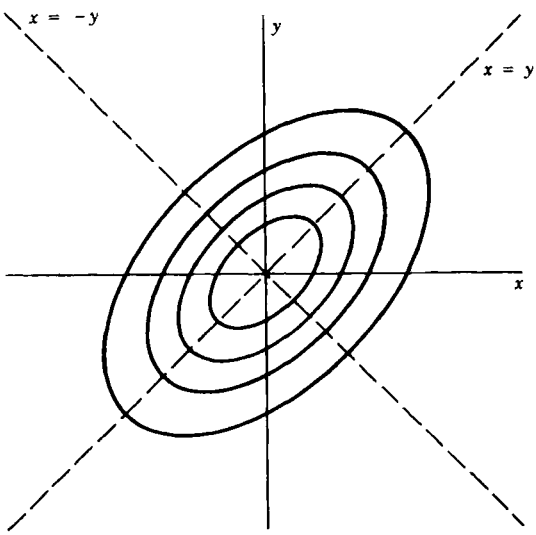


Figure 8.1 Contours of equal probability density from the bivariate Gaussian distribution in Eq. (8.1). The contours are given by $x^2 + y^2 - 2\rho xy = \text{const}$. For $\rho = 0$ they become circles, for $\rho = 1$ they merge into the line $x = y$, and for $\rho = -1$ they merge into $x = -y$.

$$p(x, y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \frac{1}{\sigma\sqrt{2\pi(1-\rho^2)}} \exp\left[\frac{-(y-\rho x)^2}{2\sigma^2(1-\rho^2)}\right]. \tag{8.3}$$

If this expression is integrated with respect to y from $-\infty$ to $+\infty$, it reduces to a Gaussian function in x. As ρ approaches unity, Eq. (8.3) becomes the product of a Gaussian in x and a Gaussian in $y - x$; the latter has a standard deviation $\sigma\sqrt{1-\rho^2}$, which approaches zero. Equations (8.1) and (8.2) will be used in examining the response of various types of samplers and correlators. For autocorrelators used with single antennas the quantity to be measured is the autocorrelation function $\langle v(t)v(t-\tau)\rangle$, where v is the received signal. This case can be treated with $x = v(t)$ and $y = v(t-\tau)$.

8.2 PERIODIC SAMPLING

We first consider the process of sampling, but without quantization, in which case the full accuracy of the signal amplitude is retained.

Nyquist Rate

If the signal is bandlimited, that is, its power spectrum is nonzero only within a finite band of frequencies, no information is lost in the sampling process as long as the sampling rate is high enough. This follows from the sampling theorem discussed in Section 5.2. Here we sample a function of time and must avoid aliasing in the frequency domain. For a baseband (lowpass) rectangular spectrum with an upper cutoff frequency $\Delta\nu$, the width of the frequency spectrum, including negative frequencies, is $2\Delta\nu$. The function is fully specified by samples spaced in time with an interval no greater than $1/(2\Delta\nu)$, that is, a sampling frequency of $2\Delta\nu$ or greater. This critical sampling frequency, $2\Delta\nu$, is called the *Nyquist rate*† (or Nyquist frequency) for the waveform. For further discussion see, for example, Bracewell (2000) or Oppenheim and Schafer (1975). In many digital systems in radio astronomy the final IF waveform has a baseband spectrum and is sampled at the Nyquist rate. For a rectangular passband of this type, the autocorrelation function, which by the Wiener–Khinchin relation is the Fourier transform of the power spectrum, is

$$R_\infty(\tau) = \frac{\sin(2\pi\Delta\nu\tau)}{2\pi\Delta\nu\tau}, \tag{8.4}$$

where the subscript ∞ indicates unquantized sampling (that is, the accuracy is not limited by a finite number of quantization levels). Nyquist sampling can also be applied to bandpass spectra, and if the spectrum is nonzero only within a range of $n\Delta\nu$ to $(n+1)\Delta\nu$, where n is an integer, the Nyquist rate is again $2\Delta\nu$. Thus,

†Shannon (1949) cites several references relevant to the development of this result, of which the earliest is Nyquist (1928).

for sampling at the Nyquist rate, the lower and upper bounds of the spectral band must be integral multiples of the bandwidth. The autocorrelation function of a signal that has a flat spectrum over such a band is

$$R_\infty(\tau) = \frac{\sin(\pi \Delta\nu\tau)}{\pi \Delta\nu\tau} \cos\left[2\pi \left(n + \tfrac{1}{2}\right) \Delta\nu\tau\right]. \tag{8.5}$$

Zeros in this function occur at time intervals τ that are integral multiples of $1/(2\Delta\nu)$. Therefore, for a rectangular passband, successive samples at the Nyquist rate are uncorrelated. Sampling at frequencies greater or less than the Nyquist rate is referred to as oversampling or undersampling, respectively.

For any signal, adjusting the center frequency so that the spectrum conforms to the bandpass sampling requirement described above minimizes the sampling rate required to avoid aliasing. If the spectrum does not conform, it is necessary to define a wider frequency band that both conforms to the sampling requirement and encompasses the required signal band. Sampling can then be performed at the Nyquist rate appropriate for the wider band.

Correlation of Sampled but Unquantized Waveforms

We now investigate the response of a hypothetical correlator for which the input signals are sampled at the Nyquist rate, but are not quantized. It is necessary to consider only single-multiplier correlators since complex correlators can be implemented as combinations of them, as indicated in Fig. 6.3. The system under discussion can be visualized as one in which the samples either remain as analog voltages, or are encoded with a sufficiently large number of bits that quantization errors are negligible. Since no information is lost in sampling or quantization, the signal-to-noise ratio of the correlation measurement may be expected to be the same as would be obtained by applying the waveforms without sampling to an analog correlator. There is probably no reason, in practice, to build a correlator for inputs with unquantized sampling. However, by comparing the results with those for quantized sampling, which we discuss later, the effects of quantization are more easily understood.

Two bandlimited waveforms, $x(t)$ and $y(t)$, are sampled at the Nyquist rate, and for each pair of samples the multiplier within the correlator produces an output proportional to the product of the input amplitudes. The integrator allows the output to be averaged for any required time interval. Now the (normalized) cross-correlation coefficient of $x(t)$ and $y(t)$ for zero time delay between the two waveforms is

$$\rho = \frac{\langle x(t)y(t)\rangle}{\sqrt{\langle [x(t)]^2\rangle \langle [y(t)]^2\rangle}}. \tag{8.6}$$

(The cross-correlation coefficient ρ should not be confused with the autocorrelation function of x or y, R_∞.) Since x and y have equal variance σ^2,

$$\langle x(t)y(t)\rangle = \rho\sigma^2. \tag{8.7}$$

The left-hand side is the averaged product of the two waveforms and thus represents the correlator output. The output of the digital correlator after N_N samples is

$$r_\infty = N_N^{-1} \sum_{i=1}^{N_N} x_i\, y_i, \tag{8.8}$$

where the subscript N denotes the Nyquist rate. Since the samples x_i and y_i obey the same Gaussian statistics as the continuous waveforms $x(t)$ and $y(t)$, we can clearly write

$$\langle r_\infty \rangle = \rho\sigma^2. \tag{8.9}$$

Thus, the output of the correlator is a linear measure of the correlation ρ. The variance of the correlator output is

$$\sigma_\infty^2 = \langle r_\infty^2 \rangle - \langle r_\infty \rangle^2, \tag{8.10}$$

and

$$\begin{aligned} \langle r_\infty^2 \rangle &= N_N^{-2} \sum_{i=1}^{N_N} \sum_{k=1}^{N_N} \langle x_i\, y_i\, x_k\, y_k \rangle \\ &= N_N^{-2} \sum_{i=1}^{N_N} \langle x_i\, y_i \rangle^2 + N_N^{-2} \sum_{i=1}^{N_N} \sum_{k \neq i} \langle x_i\, y_i\, x_k\, y_k \rangle, \end{aligned} \tag{8.11}$$

where we have separated the terms for which $i = k$ and $i \neq k$. The first summation on the right-hand side of Eq. (8.11) has a value of $\sigma^4(1 + 2\rho^2)N_N^{-1}$: from Eq. (8.3) it can be shown that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^2 y^2 p(x, y)\, dx\, dy = \sigma^4(1 + 2\rho^2). \tag{8.12}$$

The second summation term in Eq. (8.11) is readily evaluated by using the fourth-order moment relation in Eq. (6.36). Because successive samples of each signal are uncorrelated (a rectangular passband is assumed), $\langle x_i\, y_i\, x_k\, y_k \rangle = \langle x_i\, y_i \rangle \langle x_k y_k \rangle$, and the second summation term has a value of $(1 - N_N^{-1})\rho^2\sigma^4$. Returning to Eq. (8.10), we can write

$$\begin{aligned} \sigma_\infty^2 &= (1 + 2\rho^2)\sigma^4 N_N^{-1} + \left(1 - N_N^{-1}\right)\rho^2\sigma^4 - \rho^2\sigma^4 \\ &= \sigma^4 N_N^{-1}(1 + \rho^2). \end{aligned} \tag{8.13}$$

The signal-to-noise ratio with unquantized sampling is

$$\mathcal{R}_{\mathrm{sn}\infty} = \frac{\langle r_\infty \rangle}{\sigma_\infty} = \frac{\rho\sqrt{N_N}}{\sqrt{(1+\rho^2)}} \simeq \rho\sqrt{N_N}, \tag{8.14}$$

where the approximation applies for $\rho \ll 1$. Note that the condition $\rho \ll 1$ is satisfactory for most practical purposes. [The signal-to-noise ratio at the correlator output, which we are calculating here, is of interest mainly for weak signals. For a measurement period τ, $N_N = 2\Delta\nu\tau$, which is commonly 10^6–10^{12}. From Eq. (8.14) the threshold of detectability of a signal is given by $\rho\sqrt{N_N} \simeq 1$, that is, $\rho \simeq 10^{-3}$–10^{-6}.] In terms of the signal bandwidth and measurement duration, $\mathcal{R}_{\mathrm{sn}\infty} = \rho\sqrt{2\Delta\nu\tau}$. Now for observations of a point source with identical antennas and receivers, ρ is equal to the ratio of the resulting antenna temperature to the system temperature, T_A/T_S. Thus the present result is identical to that given by Eq. (6.45) for an analog correlator with continuous unsampled inputs and $T_A \ll T_S$.

Before leaving the subject of unquantized sampling, we should consider the effect of sampling at rates other than the Nyquist rate. Successive sample values from any one signal are then no longer independent. We consider a sampling frequency that is β times the Nyquist rate, and a number of samples $N = \beta N_N$. The sample interval is $\tau_s = (2\beta\Delta\nu)^{-1}$. Samples spaced by $q\tau_s$, where q is an integer, have a correlation coefficient which, from Eq. (8.4), is equal to

$$R_\infty(q\tau_s) = \frac{\beta\sin(\pi q/\beta)}{\pi q} \tag{8.15}$$

for a rectangular baseband response. Since the samples are not independent, we must reconsider the evaluation of the second summation term on the right-hand side of Eq. (8.11). For those terms for which $q = |i-k|$ is small enough that $R_\infty(q\tau_s)$ is significant, there will be additional contribution given by

$$\left[\sigma^2 R_\infty(q\tau_s)\right]^2. \tag{8.16}$$

Now R_∞^2 is very small for all but a very small fraction of the $N(N-1)$ terms in the second summation in Eq. (8.11). From Eq. (8.15), R_∞^2, at its maxima, is equal to $(\beta/\pi q)^2$ and for $q = 10^3$ is of order 10^{-6}. However, as shown above, N is likely to be as high as 10^6–10^{12}. Thus, in the second summation in Eq. (8.11) the contribution made by the terms for which the i and k samples are effectively independent remains essentially unchanged. The products for which R_∞^2 is significant make an additional contribution equal to

$$2\sigma^4 N^{-2}\sum_{q=1}^{N-1}(N-q)R_\infty^2(q\tau_s) \simeq 2\sigma^4 N^{-1}\sum_{q=1}^{\infty} R_\infty^2(q\tau_s). \tag{8.17}$$

The variance of the correlator output now becomes

$$\sigma_\infty^2 = \sigma^4 N^{-1}\left[1 + 2\sum_{q=1}^{\infty} R_\infty^2(q\tau_s)\right], \tag{8.18}$$

and the signal-to-noise ratio of the correlation measurement is

$$\mathcal{R}_{\text{sn}\infty} = \frac{\rho\sqrt{\beta N_N}}{\sqrt{1 + 2\sum_{q=1}^{\infty} R_\infty^2(q\tau_s)}}. \tag{8.19}$$

Compare this result with Eq. (8.14) for Nyquist sampling. For values of β of $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, and so on, which correspond to undersampling, $R_\infty = 0$ and the denominator in Eq. (8.19) is unity. The sensitivity thus drops as one would expect from the decrease in the data. For oversampling, $\beta > 1$ and the summation of $R_\infty^2(q\tau_s)$ in Eq. (8.19) is shown in Appendix 8.1 to be equal to $(\beta - 1)/2$. The denominator in Eq. (8.19) is then equal to $\sqrt{\beta}$, so the sensitivity is the same as that for sampling at the Nyquist rate. This is as expected, since in Nyquist sampling no information is lost. The result is different for quantized sampling, as will be explained in the following sections.

8.3 SAMPLING WITH QUANTIZATION

In some sampling schemes the signal is first quantized and then sampled, and in others it is sampled and then quantized. Ideally, the end result is the same in either case, and in analyzing the process we can choose the order that is most convenient.

Suppose that a bandlimited signal is first quantized and then sampled. Quantization generates new frequency components in the signal waveform, so it is no longer bandlimited. If it is sampled at the Nyquist rate corresponding to the unquantized waveform, as is the usual practice, some information will be lost, and the sensitivity will be less than for unquantized sampling. Also, because quantization is a nonlinear operation, we cannot assume that the measured correlation of the quantized waveforms will be a linear function of ρ, which is what we want to measure. Thus, to utilize digital signal processing there are three main points that should be investigated: (1) the relation between ρ and the measured correlation, (2) the loss in sensitivity, and (3) the extent to which oversampling can restore the lost sensitivity. Investigations of these points can be found in the work of Weinreb (1963), Cole (1968), Burns and Yao (1969), Cooper (1970), Hagen and Farley (1973), Bowers and Klingler (1974), and Jenet and Anderson (1998). The discussions given here follow most closely the approach of Hagen and Farley.

Note that in discussing sampling with quantization it is common practice to refer to Nyquist sampling when what is meant is sampling at the Nyquist rate for the *unquantized* waveform. We also follow this usage.

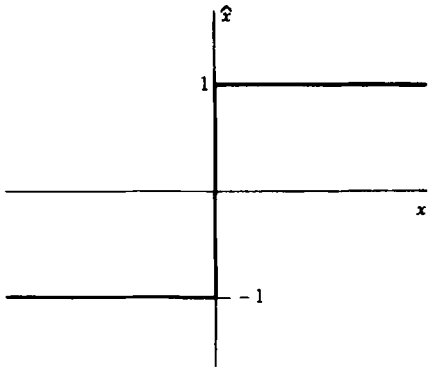


Figure 8.2 Characteristic curve for two-level quantization. The abscissa is the input voltage x and the ordinate is the quantized output $\hat{x}$.

Two-Level Quantization

The quantization characteristic for two-level (one-bit) sampling is shown in Fig. 8.2. The quantizing action senses only the sign of the instantaneous signal voltage. In many samplers the signal voltage is first amplified and strongly clipped. The zero crossings are more sharply defined in the resulting waveform, and errors that might occur if the sampling time coincides with a sign reversal are thereby minimized.

The correlator for two-level signals consists of a multiplying circuit followed by a counter that averages the products of the input samples. The input signals are assigned values of $+1$ or -1 to indicate positive or negative signal voltages, and the products at the multiplier output thus take values of $+1$ or -1 for identical or different input values, respectively. We consider sampling both at the Nyquist rate and at multiples of it, and represent by N the number of samples fed to the correlator. The average two-level correlation coefficient is

$$\rho_2 = \frac{(N_{11} + N_{\bar{1}\bar{1}}) - (N_{\bar{1}1} + N_{1\bar{1}})}{N}, \tag{8.20}$$

where N_{11} is the average number of products for which both samples have the value $+1$, $N_{1\bar{1}}$ is the average number of products in which the x sample has the value $+1$ and the y sample -1, and so on. The denominator in Eq. (8.20) is equal to the output that would occur if all pairs of input samples were identical. ρ_2 can be related to the correlation coefficient ρ of the unquantized signals through the bivariate probability distribution Eq. (8.1), from which

$$N_{11} = NP_{11} = \frac{N}{2\pi\sigma^2\sqrt{1-\rho^2}} \int_0^\infty \int_0^\infty \exp\left[\frac{-(x^2 + y^2 - 2\rho xy)}{2\sigma^2(1-\rho^2)}\right] dx\, dy, \tag{8.21}$$

where P_{11} is the probability of the two unquantized signals being simultaneously greater than zero. The other required probabilities are obtained by changing the limits of the integrals in Eq. (8.21) as follows: $\int_{-\infty}^{0}\int_{-\infty}^{0}$ for $P_{\bar{1}\bar{1}}$; $\int_{-\infty}^{0}\int_{0}^{\infty}$ for $P_{\bar{1}1}$; and $\int_{0}^{\infty}\int_{-\infty}^{0}$ for $P_{1\bar{1}}$. Note that $P_{11} = P_{\bar{1}\bar{1}}$ and $P_{1\bar{1}} = P_{\bar{1}1}$. Thus,

$$\rho_2 = 2(P_{11} - P_{1\bar{1}}). \tag{8.22}$$

The integral in Eq. (8.21) is evaluated in Appendix 8.2, from which we obtain

$$P_{11} = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho. \tag{8.23}$$

Similarly,

$$P_{1\bar{1}} = \frac{1}{4} - \frac{1}{2\pi}\sin^{-1}\rho, \tag{8.24}$$

so

$$\rho_2 = \frac{2}{\pi}\sin^{-1}\rho. \tag{8.25}$$

Equation (8.25), known as the Van Vleck relationship,[‡] allows ρ to be obtained from the measured correlation ρ_2. For small values, ρ is proportional to ρ_2.

To determine the signal-to-noise ratio of the correlation measurement, we now calculate σ_2^2, the variance of the correlator output r_2:

$$\sigma_2^2 = \langle r_2^2\rangle - \langle r_2\rangle^2, \tag{8.26}$$

where

$$r_2 = N^{-1}\sum_{i=1}^{N}\hat{x}_i\hat{y}_i. \tag{8.27}$$

Note that in this chapter the circumflex (ˆ) is used to denote quantized signal waveforms. Since $\rho_2 = \langle\hat{x}\hat{y}\rangle$, then from Eq. (8.27) $\langle r_2\rangle = \rho_2$. Note that r_2 is thus an unbiased estimator of ρ_2. The expression for $\langle r_2^2\rangle$ is equivalent to Eq. (8.11) for unquantized waveforms:

$$\langle r_2^2\rangle = N^{-2}\sum_{i=1}^{N}\langle\hat{x}_i^2\hat{y}_i^2\rangle + N^{-2}\sum_{i=1}^{N}\sum_{k\neq i}\langle\hat{x}_i\hat{y}_i\hat{x}_k\hat{y}_k\rangle. \tag{8.28}$$

[‡]This result was first derived by J. H. Van Vleck during World War II, when studying the power spectrum of strongly clipped noise which was used for electromagnetic jamming. The work was later declassified and was published by Van Vleck and Middleton (1966).

The first summation term on the right-hand side of Eq. (8.28) is equal to N^{-1} since the products $\hat{x}_i \hat{y}_i$ take values of ± 1 for two-level sampling. In evaluating the second summation term, the situation is similar to that for unquantized sampling. The factor σ^4 in Eq. (8.17) is here replaced by the square of the variance of the quantized waveform, which is unity for two-level quantization. For all except a small fraction of the terms, $q = |i - k|$ is large enough that samples i and k from the same waveform are uncorrelated. These terms make a total contribution closely equal to ρ_2^2. Those terms for which samples i and k are correlated make an additional contribution closely equal to

$$2N^{-1} \sum_{q=1}^{\infty} R_2^2(q\tau_s), \tag{8.29}$$

where $R_2(\tau)$ is the autocorrelation coefficient for a signal after two-level quantization. Thus,

$$\begin{aligned} \sigma_2^2 &= N^{-1} + (1 - N^{-1})\rho_2^2 + 2N^{-1} \sum_{q=1}^{\infty} R_2^2(q\tau_s) - \rho_2^2 \\ &\simeq N^{-1} \left[1 + 2 \sum_{q=1}^{\infty} R_2^2(q\tau_s) \right], \end{aligned} \tag{8.30}$$

where we have assumed that $\rho_2 \ll 1$ and therefore neglected the term $-N^{-1}\rho^2$. The signal-to-noise ratio is

$$\mathcal{R}_{\text{sn}2} = \frac{\langle r_2 \rangle}{\sigma_2} = \frac{2\rho\sqrt{N}}{\pi \sqrt{1 + 2 \sum_{q=1}^{\infty} R_2^2(q\tau_s)}}. \tag{8.31}$$

This ratio, relative to that for unquantized sampling at the Nyquist rate given by Eq. (8.14), defines an efficiency factor for the quantized correlation process:

$$\eta_2 = \frac{\mathcal{R}_{\text{sn}2}}{\mathcal{R}_{\text{sn}\infty}} = \frac{2\sqrt{\beta}}{\pi \sqrt{1 + 2 \sum_{q=1}^{\infty} R_2^2(q\tau_s)}}. \tag{8.32}$$

Here we have used $N = \beta N_N$, so we are considering the same observing time as in the Nyquist-sampled case, but sampling β times as rapidly.

Equation (8.25) gives the relationship between the correlation coefficients for a pair of signals before and after two-level quantization. This result includes the case of autocorrelation where the two signals differ only because of a delay. Thus,

we may write

$$R_2(q\tau_s) = \frac{2}{\pi}\sin^{-1}[R_\infty(q\tau_s)]. \tag{8.33}$$

Equation (8.15) gives $R_\infty(q\tau_s)$ for a rectangular baseband signal spectrum sampled at β times the Nyquist rate, and Eq. (8.33) becomes

$$R_2(q\tau_s) = \frac{2}{\pi}\sin^{-1}\left[\frac{\beta\sin(\pi q/\beta)}{\pi q}\right]. \tag{8.34}$$

$R_2(q\tau_s)$ thus has zeros at the same values of $q\tau_s$ that $R_\infty(q\tau_s)$ does (the principal value is taken for the inverse sine function), and for $\beta = 1, \frac{1}{2}, \frac{1}{3}$, and so on, we obtain

$$\sum_{q=1}^{\infty} R_2^2(q\tau_s) = 0. \tag{8.35}$$

In these cases the signal-to-noise ratio is $2/\pi (= 0.637)$ times that for unquantized sampling at the same rate. For oversampling with $\beta = 2$ and $\beta = 3$, the signal-to-noise ratios from Eqs. (8.32) and (8.34) are 0.744 and 0.773, respectively. Similar results have been obtained by Burns and Yao (1969) and others. Note that in the calculations given above there is an implicit dependence on the bandpass shape of the signal through the assumption that $\rho_2 \ll 1$ for samples for which i is not equal to k in Eq. (8.28). For $\beta \geq 2$, a further dependence on the bandpass shape enters through the autocorrelation function $R_2(q\tau_s)$.

It has been mentioned that quantization generates additional spectral components. We can now compare the power spectra of a signal before and after quantization since these spectra are the Fourier transforms of autocorrelation functions that are related by Eq. (8.25). Figure 8.3 shows the spectrum, after two-level quantization, of noise with an originally rectangular spectrum. A fraction of the original bandlimited spectrum is converted into a broad, low-level skirt that dies away very slowly with frequency.

Four-Level Quantization

The use of two digital bits to represent the amplitude of each sample results in less degradation of the signal-to-noise ratio than is obtained with one-bit quantization. Consideration of two-bit sampling leads naturally to four-level quantization, the performance of which has been investigated by several authors, notably Cooper (1970) and Hagen and Farley (1973). The quantization characteristic is shown in Fig. 8.4, where the quantization thresholds are $-v_0$, 0, and v_0. The four quantization states have designated values $-n$, -1, $+1$, and $+n$, where n, which is not necessarily an integer, can be chosen to optimize the performance. Products of two samples can take the value ± 1, $\pm n$, or $\pm n^2$. The four-level correlation coef-

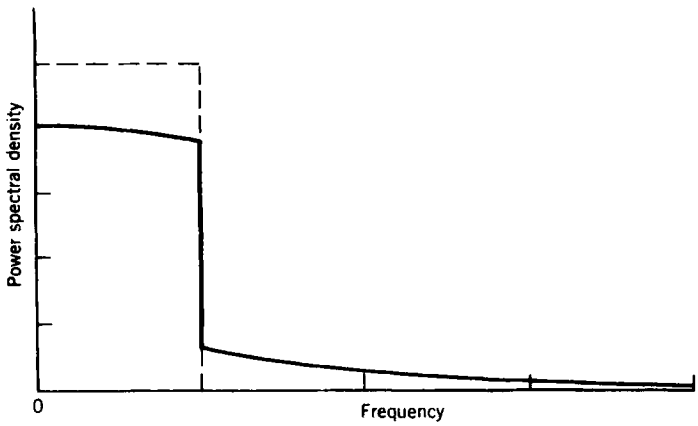


Figure 8.3 Spectra of rectangular bandpass noise before and after two-level quantization. The unquantized spectrum is of lowpass form, as shown by the broken line. The spectrum after quantization is shown by the solid curve. The power levels of the two waveforms (represented by the areas under the curves) are equal, and the Fourier transforms of their spectra are related by Eq. (8.25).

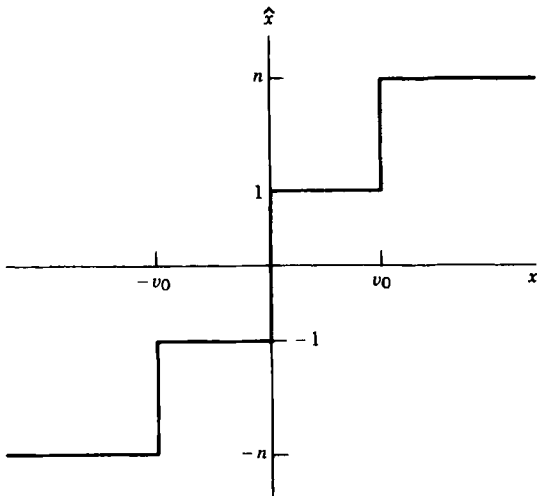


Figure 8.4 Characteristic curve for four-level quantization. The abscissa is the unquantized voltage x and the ordinate is the quantized output $\hat{x}$. v_0 is the threshold voltage.

ficient ρ_4 can be specified by an expression similar to Eq. (8.20) for the two-level case, that is,

$$\rho_4 = \frac{2n^2 N_{nn} - 2n^2 N_{n\bar{n}} + 4nN_{1n} - 4nN_{1\bar{n}} + 2N_{11} - 2N_{1\bar{1}}}{(2n^2 N_{nn} + 2N_{11})_{\rho=1}}, \tag{8.36}$$

where a bar on the subscript indicates a negative sign. The numerator is proportional to the correlator output, and reduces to the form in the denominator for $\rho = 1$, that is, when the two input waveforms are identical. The numbers of the various level combinations can be derived from the corresponding joint probabilities. Thus, for example,

$$\begin{aligned} N_{nn} &= N P_{nn} \\ &= \frac{N}{2\pi\sigma^2\sqrt{1-\rho^2}} \int_{v_0}^{\infty} \int_{v_0}^{\infty} \exp\left[\frac{-(x^2 + y^2 - 2\rho xy)}{2\sigma^2(1-\rho^2)}\right] dx\, dy, \end{aligned} \tag{8.37}$$

and, as in the two-level case, the other probabilities are obtained by using the appropriate limits for the integrals. For the case of $\rho \ll 1$, the approximate form of the probability distribution in Eq. (8.2) simplifies the calculation.

Although ρ_4 can be evaluated from Eq. (8.36) in the above manner, an alternative derivation that provides a more rapid approach to the desired result is used here. This approach follows the treatment of Hagen and Farley (1973) and is based on a theorem by Price (1958). The form of the theorem that we require is

$$\frac{d\langle r_4\rangle}{d\rho} = \sigma^2 \left\langle \frac{\partial \hat{x}}{\partial x} \frac{\partial \hat{y}}{\partial y} \right\rangle, \tag{8.38}$$

where r_4 is the unnormalized correlator output, and $\hat{x}$ and $\hat{y}$ are again the quantized versions of the input signals. For four-level sampling,

$$\frac{\partial \hat{x}}{\partial x} = (n-1)\delta(x + v_0) + 2\delta(x) + (n-1)\delta(x - v_0), \tag{8.39}$$

where δ is the delta function, and a similar expression can be written for $\partial\hat{y}/\partial y$. Equation (8.39) is the derivative of the function in Fig. 8.4. To determine the expectation of the product of the two derivatives on the right-hand side of Eq. (8.38), the magnitudes of each of the nine terms in the product of the derivatives must be multiplied by the probability of occurrence. Thus, for example, the term $(n-1)^2\delta(x + v_0)\delta(y + v_0)$ has a magnitude of $(n-1)^2$ and probability

$$\frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left[\frac{-2v_0^2}{2\sigma^2(1+\rho)}\right]. \tag{8.40}$$

By consolidating terms with equal probabilities we obtain

$$\frac{d\langle r_4\rangle}{d\rho} = \frac{1}{\pi\sqrt{1-\rho^2}}\left\{(n-1)^2\left[\exp\left(\frac{-v_0^2}{\sigma^2(1+\rho)}\right)+\exp\left(\frac{-v_0^2}{\sigma^2(1-\rho)}\right)\right]\right.$$
$$\left. + 4(n-1)\exp\left(\frac{-v_0^2}{2\sigma^2(1-\rho^2)}\right)+2\right\}, \tag{8.41}$$

and

$$\langle r_4\rangle = \frac{1}{\pi}\int_0^{\rho}\frac{1}{\sqrt{1-\xi^2}}\left\{(n-1)^2\left[\exp\left(\frac{-v_0^2}{\sigma^2(1+\xi)}\right)+\exp\left(\frac{-v_0^2}{\sigma^2(1-\xi)}\right)\right]\right.$$
$$\left. + 4(n-1)\exp\left(\frac{-v_0^2}{2\sigma^2(1-\xi^2)}\right)+2\right\}d\xi, \tag{8.42}$$

where ξ is a dummy variable of integration. To obtain the correlation coefficient ρ_4, $\langle r_4\rangle$ must be divided by the expectation of the correlator output when the inputs are identical four-level waveforms, as in Eq. (8.36):

$$\rho_4 = \frac{\langle r_4\rangle}{\Phi + n^2(1-\Phi)}, \tag{8.43}$$

where Φ is the probability that the unquantized level lies between $\pm v_0$, that is

$$\Phi = \frac{1}{\sigma\sqrt{2\pi}}\int_{-v_0}^{v_0}\exp\left(\frac{-x^2}{2\sigma^2}\right)dx = \operatorname{erf}\left(\frac{v_0}{\sigma\sqrt{2}}\right). \tag{8.44}$$

Equations (8.42)–(8.44) provide a relationship between ρ_4 and ρ that is equivalent to the Van Vleck relationship for two-level quantization.

The choice of values for n and v_0 is usually made to maximize the signal-to-noise ratio for weak signals, which we now derive. For $\rho \ll 1$, Eqs. (8.42) and (8.43) reduce to

$$(\rho_4)_{\rho\ll 1} = \rho\frac{2[(n-1)E+1]^2}{\pi[\Phi+n^2(1-\Phi)]}, \tag{8.45}$$

where $E = \exp(-v_0^2/2\sigma^2)$. The variance in the measurement of r_4 is

$$\sigma_4^2 = \langle r_4^2\rangle - \langle r_4\rangle^2 = \langle r_4^2\rangle - \rho_4^2\left[\Phi + n^2(1-\Phi)\right]^2. \tag{8.46}$$

The factor $[\Phi + n^2(1-\Phi)]$ is the variance of the quantized waveform and here takes the place of σ^2 in the corresponding equations for unquantized sampling. Again we follow the procedure explained for the unquantized case, and write

$$\langle r_4^2 \rangle = N^{-2} \sum_{i=1}^{N} \langle \hat{x}_i^2 \hat{y}_i^2 \rangle + N^{-2} \sum_{i=1}^{N} \sum_{i \neq k} \langle \hat{x}_i \hat{y}_i \hat{x}_k \hat{y}_k \rangle. \tag{8.47}$$

To evaluate the first summation, note that $(\hat{x}_i \hat{y}_i)^2$ can take values of 1, n^2, or n^4, and the sum of these values multiplied by their probabilities is equal to $[\Phi + n^2(1 - \Phi)]^2$. The contribution of the second summation is

$$(1 - N^{-1})\rho_4^2 \left[\Phi + n^2(1 - \Phi)\right]^2 + 2N^{-1}[\Phi + n^2(1 - \Phi)]^2 \sum_{q=1}^{\infty} R_4^2(q\tau_s), \tag{8.48}$$

where the second term represents the effect of oversampling and is similar to Eq. (8.17), and R_4 is the autocorrelation function after four-level quantization. Thus from Eq. (8.46) we have

$$\sigma_4^2 = N^{-1}[\Phi + n^2(1 - \Phi)]^2 \left[1 + 2 \sum_{q=1}^{\infty} R_4^2(q\tau_s) - \rho_4^2 \right]. \tag{8.49}$$

Since we have assumed $\rho \ll 1$ the ρ_4^2 term can be neglected, and the signal-to-noise ratio for the four-level correlation measurement is

$$\mathcal{R}_{\text{sn4}} = \frac{\langle r_4 \rangle}{\sigma_4} = \frac{2\rho[(n-1)E + 1]^2 \sqrt{N}}{\pi \left[\Phi + n^2(1 - \Phi)\right] \sqrt{1 + 2 \sum_{q=1}^{\infty} R_4^2(q\tau_s)}}. \tag{8.50}$$

The signal-to-noise ratio relative to that for unquantized Nyquist sampling is obtained from Eq. (8.14) for $N = \beta N_N$, and is

$$\eta_4 = \frac{\mathcal{R}_{\text{sn4}}}{\mathcal{R}_{\text{sn}\infty}} = \frac{2[(n-1)E + 1]^2 \sqrt{\beta}}{\pi[\Phi + n^2(1 - \Phi)] \sqrt{1 + 2 \sum_{q=1}^{\infty} R_4^2(q\tau_s)}}. \tag{8.51}$$

For sampling at the Nyquist rate, $\beta = 1$ and

$$\eta_4 = \frac{\mathcal{R}_{\text{sn4}}}{\mathcal{R}_{\text{sn}\infty}} = \frac{2[(n-1)E + 1]^2}{\pi[\Phi + n^2(1 - \Phi)]}. \tag{8.52}$$

Values of η_4 very close to optimum sensitivity are obtained for $n = 3$ with $v_0 = 0.996\sigma$, and for $n = 4$, with $v_0 = 0.942\sigma$: see Table A8.1 in Appendix 8.3. Note that the choice of an integer for the value of n simplifies the correlator. For these two cases, η_4, the signal-to-noise ratio relative to that for unquantized sampling,

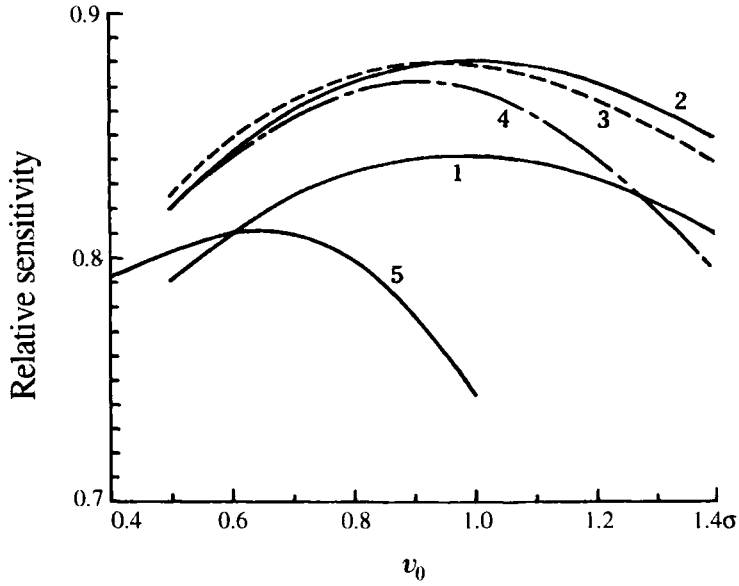


Figure 8.5 Signal-to-noise ratio relative to that for unquantized correlation for the four-level system and several modifications of it. The abscissa is the quantization threshold v_0 in units of the rms level of the waveforms at the quantizer input. The ordinate is sensitivity (signal-to-noise ratio) relative to an unquantized system. The curves are for (1) full four-level system with $n = 2$; (2) full four-level system with $n = 3$; (3) full four-level system with $n = 4$; (4) four-level system with $n = 3$ and low-level products omitted; (5) three-level system. From Cooper (1970).

is equal to 0.881 and 0.880, respectively. Curves of the relative sensitivity as a function of v_0/σ for $n = 2$, 3, and 4 are shown in Fig. 8.5. Similar conclusions are derived by Hagen and Farley (1973) and Bowers and Klingler (1974).

Having chosen values for n and v_0, we can now return to Eqs. (8.42) and (8.43) to examine the relationship of ρ and ρ_4. Curve 1 of Fig. 8.6 shows a plot of ρ and ρ_4. Extrapolation of a linear relationship with slope chosen to fit low values of ρ results in errors of 1% at $\rho = 0.5$, 2% at $\rho = 0.7$, and 2.8% at $\rho = 0.8$, where the error is a percentage of the true value of ρ. Thus for many purposes a linear approximation is satisfactory for values of ρ up to ~0.6. This linearity assumption simplifies the final step that we require in discussing four-level sampling, namely, calculation of the improvement in sensitivity resulting from oversampling.

The relationship between the autocorrelation function for unquantized noise R_∞ and that for the same waveform after four-level quantization is the same as for the corresponding cross-correlation functions in Eq. (8.45), so we can write

$$R_4 = \frac{2[(n-1)E+1]^2 R_\infty}{\pi[\Phi + n^2(1-\Phi)]}, \tag{8.53}$$

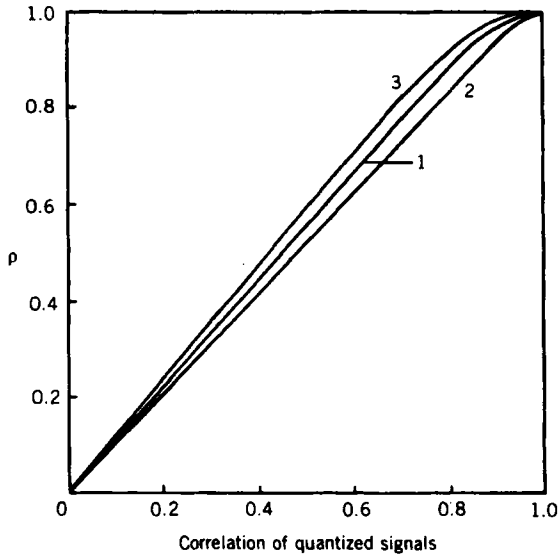


Figure 8.6 Correlation coefficient ρ for unquantized signals plotted as a function of the correlation that would be measured after quantization. The curves are for: (1) full four-level system with $n = 3$ and $v_0 = \sigma$, or $n = 4$ and $v_0 = 0.95\sigma$; (2) four-level system with low-level products omitted, $n = 4$ and $v_0 = 0.9\sigma$; (3) three-level system with $v_0 = 0.6\sigma$. From Cooper (1970).

provided that $R_\infty < \sim 0.6$. Now R_∞ as given by Eq. (8.15) fulfills this condition for $q = 1$ with an oversampling factor $\beta = 2$. For $n = 3$ and the corresponding optimum value of v_0, $E = 0.6091$, $\Phi = 0.6806$, and $R_4 = 0.881 R_\infty$. For $\beta = 2$, we use Eqs. (8.15) and (8.53) and Eq. (A8.5) of Appendix 8.1 to evaluate the summation in the denominator of Eq. (8.51), and obtain $\eta_4 = 0.935$, which is a factor of 1.06 greater than for $\beta = 1$. Bowers and Klingler (1974) have pointed out that the optimum value of the quantization level v_0 changes slightly with the oversampling factor. However, the optimum values are rather broad (see Fig. 8.5), and the effect on the sensitivity is very small.

In a discussion of two-bit quantization, Cooper (1970) considered the effect of omitting certain products in the multiplication process. For example, if all products of the two low-level bits are counted as zero instead of ± 1, the loss in signal-to-noise ratio is approximately 1% as shown in curve 4 of Fig. 8.5. The products to be accumulated are then only those counted as $\pm n$ and $\pm n^2$ in the full four-level system described above, and in the modified system they can be assigned values of ± 1 and $\pm n$, respectively, thereby simplifying the counter circuitry of the integrator. An even greater simplification can be accomplished by omitting the intermediate-level products also and assigning values ± 1 to the high-level products. This last type of modification yields 92% of the sensitivity of a full four-level correlator. We shall not analyze the case where only the low-level products are omitted, but we note that to derive the correlation coefficient as a function of ρ, one can express the action of the correlator in terms of two

different quantization characteristics (Hagen and Farley 1973) or else return to Eq. (8.36) and omit the appropriate terms. If both the low- and intermediate-level products are omitted, however, the action can be described more simply in terms of a new quantization characteristic, known as three-level quantization, without arbitrary omission of product terms.

Three-Level Quantization

Three-level quantization has proved to be an important practical technique, and the quantization characteristic is shown in Fig. 8.7. In this case the approach using Price's theorem will again be followed.

The expressions for the operating characteristics of a three-level correlator can be obtained from those in the preceding section by omitting the terms that refer to low- and intermediate-level products and adjusting the weighting factors as appropriate. Thus, the equivalent derivative needed in Price's theorem is

$$\frac{\partial \hat{x}}{\partial x} = \delta(x - v_0) + \delta(x + v_0), \tag{8.54}$$

and the expectation of the correlator output $\langle r_3 \rangle$ is, from Price's theorem,

$$\langle r_3 \rangle = \frac{1}{\pi} \int_0^{\rho} \frac{1}{\sqrt{1 - \xi^2}} \left[\exp\left(\frac{-v_0^2}{\sigma^2(1 + \xi)} \right) + \exp\left(\frac{-v_0^2}{\sigma^2(1 - \xi)} \right) \right] d\xi, \tag{8.55}$$

where ξ is a dummy variable of integration. The normalized correlation coefficient is

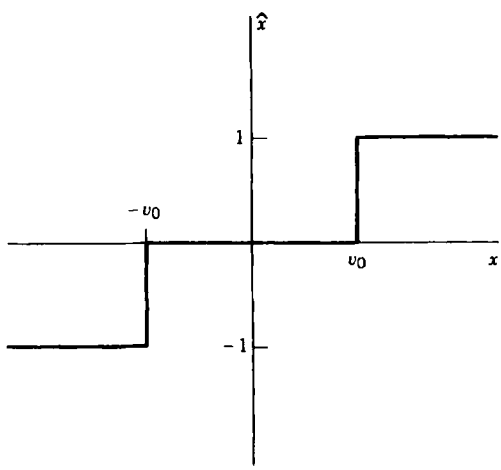


Figure 8.7 Characteristic curve for three-level quantization. The abscissa is the unquantized voltage x and the ordinate is the quantized output $\hat{x}$. v_0 is the threshold voltage. Since the magnitude of $\hat{x}$ takes only one nonzero value, it is perfectly general to set this value to unity.

$$\rho_3 = \frac{\langle r_3 \rangle}{1 - \Phi}, \tag{8.56}$$

where Φ is given by Eq. (8.44). For $\rho \ll 1$, Eqs. (8.55) and (8.56) yield

$$(\rho_3)_{\rho \ll 1} = \rho \frac{2E^2}{\pi(1 - \Phi)}, \tag{8.57}$$

where E is defined following Eq. (8.45). The variance of r_3 is

$$\sigma_3^2 = \langle r_3^2 \rangle - \langle r_3 \rangle^2 = N^{-1}(1 - \Phi)^2 \left[1 + 2 \sum_{q=1}^{\infty} R_3^2(q\tau_s) - \rho_3^2 \right], \tag{8.58}$$

where R_3 is the autocorrelation coefficient after three-level quantization. If ρ_3^2 in Eq. (8.58) can be neglected, the signal-to-noise ratio relative to a nonquantizing correlator is

$$\eta_3 = \frac{\mathcal{R}_{\text{sn}3}}{\mathcal{R}_{\text{sn}\infty}} = \frac{\langle r_3 \rangle}{\sigma_3 \mathcal{R}_{\text{sn}\infty}} = \frac{2\sqrt{\beta} E^2}{\pi(1 - \Phi)\sqrt{1 + 2\sum_{q=1}^{\infty} R_3^2(q\tau_s)}}. \tag{8.59}$$

For Nyquist sampling the maximum sensitivity relative to the nonquantizing case is obtained with $v_0 = 0.6120\sigma$, for which η_3 is equal to 0.810 (see curve 5 of Fig. 8.5). With this optimized threshold value, $\Phi = 0.4595$, $E = 0.8292$, and we can write $R_3(q\tau_s) = 0.810 R_\infty(q\tau_s)$, assuming that ρ is an approximately linear function of r_3. Then from Eqs. (8.15), (8.59), and (A8.5), we find that for a rectangular baseband spectrum with the oversampling factor $\beta = 2$, η_3 becomes 0.890, which is a factor of 1.10 greater than for $\beta = 1$. Table 8.1 summarizes the results for two-, three-, and four-level quantization.

TABLE 8.1 Efficiency Factor η_Q for Various Quantization Schemes

Number of Quantization Levels (Q)	η_Q, Sensitivity Relative to Unquantized Case	
	$\beta = 1$	$\beta = 2$
2	0.637	0.744
3	0.810	0.890
4	0.881[a]	0.935

[a] See also Table A8.1 (Appendix 8.3).

Quantization with Eight or More Levels

For eight or more quantization levels, the loss in efficiency resulting from the quantization is small, and a simpler, approximate method of calculating the loss can be used (Thompson 1998). The principle of the method is to calculate the fractional increase in the variance of a signal that results from the quantization. The signal-to-noise ratio at the correlator output is inversely proportional to this variance. This approach is of particular interest because it shows how the loss of sensitivity is related to the errors introduced by the quantization process.

Figure 8.8 shows a piecewise linear approximation of the Gaussian probability distribution of a signal from one antenna. This approximation simplifies the analysis. The intersections with the vertical lines indicate exact values of the Gaussian. For eight-level sampling, the quantization thresholds are indicated by the positions of the vertical lines between the numbers ± 3.5 on the abscissa. The widths of the levels are $\epsilon\sigma$ in voltage, that is, ϵ in units of σ where σ^2 is the unquantized variance. We consider first the case where the number of levels is even, as in Fig. 8.8. The probability that any one sample will fall between the two consecutive thresholds at $m\epsilon\sigma$ and $(m+1)\epsilon\sigma$, where m is an integer, is

$$\frac{1}{\sqrt{2\pi}\sigma}\left[e^{-m^2\epsilon^2/2} + e^{-(m+1)^2\epsilon^2/2}\right]\frac{\epsilon}{2}. \tag{8.60}$$

Any sample within this interval would be assigned the magnitude $(m + \frac{1}{2}\epsilon\sigma)$. For example, any voltage that falls within the level from 2 to 3 is encoded as amplitude 2.5, in units of $\epsilon\sigma$. The mean increase in the variance resulting from this representation is

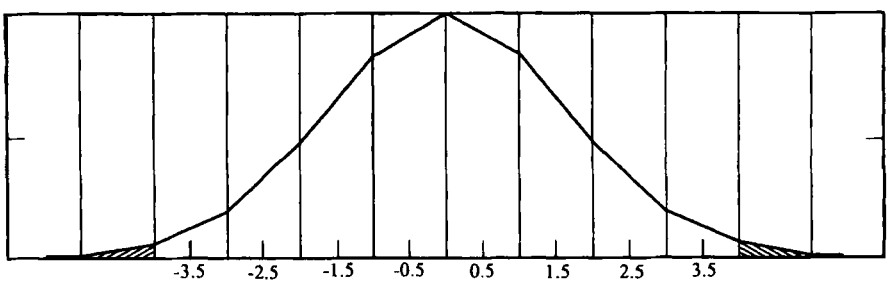


Figure 8.8 Piecewise linear representation of the Gaussian probability distribution of the amplitude of an IF signal. The intersections of the curve with the vertical lines denote exact values of the Gaussian. The abscissa is the signal amplitude (voltage) in units of $\epsilon\sigma$, and the numbers indicate the values assigned to the levels after quantization. For eight-level sampling the quantization thresholds are indicated by the seven vertical lines that lie between $-3.5\epsilon\sigma$ and $3.5\epsilon\sigma$ on the abscissa. For signal levels outside the range $\pm 4\epsilon\sigma$, indicated by the shaded areas, the assigned values are $\pm 3.5\epsilon\sigma$.

$$\frac{2}{\epsilon\sigma}\int_0^{\epsilon\sigma/2} x^2\,dx = \frac{1}{3}\left(\frac{\epsilon\sigma}{2}\right)^2. \tag{8.61}$$

This is exact for the piecewise linear probability curve in Fig. 8.8. The same increase in variance resulting from quantization applies to the range of signal levels from $-4\epsilon\sigma$ to $4\epsilon\sigma$ in Fig. 8.8. The fraction of the area under the Gaussian probability curve that lies between these levels is

$$\frac{1}{\sqrt{2\pi}}\int_{-4\epsilon\sigma}^{4\epsilon\sigma} e^{-x^2/2\sigma^2}\,dx = \operatorname{erf}\left(\frac{4\epsilon}{\sqrt{2}}\right). \tag{8.62}$$

Thus the variance resulting from quantization of the signal samples with amplitudes in the range $\pm 4\epsilon\sigma$ is

$$\frac{1}{3}\left(\frac{\epsilon\sigma}{2}\right)^2 \operatorname{erf}\left(\frac{4\epsilon}{\sqrt{2}}\right). \tag{8.63}$$

We shall assume that the quantization error is essentially uncorrelated with the unquantized signal. In the extreme case of two-level sampling the quantization error is highly correlated with the unquantized signal, so the treatment used here would not apply. Consider, however, the case of multilevel quantization as in Fig. 8.8. If the signal voltage is increased steadily, the quantization error decreases from a maximum at each quantization threshold to zero when the voltage is equal to the mid-point of two thresholds. At each threshold the quantization error changes sign and the cycle repeats. This behavior greatly reduces any correlation between the quantization error and the signal waveform.

It is also necessary to take account of the effect of counting all signals below $-4\epsilon\sigma$ as level $-3.5\epsilon\sigma$, and those above $+4\epsilon\sigma$ as $+3.5\epsilon\sigma$. To make an approximate estimate of this effect, we divide the range of signal level outside of $\pm 4\epsilon\sigma$ into intervals of width $\epsilon\sigma$. Consider, for example, the interval centered on $6.5\epsilon\sigma$. The probability of the signal falling within this level is equal to the corresponding area under the curve, which for the piecewise linear approximation is

$$\frac{1}{2}\frac{\epsilon}{\sqrt{2\pi}}\left[e^{-(6\epsilon)^2/2} + e^{-(7\epsilon)^2/2}\right]. \tag{8.64}$$

The square of the mean error resulting from quantization of the signal within this range is closely approximated by $[(6.5 - 3.5)\epsilon\sigma]^2$, so the total variance of the quantization error for signals outside the range $\pm 4\epsilon\sigma$ is

$$\frac{\epsilon^3\sigma^2}{\sqrt{2\pi}}\sum_{m=4}^{20}(m-3)^2\left[e^{-m^2\epsilon^2/2} + e^{-(m+1)^2\epsilon^2/2}\right]. \tag{8.65}$$

The upper limit of the summation in (8.65) is chosen to be large enough that increasing it does not significantly change the result. The quantization error re-

sulting from the truncation of the signal values outside the range $\pm 4\epsilon\sigma$ clearly has some degree of correlation with the unquantized signal level. However, this is a small effect because the fraction of samples for which the signal lies outside $\pm 4\epsilon\sigma$ is less than 1.6% for eight-level quantization, with ϵ optimized for sensitivity. The percentage decreases as the number of quantization levels increases. We shall therefore treat the quantization error resulting from the truncation of the signal peaks as uncorrelated with the signal, but bear in mind that this assumption may introduce a small uncertainty into the calculation.

The variance of the quantized signal is equal to the variance of the unquantized signal (σ^2) plus the variance of the quantization errors in (8.63) and (8.65), that is,

$$\sigma^2 + \frac{1}{3}\left(\frac{\epsilon\sigma}{2}\right)^2 \operatorname{erf}\left(\frac{4\epsilon}{\sqrt{2}}\right) + \frac{\epsilon^3\sigma^2}{\sqrt{2\pi}} \sum_{m=4}^{20} (m-3)^2 \left[e^{-m^2\epsilon^2/2} + e^{-(m+1)^2\epsilon^2/2}\right]. \tag{8.66}$$

If the variance is the same for both signals at the correlator input, and if the correlation of the signals is small (i.e., $\rho \ll 1$ as assumed for the two-, three-, and four-level cases), then the signal-to-noise ratio at the correlator output is inversely proportional to the variance. Thus the quantization efficiency is

$$\eta_{(2\mathcal{N})} = \left\{1 + \frac{1}{3}\left(\frac{\epsilon}{2}\right)^2 \operatorname{erf}\left(\frac{\mathcal{N}\epsilon}{\sqrt{2}}\right) + \frac{\epsilon^3}{\sqrt{2\pi}} \sum_{m=\mathcal{N}}^{\mathcal{N}+20} (m-\mathcal{N}+1)^2 \left[e^{-m^2\epsilon^2/2} + e^{-(m+1)^2\epsilon^2/2}\right]\right\}^{-1}. \tag{8.67}$$

Here the equation has been generalized for $2\mathcal{N}$ levels. For an odd number of levels, $2\mathcal{N}+1$, one of which is centered on zero signal level, the equivalent equation for the quantization efficiency is

$$\eta_{(2\mathcal{N}+1)} = \left\{1 + \frac{1}{3}\left(\frac{\epsilon}{2}\right)^2 \operatorname{erf}\left(\frac{(\mathcal{N}+\frac{1}{2})\epsilon}{\sqrt{2}}\right) + \frac{\epsilon^3}{\sqrt{2\pi}} \sum_{m=\mathcal{N}+1}^{\mathcal{N}+20} (m-\mathcal{N})^2 \left[e^{-[m-(1/2)]^2\epsilon^2/2} + e^{-[m+(1/2)]^2\epsilon^2/2}\right]\right\}^{-1}. \tag{8.68}$$

Results from Eqs. (8.67) and (8.68) are given in Table 8.2.

The second column of Table 8.2 gives the value of $\mathcal{N}$, and the third column gives ϵ. The values of ϵ have been chosen to minimize, approximately, the values of η_Q, and are by no means critical because the minimum is broad. The fourth column of the table gives P, which is the fraction of samples for which the signal amplitude is greater than $\pm\mathcal{N}\epsilon\sigma$ for an even number of levels or greater than

TABLE 8.2 Quantization Efficiency and Other Factors for Eight or More Levels

Number of Levels (Q)	$\mathcal{N}$	ϵ	P	η_Q
8	4	0.60	0.016	0.960
9	4	0.55	0.013	0.968
16	8	0.34	0.006	0.988
32	16	0.19	0.002	0.996

$\pm(\mathcal{N} + \frac{1}{2})\epsilon\sigma$ for an odd number of levels. For eight levels, P is the fraction of signal samples that contribute to the variance in (8.65).

The result for nine-level quantization can be compared with a corresponding result computed by F. R. Schwab using the more precise methods described for three and four levels. Schwab obtained $\eta_9 = 0.969$ for $\epsilon = 0.534$. The quantization efficiency varies only slowly with ϵ, and the value of η_9 from Eq. (8.68) agrees with the value obtained by Schwab to within ~0.1%, or ~3% in the degradation factor $(1 - \eta_9)$. This agreement verifies the present method within these limits of accuracy. Schwab also found that the quantization efficiency can be slightly improved by allowing the weights and level widths to increase with increasing signal amplitude, rather than using constant increments in weights and constant level widths. By optimizing in this manner, he obtained a value of 0.9655 for eight levels, which is about 0.5% higher than the value in Table 8.2. The values of η_Q in Table 8.2 are in similar agreement with results by Jenet and Anderson (1998), who give detailed calculations of performance for two- to eight-bit quantization, for both uniform and nonuniform threshold spacing. For cases with more than four levels, they use a Monte Carlo analysis. With values of quantization efficiency approaching unity, other effects, such as the departure of the bandpass responses from the ideal rectangular shape, become limiting factors.

Quantization Correction

As a result of quantization, the output of the correlator is only an approximately linear function of the cross-correlation, ρ. In particular, for two-level quantization, the correlation is proportional to a sine function of the correlator output, as in Eq. (8.25). For weak signals for which $\rho \ll 1$, the nonlinearity can be neglected, but for some observations a correction, sometimes referred to as the Van Vleck correction, is required. The corresponding precise relationship for four-level quantization can be obtained from Eqs. (8.42)–(8.44), and the linear approximation for the case $\rho \ll 1$ is given by Eq. (8.45). Similarly, for three-level quantization the precise relationship between the correlation and the correlator output is obtainable from Eqs. (8.44), (8.55), and (8.56), and the linear approximation is given by Eq. (8.57). As shown in Fig. 8.6, the relationships for three- and four-level quantization are substantially linear for values of ρ up to ~0.6. The linear approximation is therefore usually satisfactory for cross-correlation of signals from two antennas for which the uncorrelated noise on the signals,

and possible resolution of the source, limit the magnitude of the correlation. For very strong sources, or for measurement of autocorrelation of the signal from one antenna, the quantization correction is necessary. For each correlator output the correction must be applied once for each averaging period, which is likely to be in the range 10 ms to 10 s. For ease of computation the correlation can be expressed as a rational function, or similar approximation, of the correlator output; see Appendix 8.3 for four-level quantization. For three-level quantization, procedures for determination of ρ from the correlator output are given by Kulkarni and Heiles (1980) and D'Addario et al. (1984).

Comparison of Quantization Schemes

At this point it is useful to put into perspective the characteristics of quantization schemes, which are summarized in Tables 8.1 and 8.2. It should be remembered that the assumption $\rho \ll 1$ was used in determining these values. In considering the relative advantages of different quantization schemes, we note first that both the efficiency factor η_Q and the receiving bandwidth $\Delta\nu$ may be limited by the size and speed of the correlator system. The overall sensitivity is proportional to $\eta_Q\sqrt{\Delta\nu}$. Consider two conditions. In the first, the observing bandwidth is limited by factors other than the capacity of the digital system. This can occur in spectral line observing or when the interference-free band is of limited width. The sensitivity limitation imposed by the correlator system then involves only the efficiency factor η_Q in Table 8.1, and the choice of quantization scheme is one between simplicity and sensitivity. In the second case, the observing bandwidth is set by the maximum bit rate that the digital system can handle, as may occur in continuum observation in the higher-frequency bands. For a fixed bit rate ν_b the sample rate is ν_b/N_b, where N_b is the number of bits per sample, and the maximum signal bandwidth $\Delta\nu$ is $\nu_b/(2\beta N_b)$. Thus the sensitivity is proportional to $\eta_Q/\sqrt{\beta N_b}$, and this factor is listed for various systems in Table 8.3, in which $N_b = 1$ for $Q = 2$ and $N_b = 2$ for $Q = 3$ or 4. Note that oversampling always reduces the performance under these conditions. For those situations in which the capacity of the data processing is limited by the correlator system, the value of 0.64 for Nyquist sampling with two-level quantization results in the highest overall performance. Four-level sampling is almost as good, and four or more levels would be preferred if the bandwidth is limited as in spectral line observations. Encoding schemes involving nonintegral values of N_b are also of interest, for example, in

TABLE 8.3 Sensitivity Factor $\eta_Q/\sqrt{\beta N_b}$ for a Correlator-Limited System

Number of Quantization Levels (Q)	$\eta_Q/\sqrt{\beta N_b}$	
	$\beta = 1$	$\beta = 2$
2	0.64	0.52
3	0.57	0.45
4	0.62	0.47

tape recording of data. In that case the amount of information stored per bit is a prime consideration as discussed in Section 9.6.

A three-level × five-level correlator, for which the quantization efficiency factor η_Q is 0.86, has been constructed by Bowers et al. (1973) for spectral line mapping with a two-element interferometer. The use of different numbers of bits at the correlator inputs offers some simplification for a two-antenna system if the instrumental delay is applied to the signal with fewer quantization levels.

System Sensitivity

The relative sensitivity of different interferometer schemes resulting from characteristics of the analog processing is discussed in Section 6.2 (see Table 6.1). In systems with digital processing the quantization noise introduces further considerations. For example, with an analog correlator the sine × sine and cosine × cosine products for signals from two antennas provide, in principle, exactly the same information. However, with a digital correlator the quantization noise is largely uncorrelated between the sine and cosine components of the signal, so the quantization loss can be reduced by generating both products and averaging them.

8.4 ACCURACY IN DIGITAL SAMPLING

Principal Causes of Error

Deviations from ideal performance in practical samplers result in errors that, if not corrected for, can limit the accuracy of maps synthesized from the data. Once the signal is in digital form, however, the rate at which errors are introduced is usually negligibly small.

Two-level samplers, which sense only the sign of the signal voltages, are the simplest samplers to construct. The most serious error that is likely to occur is in the definition of the zero level, in which a small voltage offset may occur. The effect of offsets in the samplers is to produce small offsets of positive or negative polarity in the correlator outputs, which can be largely eliminated by phase switching, as described in Section 7.5. Alternatively, the offsets in the samplers can be measured by incorporating counters to compare the numbers of positive and negative samples produced. Correction for the offsets can then be applied to the correlator output data [see, e.g., Davis (1974)].

In samplers with three or more quantization levels, the performance depends on the specification of the levels with respect to the rms signal level, σ. An automatic level control (ALC) circuit is therefore sometimes used at the sampler input. Errors resulting from incorrect signal amplitude become less important as the number of quantization levels is increased; with many levels the signal amplitude becomes simply a linear factor in the correlator output. In systems using complex correlators, two samplers are usually required for each signal, one at each output of a quadrature network. The accuracy of the quadrature network, and the possible errors in the relative timing of the two sample pulses, can also introduce errors in the data.

Tolerances in Three-Level Sampling

The results in this section are based largely on a study of accuracy requirements in three-level sampling by D'Addario et al. (1984). We start by considering the diagram in Fig. 8.9, which shows the sampling thresholds for a pair of signals to be correlated. Thresholds v_1 and $-v_2$ apply to the signal waveform $x(t)$ and v_3 and $-v_4$ to $y(t)$. The probability distribution of x and y is given by Eq. (8.1), and the correlator output is proportional to this probability integrated over the (x, y) plane with the weighting factors ± 1 and zero indicated in the figure. This approach enables us to investigate the effect of deviations of the sampler thresholds from the optimum, $v_0 = 0.612\sigma$. For three-level sampling the correlator output can be written

$$\langle r_3(\boldsymbol{\alpha}, \rho)\rangle = [L(\alpha_1, \alpha_3, \rho) + L(\alpha_2, \alpha_4, \rho) - L(\alpha_1, \alpha_4, -\rho) - L(\alpha_2, \alpha_3, -\rho)], \tag{8.69}$$

where $\alpha_i = v_i/\sigma$, and

$$L(\alpha_i, \alpha_k, \rho) = \int_{\alpha_i}^{\infty}\int_{\alpha_k}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[\frac{-(X^2 + Y^2 - 2\rho XY)}{(1-\rho^2)}\right] dX\, dY. \tag{8.70}$$

Here $X = x/\sigma$, $Y = y/\sigma$, and the integrand in Eq. (8.70) is the right-hand side of Eq. (8.1) with the variables measured in units of σ.

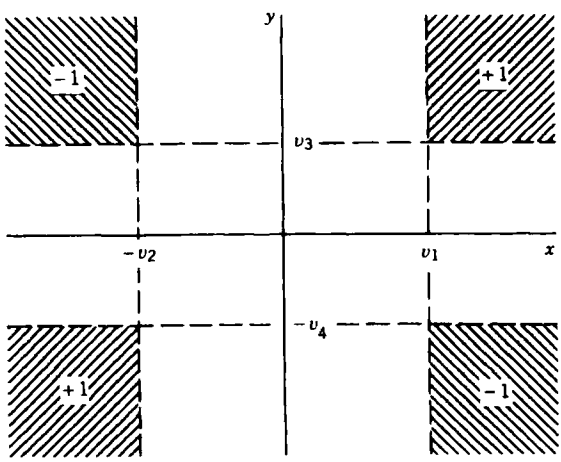


Figure 8.9 Threshold diagram for a correlator, the inputs of which are three-level quantized signals. x and y represent the unquantized signals, and the shaded areas show the combinations of input levels for which the output is nonzero.

D'Addario et al. (1984) point out that since less than 5% loss in signal-to-noise ratio occurs for threshold departures of ±40% from optimum, the required accuracy of the threshold settings, in practice, depends mainly on the correction algorithm. Suppose that the thresholds are kept close to, but not exactly equal to, the optimum value. For the x sampler in Fig. 8.9 the deviations from the ideal threshold value α_0 can be expressed in terms of an even part

$$\Delta_{gx} = \tfrac{1}{2}(\alpha_1 + \alpha_2) - \alpha_0, \tag{8.71}$$

and an odd part

$$\Delta_{ox} = \tfrac{1}{2}(\alpha_1 - \alpha_2). \tag{8.72}$$

For the y sampler Δ_{gy} and Δ_{oy} are similarly defined. The Δ_g terms produce gain errors. They are equivalent to an error in the level of the signal at the sampler, and they have the effect of introducing a multiplicative error in the measured cross-correlation. The Δ_o terms produce offset errors in the correlator output and are potentially more damaging since such errors can be large compared with the low levels of cross-correlation resulting from weak sources. The offset errors, however, can be removed with high precision by phase switching. The cancellation of the offset results from the sign reversal of the digital samples, or of the correlator output, as described in Section 7.5. The correlator output of a phase-switched system is of the form

$$r_{3s}(\boldsymbol{\alpha}, \rho) = \tfrac{1}{2}\left[r_3(\boldsymbol{\alpha}, \rho) - r_3(\boldsymbol{\alpha}, -\rho)\right]. \tag{8.73}$$

If all α values are within ±10% of α_0, the output is always within 10^{-3} (relative error) of the output of a correlator with the same gain errors, but no offset errors, in the samplers. Note also from Fig. 8.5 that 10% errors in the threshold settings result in less than 1% loss in signal-to-noise ratio. Thus, with phase switching, 10% errors are tolerable in the thresholds, and the effects of the gain errors can be corrected for if the actual threshold levels are known. Since the probability density distribution of the signal amplitudes can be assumed to be Gaussian, the threshold levels can be determined by counting the relative numbers of +1, 0, and −1 outputs from each sampler. When ρ is small (a few percent), a simple correction for the gain error can be obtained by dividing the correlator output by the arithmetic mean of the numbers of high-level (±1) samples for the two signals. Then 10% errors in the threshold settings result in errors of less than 1% in ρ.

Another nonideal aspect of the behavior of the sampler and quantizer is that the threshold level may not be precisely defined but may be influenced by the direction and rate of change of the signal voltage, the previous sample value (hysteresis), and other effects. The result can be modeled by including an indecision region in the sampler response extending from $\alpha_k - \Delta$ to $\alpha_k + \Delta$. It is assumed that a signal that falls within this region results in an output that takes either of

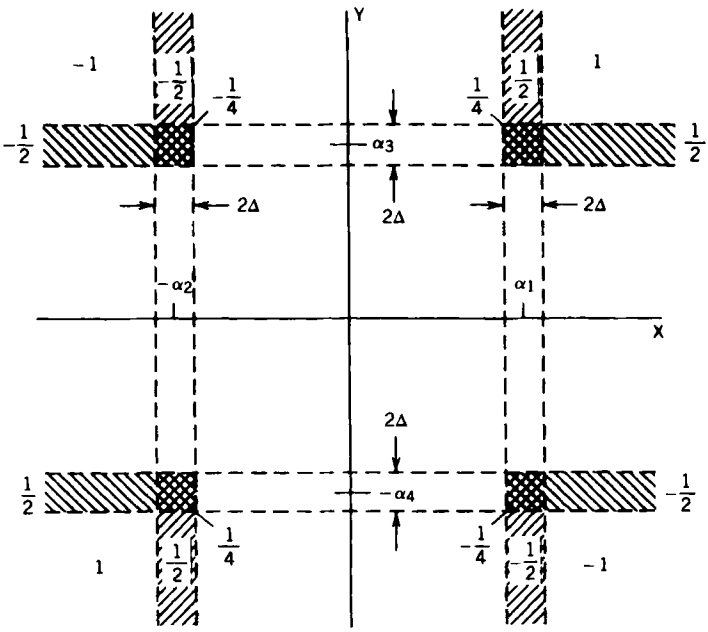


Figure 8.10 Threshold diagram for a three-level correlator showing indecision regions and the shaded areas within them for which the response is nonzero. The figures ± 1, $\pm\frac{1}{2}$, and $\pm\frac{1}{4}$ indicate the correlator response. The diagram shows the (X, Y) plane in which the signals are normalized to the rms value σ.

the two values associated with the threshold randomly and with equal probability. The threshold diagram with indecision regions included is shown in Fig. 8.10.

The weighting in the indecision regions depends on the probability of the random sample values and is 1/4 when both signals fall within indecision regions, and 1/2 when one signal is within an indecision region and the other produces a nonzero output. As before, the correlator output can be obtained by integrating the weighted probability of the signal values over the (X, Y) plane. Figure 8.11 shows the decrease in the correlator output as a function of Δ for several values of ρ, computed by expressing the output decrease as a Maclaurin series in Δ (D'Addario et al. 1984). For all cases except those where ρ approaches unity, the relatively small decrease in output results from the fact that when one input waveform falls within an indecision region, the other generally does not. For the particular case of $\rho = 1$, the input waveforms are identical and fall within these regions simultaneously. The output decrease is then proportional to Δ as shown by the broken line in Fig. 8.11: however, this case is not of much practical importance. For a 1% maximum error, Δ must not exceed 0.11σ, so the indecision region can be as large as $\pm 18\%$ of the threshold value. For a maximum error of 0.1% the above limits must be divided by $\sqrt{10}$. Thus the indecision regions have large enough tolerances that their effect can often be neglected.

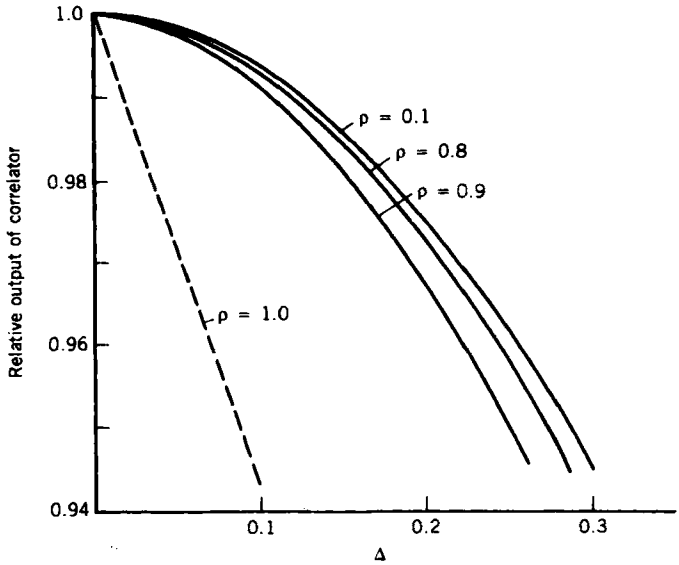


Figure 8.11 Effect of indecision regions on the output of a three-level correlator. The thresholds are assumed to be set to the optimum value 0.612σ, and the widths of the indecision regions are $2\sigma\Delta$. The output is given as a fraction of the output for $\Delta = 0$.

8.5 DIGITAL DELAY CIRCUITS

Time delays that are multiples of the sample interval can be applied to streams of digital bits by passing them through shift registers that are clocked at the sampling frequency. Shift registers with different numbers of stages thus provide different fixed delays. A method of using two shift registers to obtain a delay that is variable in increments of the clock pulse interval is described by Napier, Thompson, and Ekers (1983). However, integrated circuits for random access memory (RAM), developed for computer applications, provide the most economical solution for large digital delays.

Another useful technique is serial-to-parallel conversion, that is, the division of a bit stream at frequency ν into n parallel streams at frequency ν/n, where n is a power-of-two integer. This allows the use of slower and more economical types of digital circuits for delay, correlation, and other processes.

The precision required in setting a delay has been discussed in Section 7.3 under *Delay-Setting Tolerances*, and is usually some fraction of the reciprocal analog bandwidth. In any form of delay that operates at the frequency of the sampler clock, the basic delay increment is the reciprocal of the sampling frequency. A finer delay step can be obtained digitally by varying the timing of the sample pulse in a number of steps, for example, 16, between the basic timing pulses. Thus, if an extra delay of, say, 5/16 of a clock interval is required, the sampler is activated 11/16 of a clock interval after the previous clock pulse, and the data are held for 5/16 of an interval to bring them into phase with the clock-pulse timing.

Correction for delay steps equal to the sampling interval can also be made after the signals have been cross-correlated by applying a phase correction to the cross power spectrum, as described in Section 9.7 under *Discrete Delay Step Loss.*

8.6 QUADRATURE PHASE SHIFT OF A DIGITAL SIGNAL

We have mentioned that complex correlators for digital signals can be implemented by introducing the quadrature phase shift in the analog signal, as in Fig. 6.3, and then using separate samplers for the signal and its phase-shifted version. The Hilbert transformation that the phase shift represents can also be performed on the digital signal, thus eliminating the quadrature network and saving samplers and delay lines, but the accuracy is limited. Hilbert transformation is mathematically equivalent to convolution with the function $(-\pi\tau)^{-1}$, which extends to infinity in both directions [see, e.g., Bracewell (2000) p. 364]. A truncated sequence of the same form, for example, $\frac{1}{3}, 0, 1, 0, -1, 0, -\frac{1}{3}$, provides a convolving function for the digital data that introduces the required phase shift. However, the truncation results in convolution of the resulting signal spectrum with the Fourier transform of the truncation function, that is, a sinc function. This introduces ripples and degrades the signal-to-noise ratio by a few percent. Also, the summation process in the digital convolution increases the number of bits in the data samples, but the low-order bits can be discarded to avoid a major increase in the complexity of the correlator. This results in a further quantization loss. The overall result is that the imaginary output of the correlator suffers spectral distortion and some loss in signal-to-noise ratio relative to the real output. These effects are most serious in broad bandwidth systems, in which the high data rate permits only simple processing. Lo et al. (1984) have described a system in which the real part of the correlation is measured as a function of time offset, as described below for the spectral correlator, and the imaginary part is then computed by Hilbert transformation.

8.7 DIGITAL CORRELATORS

Correlators for Continuum Observations

In continuum observations the average correlation over the signal bandwidth is measured, and data on a finer frequency scale may not be required. In such cases the correlation of the signals is measured only for zero time-delay offset. Digital correlators can be designed to run at the sampling frequency of the signals, or at a submultiple resulting from dividing the bit stream from the sampler into a number of parallel streams. In the latter case the number of correlator units must be proportionally increased, and their outputs can subsequently be additively combined. Two-level and three-level correlators, for which the products are represented by values of -1, 0, and $+1$, are the simplest to construct. Correlators in which one of the inputs is a two-level or three-level signal and the other input is more highly quantized also have a degree of simplicity. In this case, the correlator is essentially

an accumulating register into which the higher-quantization value is entered. The two-level or three-level value is used to specify whether the other number is to be added, subtracted, or ignored. In correlators in which both inputs have more than three levels of quantization, the multiplier output for any single product can be one of a range of numbers. One method of implementing such a multiplier is to use a read-only memory unit as a lookup table in which the possible product values are stored. The input bits to be multiplied are used to specify the address of the required product in the memory.

The output of a multiplier can take both positive and negative values, and, ideally, an up–down counter is required as an integrator. Since such counters are usually slower than simple adding counters, two of the latter are sometimes used to accumulate the positive and negative counts independently. Another technique is to count, for example, -1, 0, and $+1$ as 0, 1, or 2, and then subtract the excess values, in this case equal to the number of products, in the subsequent processing.

Spectral line correlators are also often used for continuum observations. They offer advantages such as the ability to reject narrowband interfering signals, or to subdivide the band into narrower sub-bands to reduce the smearing effect resulting from wide bandwidth, as discussed in Section 6.3.

Principles of Digital Spectral Measurements

In spectral line observations, measurements at different frequencies across the signal band are required. These measurements can be obtained by digital techniques using a spectral correlator system, which is most commonly implemented by measuring the correlation of the signals as a function of time offset (also referred to as time lag). The Fourier transform of this quantity is the cross power spectrum, which can be regarded as the complex visibility as a function of frequency. This Fourier transform relationship, which is related to the Wiener–Khinchin relation, is discussed in Section 3.2. In the case of an autocorrelator (for use with a single antenna), the two input signals are the same waveform with a time offset. Thus the autocorrelation function is symmetric, and the power spectrum is entirely real and even. However, the cross power spectrum of the signals from two different antennas is complex, and the cross-correlation function has odd as well as even parts.

The output of a spectral correlator system provides values of the visibility at N frequency intervals across the signal band. These intervals are sometimes spoken of as frequency channels, and their spacing as the channel bandwidth, by analogy with the analog type of spectral correlator in which the signal band is broken up into channels by a bank of N analog filters with separate correlators for each filter channel. To explain the action of a digital spectral correlator, we consider the cross power spectrum $\mathcal{S}(\nu)$ of the signals from two antennas, as shown in idealized form in Fig. 8.12. Here it is assumed that the source under observation has a flat spectrum with no line features, and the final IF amplifier before the sampler has a rectangular baseband response. In Fig. 8.12 we have included the negative frequencies since they are necessary in the Fourier transform relationships. For $-\Delta\nu \leq \nu \leq \Delta\nu$, the real and imaginary parts of $\mathcal{S}(\nu)$ have magnitudes a and

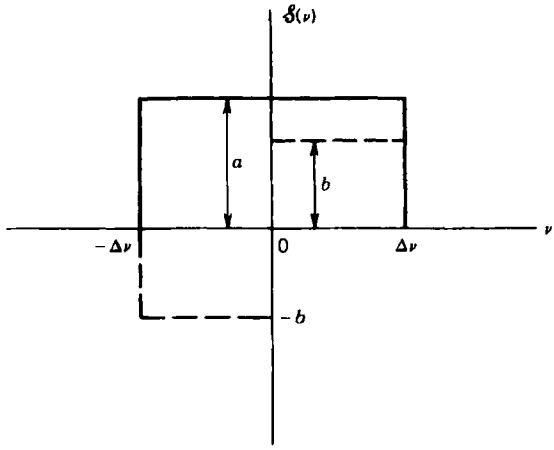


Figure 8.12 Cross power spectrum $\mathcal{S}(\nu)$ of two signals for which the power spectra are rectangular bands extending in frequency from zero to $\Delta\nu$. Negative frequencies are included. The full line represents the real part of $\mathcal{S}(\nu)$ and the broken line the imaginary part. The corresponding correlation function is derived in Eq. (8.74).

b, respectively, and the corresponding visibility phase is $\tan^{-1}(b/a)$. The cross-correlation function $\rho(\tau)$ is the Fourier transform of $\mathcal{S}(\nu)$, where τ is the time offset:

$$\begin{aligned}\rho(\tau) &= (a - jb)\int_{-\Delta\nu}^{0} e^{j2\pi\nu\tau}d\nu + (a + jb)\int_{0}^{\Delta\nu} e^{j2\pi\nu\tau}d\nu \\ &= 2\Delta\nu\left[a\frac{\sin(2\pi\Delta\nu\tau)}{2\pi\Delta\nu\tau} - b\frac{1-\cos(2\pi\Delta\nu\tau)}{2\pi\Delta\nu\tau}\right].\end{aligned} \tag{8.74}$$

Thus $\rho(\tau)$ has an even component of the form $\sin x/x$, which is related to the real part of $\mathcal{S}(\nu)$, and an odd component of the form $(1 - \cos x)/x$, which is related to the imaginary part. The spectral correlator measures $\rho(\tau)$ for integral values of the sampling interval τ_s. We consider the case of Nyquist sampling, for which $\tau_s = 1/(2\Delta\nu)$. The measured cross-correlation refers to the quantized waveforms, and the results in Section 8.3 show how this is related to the cross-correlation of the unquantized waveforms. For correlation levels that are not too large, the two quantities are closely proportional, so for simplicity we assume that Eq. (8.74) represents the behavior of the measured cross-correlation. The measurements are made with $2N$ time offsets from $-N\tau_s$ to $(N-1)\tau_s$ between the signals, and Fourier transformation of these discrete values yields the cross power spectrum at frequency intervals of $(2N\tau_s)^{-1} = \Delta\nu/N$ for Nyquist sampling. The N complex values of the positive frequency spectrum are the data required. Of these, the imaginary part comes from the odd component of $r(\tau)$. Thus, in the correlation measurement it suffices to use single-multiplier correlators to measure $2N$ real values of $r(\tau)$ over both positive and negative values of τ for one antenna with respect to the other. As an alternative to measuring only the real part of

the correlation, complex correlators could be used to measure both the real and imaginary parts, as in Fig. 6.3, for a range of time offsets from zero to $(N-1)\tau_s$. From a practical viewpoint, it is generally preferable to use single-multiplier correlators to avoid the broadband quadrature networks required in most complex correlators. Errors in the frequency responses of such networks can be a limiting factor in array performance.

Measurement of the cross-correlation over the limited time offset range is equivalent to measuring $r(\tau)$ multiplied by a rectangular function of width $2N\tau_s$. The cross power spectrum derived from the limited measurements is therefore equal to the true cross power spectrum convolved with the Fourier transform of the rectangular function, that is, with the sinc function

$$\frac{\sin(\pi \nu N/\Delta\nu)}{\pi \nu}, \tag{8.75}$$

which is normalized to unit area with respect to ν. Any line feature within the spectrum is broadened by the sinc function (8.75) and, depending on its frequency profile, may show the characteristic oscillating skirts. The width of the sinc function at the half-maximum level is $1.2\ \Delta\nu/N$, that is, 1.2 times the channel separation, and this width defines the effective frequency resolution.

The oscillations of the sinc function introduce structure in the frequency spectrum similar to the sidelobe responses of an antenna beam. They result from the sharp edges of the rectangular function that multiplies the correlation function. Such sidelobes are undesirable and can be reduced by choosing weighting functions, other than rectangular truncation, that are constrained to be zero outside the measurement range. Weighting functions are generally chosen to taper smoothly to zero at $|\tau| = N\tau_s$, thereby reducing unwanted ripples in the smoothing (convolving) function, but also to be as wide as possible in order to keep the width of the smoothing function as narrow as possible. These requirements are not generally compatible, so weighting functions that produce smoothing functions with very low sidelobes have poor frequency resolution. Some commonly used weighting functions are listed in Table 8.4. Hanning weighting, also known as raised cosine weighting, reduces the first sidelobe by a factor of 9, but degrades the resolution by 1.67, compared to uniform weighting. The Fourier transform of the

TABLE 8.4 Commonly Used Smoothing Functions

Weighting Function $w(\tau)$	$[w(\tau) = 0, \lvert\tau\rvert > \tau_l = N\tau_s]$	Half-Amplitude Width (Unit = $\Delta\nu/N$)	Peak Sidelobe
Uniform	$w(\tau) = 1$	1.21	0.22
Bartlett	$w(\tau) = 1 - (\lvert\tau\rvert/\tau_l)$	1.77	0.047
Hanning	$w(\tau) = 0.5 + 0.5\cos(\pi\tau/\tau_l)$	2.00	0.027
Hamming	$w(\tau) = 0.54 + 0.46\cos(\pi\tau/\tau_l)$	1.82	0.0073
Blackman	$w(\tau) = 0.42 + 0.50\cos(\pi\tau/\tau_l) + 0.08\cos(2\pi\tau/\tau_l)$	2.30	0.0012

Hanning weighting function is the sum of three sinc functions of relative amplitudes 0.25, 0.5, and 0.25. This is the smoothing function in the spectral domain, shown in Fig. 8.13b, which corresponds to the Hanning weighting. For the usual case where the number of points in the discretely sampled spectrum equals the number of points in the correlation function (i.e., no zero padding: see *FX Correlator*), the smoothing or convolution can be implemented as a three point running mean with relative weights of 0.25, 0.5, and 0.25. Thus the smoothed value of the cross power spectrum at frequency channel n is given by,

$$\mathcal{S}'\left(\frac{n\Delta\nu}{N}\right) = \tfrac{1}{4}\mathcal{S}\left[\frac{(n-1)\Delta\nu}{N}\right] + \tfrac{1}{2}\mathcal{S}\left(\frac{n\Delta\nu}{N}\right) + \tfrac{1}{4}\mathcal{S}\left[\frac{(n+1)\Delta\nu}{N}\right]. \quad (8.76)$$

The Hamming weighting function is very similar to the Hanning function and would appear to be superior because it produces a better resolution and a lower peak sidelobe level. However, the sidelobes of the Hamming smoothing function do not decrease in amplitude as rapidly as those of the Hanning smoothing function. Weighting functions are discussed in detail by Blackman and Tukey (1959) and Harris (1978).

A further effect of the finite time offset range complicates the calibration of the instrumental frequency response in the following way (Willis and Bregman 1981). The frequency responses of the amplifiers associated with the different an-

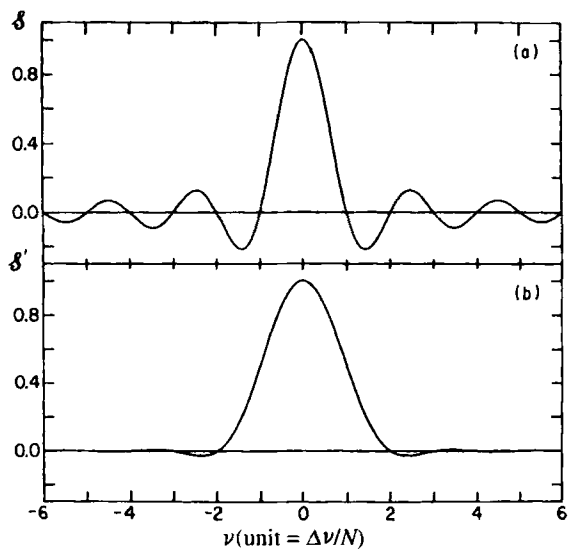


Figure 8.13 (a) The ordinate is the sinc function $\sin(\pi\nu N/\Delta\nu)/(\pi\nu N/\Delta\nu)$, which represents the frequency response of a spectral correlator with channels of width $\Delta\nu/N$ to a narrow line at $\nu = 0$. The abscissa is frequency ν measured with respect to the center of the received signal band. (b) The same curve after the application of Hanning smoothing as in Eq. (8.76).

tennas are seldom identical, as discussed in Section 7.3. To calibrate the response of each antenna pair over the spectral channels, it is usual to measure the cross power spectrum of an unresolved source for which the actual radiated spectrum is known to be flat across the receiving passband. We can consider the result in terms of the idealized power spectra in Fig. 8.12. If no special weighting function is used, the real and imaginary parts are both convolved with the sinc function (8.75). When a function with a sharp edge is convolved with a sinc function, the result is the appearance of oscillations (the Gibbs phenomenon) near the edge, as shown in Fig. 8.14. The point here is that the real component of $\mathcal{S}(\nu)$ in Fig. 8.12 is continuous through zero frequency, but the imaginary part shows a sharp sign reversal. Thus, near zero frequency the observed imaginary part of $\mathcal{S}(\nu)$ will show oscillations that may be as high as 18% in peak amplitude, whereas the real component will show relatively small oscillations at that point (see also Fig. 10.6b and associated text). As a result, the magnitude and phase measured for $\mathcal{S}(\nu)$ will show oscillations or ripples, the amplitude of which will depend on the relative amplitudes of the real and imaginary parts, that is, on the phase of the uncalibrated visibility. The uncalibrated phase measured for any source depends on instrumental factors such as the lengths of cables as well as the source position, which may not be known. In general, the phase will not be the same for the source under investigation and the calibrator. Hence, near zero frequency some precautions must be taken in applying the calibration. Possible solutions to the problem include (1) calibrating the real and imaginary parts separately, (2) observing over a wide enough band that the end channels in which the ripples are strongest can be discarded, or (3) applying smoothing in frequency to reduce the ripples.

Another problem encountered when observing a spectral line in the presence of a continuum background is caused by reflections in the antenna structure. These reflections cause a sinusoidal gain variation across the passband, the period of which is equal to the reciprocal of the delay of the signal caused by the reflection. In a correlation interferometer the magnitude of the ripple is a nearly constant fraction of the correlated continuum flux density, and the ripple is removed when the spectrum of the source under investigation is divided by the spectrum of the calibration source.

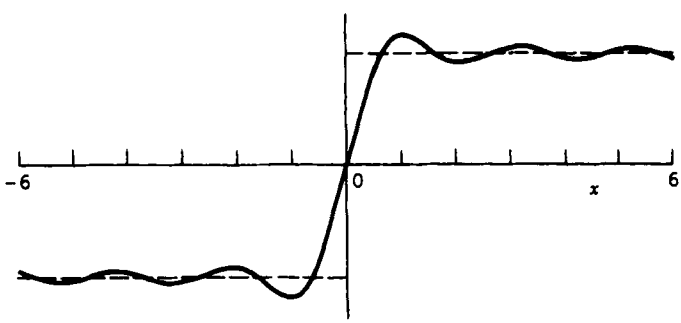


Figure 8.14 Convolution of a step function at the origin (broken line) with the sinc function $\sin(\pi x)/\pi x$. Here $x = \nu N/\Delta\nu$ and the half-cycle period of the ripple is approximately equal to the width of a spectral channel.

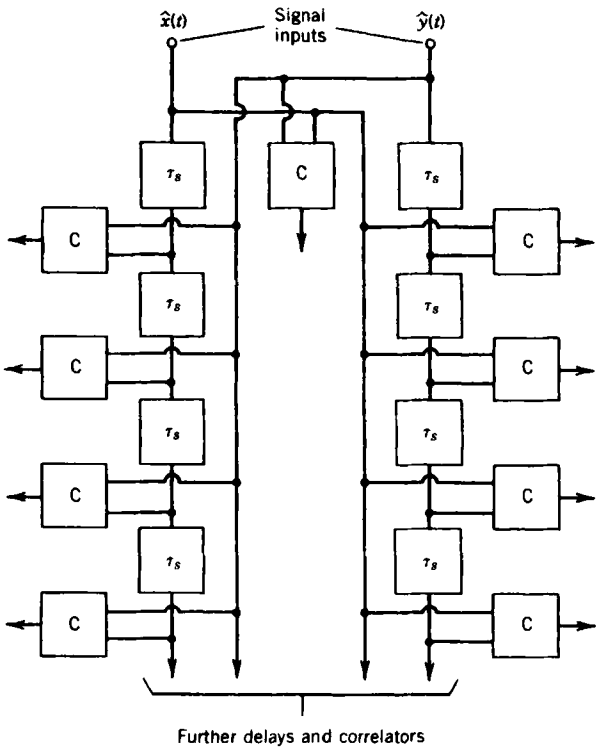


Figure 8.15 Simplified schematic diagram of a lag (XF) spectral correlator for two sampled signals. τ_s indicates a time delay equal to the sample interval and C indicates a correlator. The correlation is measured for zero delay, for the $\hat{x}$ input delayed with respect to the $\hat{y}$ input (left-hand correlator bank), and for $\hat{y}$ delayed with respect to $\hat{x}$ (right-hand correlator bank). The delays are integral multiples of τ_s.

Lag (XF) Correlator

A simplified schematic diagram of a lag correlator that measures the cross-correlation of two signals is shown in Fig. 8.15. Practical systems are often more complicated and are designed to take full advantage of the flexibility of digital processing techniques. The bandwidths of channels required for spectral line studies vary greatly, from a few hundred hertz to tens of megahertz. This versatility is necessary because the widths of spectral features are influenced by Doppler shifts, which are proportional to the rest frequencies of the lines and the velocities of the emitting atoms and molecules. Effects such as pressure broadening are also important. For this reason many digital spectral line systems incorporate a series of filters in the IF amplifiers so that the overall signal bandwidth can be reduced by factors $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$, and so on. When the signal bandwidth is halved, the Nyquist frequency is halved, and the samplers can be run at half the maximum frequency (or else every other sample can be deleted). However, if the correlators are run at the frequency used for the maximum bandwidth, the

data samples can be processed by the correlators twice, and the range of time offsets can thereby be doubled. As a result, the number of channels is doubled and the channel bandwidth decreased by a factor of 4. The use of this principle allows the signal bandwidths to be further decreased, and the number of channels increased, as required. Usually the number of channels is an integral power of 2, and the signal bandwidths are decreased by powers of 2 to be compatible with digital computing techniques. To implement the above scheme, recirculator units are required, which are basically memories that store blocks of input samples and allow them to be read out at the correlator input rate. These memory units are required in pairs, so that one is filled with data at the Nyquist rate appropriate to the chosen signal bandwidth, while the other is being read at the maximum data rate. One memory becomes filled in the time that the other is read for the required number of times, and the two are then interchanged. Correlators that incorporate the above principles are described as recirculating correlators (Ball 1973). An example of a recirculating lag correlator for a millimeter wavelength array is described by Okumura et al. (2000).

FX Correlator

The designation FX indicates a correlator in which Fourier transformation to the frequency domain is performed before cross multiplication of data from different antennas. In such a correlator the input bit stream from each antenna is converted to a frequency spectrum by a real-time FFT, and then for each antenna pair the complex amplitudes for each frequency are multiplied to produce the cross power spectrum. A major part of the computation occurs in the Fourier transformation, for which the total number of operations is proportional to the number of antennas. In comparison, in a lag correlator (also sometimes called an XF correlator), the total computation is largely proportional to the number of antenna pairs. Thus the FX scheme offers some economy in hardware, especially if the number of antennas is large. The principle of the FX correlator, based on the use of a special FFT computer, was discussed by Yen (1974) and first used in a large practical system by Chikada et al. (1984, 1987). A description of a system designed for a VLBI array is given by Benson (1995).

Two slightly different implementations of the FX correlator have been used. In one, both in-phase and quadrature components of the signal are sampled to provide a sequence of N complex samples, which is then Fourier-transformed to provide N values of complex amplitude, distributed in frequency. In the other, $2N$ real samples are transformed to provide N values of complex amplitude. In either case, N is the number of frequency channels (or points across the spectrum). Considerations are almost identical in the two cases, and we follow the second scheme in the discussion below.

Figure 8.16 is a schematic diagram of the basic operations of an FX correlator. The input sample stream from an antenna is Fourier transformed in contiguous sequences of length-$2N$ samples, where N is usually a power-of-two integer for efficiency in the FFT algorithm. The output of each transformation is a series of N complex signal amplitudes as a function of frequency. The frequency spacing

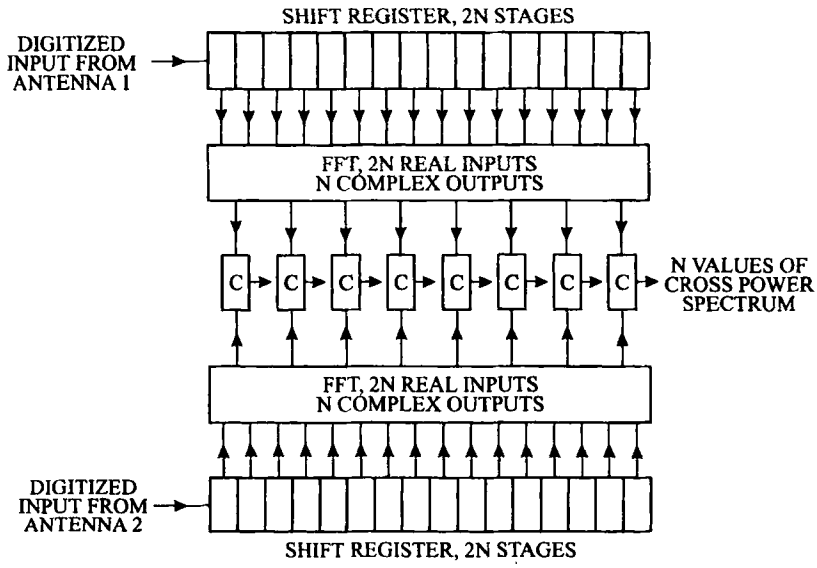


Figure 8.16 Simplified schematic diagram of an FX correlator for two antennas. The digitized signals are read into the shift registers and an FFT performed at intervals of $2N$ sample periods. The correlator elements, indicated by C, form products of one signal with the complex conjugate of the other. In an array with n_a antennas, the outputs of each FFT are split $(n_a - 1)$ ways for combination with the complex amplitudes from all other antennas.

of the data after transformation is $1/(2N\tau_s)$, where τ_s is the time interval between samples of the signals. In the cross-multiplication process that follows the FFT stage, the complex amplitude from one antenna of each pair is multiplied by the complex conjugate of the amplitude of the other. These multiplications occur in the correlator elements in Fig. 8.16. Note that the data in any one input sequence are combined only with data from other antennas for the same time sequence, and this restriction results in some loss of information. This loss can best be illustrated by considering the cross-correlation measurements (as they would be obtained in a lag correlator) that result from one $2N$-sample sequence for two antennas. Measurements of the correlation at the longest lags of $\pm(2N-1)\tau_s$ are obtained only once each, the next longest, $\pm(2N-2)\tau_s$, twice each, and so on to the unit lag interval for which there are $2N-1$ measurements. There is thus a triangular weighting of the correlation measurements as a function of lag, as shown in Fig. 8.17. The data for the longer lags are poorly represented. The situation can be improved by allowing contiguous sequences to overlap. With an overlap of 50%, which is one of the more common implementations, the data at the center of one sequence become those at the edge of a neighboring one. Weighting of the data across each sequence may also be introduced to reduce sidelobes in the frequency response, and the overlapping allows data that receive low weight in one sequence to be more highly weighted in the next. This type of procedure, known as weighted overlapping segment averaging, was originally described by

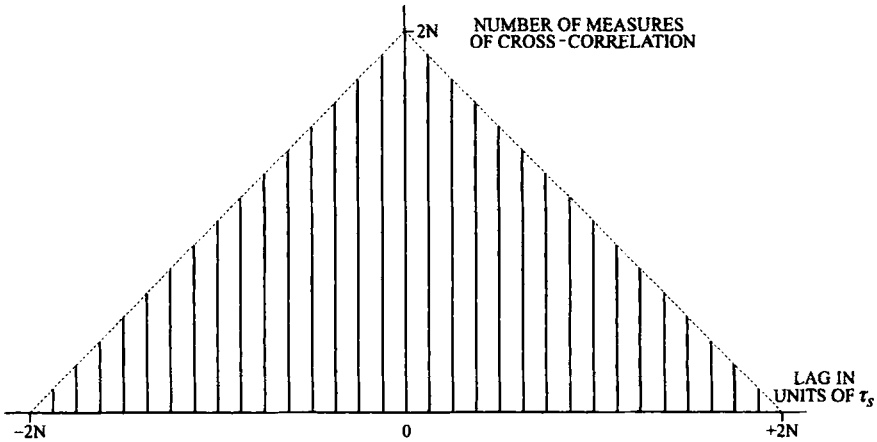


Figure 8.17 Number of effective cross-correlations for an FX correlator as a function of time delay between samples. N is the number of frequency channels (spectral points) across the signal band. For illustration we use $N = 8$, but in practice, values in the range 128 to 1024 are more typical.

Welch (1967) as a technique in general spectral analysis; for a detailed analysis, see Percival and Walden (1993).

It is interesting to show how the FX and lag methods can be made to give the same cross-correlation in the discrete signal case (Moran 1976). In the FX correlator, the Fourier transforms of the $2N$-point signals $\hat{x}(i)$ and $\hat{y}(i)$ are:

$$X(\nu) = \frac{1}{2N}\sum_{i=0}^{2N-1}\hat{x}(i)e^{-j2\pi\nu t_i}, \qquad Y(\nu) = \frac{1}{2N}\sum_{k=0}^{2N-1}\hat{y}(k)e^{-j2\pi\nu t_k}, \tag{8.77}$$

where the circumflex accent denotes a quantized variable. Let $t_i = i/2\Delta\nu$ and $t_k = k/2\Delta\nu$, where $\Delta\nu$ is the bandwidth, and let $\nu = \ell\Delta\nu/N$. Then the cross power spectrum, $\mathcal{S}(\nu) = X(\nu)Y^*(\nu)$, is

$$\mathcal{S}(\ell) = \frac{1}{(2N)^2}\sum_{i=0}^{2N-1}\sum_{k=0}^{2N-1}\hat{x}(i)\hat{y}(k)e^{-j\pi\ell(i-k)/N}. \tag{8.78}$$

The cross-correlation function is

$$\rho(n) = \sum_{\ell=0}^{2N-1}\mathcal{S}(\ell)e^{j\pi\ell n/N}. \tag{8.79}$$

Substituting Eq. (8.78) into Eq. (8.79), interchanging the order of summation, and noting that

$$\sum_{\ell=0}^{2N-1} e^{j\pi\ell(i-k-n)/N} = \begin{cases} 0 & \text{for } i-k-n \neq 0, \\ 2N & \text{for } i-k-n = 0, \end{cases} \tag{8.80}$$

we obtain

$$\rho(n) = \frac{1}{2N} \sum_{i=0}^{2N-1} \hat{x}(i)\hat{y}(i+n), 0 < (i+n) < (2N-1), \tag{8.81}$$

for n ranging between $-(2N-1)$ and $(2N-1)$. Thus from the response of the FX correlator we obtain the cross-correlation in the form in which it is measured by a lag correlator, except for the triangular weighting of the FX system, discussed with respect to Fig. 8.17.

The triangular weighting of the correlation data results in the convolution of the cross power spectrum with the function $\text{sinc}^2(2N\tau_s\nu)$. In comparison, in the lag correlator the cross-correlation measurements are uniformly weighted as a function of lag, and the cross power spectrum is convolved with a sinc function as in (8.75). The range of lags over which the cross-correlation is measured is $4N\tau_s$, which is the width of the triangle in Fig. (8.17), so according to the sampling theorem, the cross power spectrum should be represented by samples at intervals $1/(4N\tau_s)$ in frequency. The spectrum obtained as described above for the FX correlator is, by this criterion, undersampled by a factor of two. Thus if one wishes to obtain the cross-correlation function by Fourier transformation of the cross power spectrum values (e.g., to apply quantization correction), the undersampling results in aliasing and hence erroneous results. This situation can be avoided by using a Fourier transform of length $4N$ for the length-$2N$ input sequences, putting the data in positions 0 to $2N$ and zeros in positions $(2N+1)$ to $4N$ (O'Sullivan 1982, Granlund 1986). This practice is often referred to as *zero padding*.

With an FFT of length $2N$, equal to the length of the input data sequence, the response to a narrow spectral line that falls midway between two frequencies of the output spectrum is reduced by a factor $\text{sinc}^2(\frac{1}{2}) = 0.41$. With an FFT of length $4N$ and half the input sequence padded with zeros, a line that falls midway between two output points is reduced by $\text{sinc}^2(\frac{1}{4}) = 0.81$. If the amplitude loss resulting from the coarser sampling in frequency is acceptable, and if it is not necessary to transform back to cross-correlation for quantization correction, then the use of a length-$2N$ FFT is satisfactory. Otherwise the length-$4N$ FFT with zero padding of half the input range is advantageous.

Comparison of Lag and FX Correlators

Number of Operations. In a lag correlator the number of operations per second in the cross-correlation is equal to the product of the sample rate, $2\Delta\nu$, the number of different lags per baseline, $2N$, and the number of antenna pairs, $n_a(n_a-1)/2$, where N is the number of spectral points in the IF bandwidth $\Delta\nu$ and n_a is the number of antennas. The Fourier transformation to the frequency domain occurs

after integration of the cross-correlation data for a large number of cycles (typically $> 10^4$), and so its contribution to the total computation can be neglected. Thus as an estimate of the number of operations per second we have, for the lag correlator,

$$n_{\text{LAG}} \approx 2N\Delta\nu n_a(n_a - 1), \tag{8.82}$$

where each operation involves one real-number multiplication and one addition. In the FX correlator the Fourier transformation occurs once for each antenna every $2N$ samples, so the total rate of transformations is $\Delta\nu n_a/N$ per second. Each transform requires $N \log_2 N$ complex multiplications. Also, for the FX correlator there are $Nn_a(n_a - 1)/2$ products of the complex amplitude every $2N$ samples. We count each complex operation as four real-number operations, so the total number of equivalent real-number operations is

$$n_{\text{FX}} \approx 2\Delta\nu n_a(2\log_2 N + n_a - 1). \tag{8.83}$$

Equations (8.82) and (8.83) are regarded as approximations since, in practice, parameters depending on the details of implementation of the various operations may differ. However, the ratio $n_{\text{LAG}}/n_{\text{FX}} \approx Nn_a/(2\log_2 N + n_a)$ indicates the major effects, in which the factor N usually dominates. For example, if $N = 1000$ and $n_a = 10$, $n_{\text{LAG}}/n_{\text{FX}} \approx 330$. Note also that in the lag correlator the operations mainly involve numbers as they are generated by the sampler, that is, usually consisting of only one or two bits each. In the FX correlator the numbers are rapidly transformed in the FFT to values represented by many more bits, for example, 16 (six each for the real and imaginary parts and four for the mantissa). The advantage of the FX correlator is thus substantially less than indicated by the $n_{\text{LAG}}/n_{\text{FX}}$ ratio.

Integrated Circuit Implementation. Large correlators, that is, those for which the numbers of antennas and spectral channels are large, are practical because very large-scale integrated circuits (VLSI chips) can be designed specifically for a particular correlator design. These allow large numbers of parallel operations to be performed. For example, VLSI chips have been designed with 1024 lags of one input relative to the other. Two such chips would be required to implement a system of the form shown in Fig. 8.15, one for each of the correlator banks at the left- and right-hand sides of the diagram. The architecture of the lag correlator is somewhat more easily adapted to VLSI implementation than is that of the FX correlator. In particular, the number of bits per data point in the lag correlator is small compared with the multibit data representation required in the FX correlator, as explained in the previous paragraph. The smaller number of bits in the lag correlator greatly simplifies data input to the VLSI chip and interconnections between the circuits within it.

Digital Fringe Rotation. Although fringe rotation is often applied to the signal as an analog process, in some cases it is advantageous to implement it after dig-

itization. For example, in VLBI observations in which the data are recorded as digital samples, it is useful to be able to repeat the analysis with different fringe rates if the position of the source on the sky is not known with sufficient accuracy before the observation. Digital fringe rotation is applied to the digitized IF waveform just before it goes to the correlator, and involves multiplication with a digitized fringe rotation waveform as described in Section 9.7 under *Fringe Rotation Loss*. It is desirable to use a multibit representation for the rotated data to maintain the required accuracy, and thus the number of bits in the input data to the correlator is increased. As mentioned above, increasing the number of bits per sample in a lag correlator results in a proportional increase in complexity, and is not easily accommodated. Thus it may be necessary to truncate the data before input to the correlator, which effectively introduces the quantization loss a second time. In contrast, in the FX design multibit data representation is required in the FFT processing, so the bit increase that fringe rotation presents is more easily accommodated.

Fractional Sample Delay Correction. In digital implementation of the compensating delays, one way of adjusting the delay in steps smaller than the sampling interval is to adjust the timing of the sampler pulses, as described in Section 8.5. Another way of introducing a fractional sample period delay is done after transformation to the frequency domain by incrementing the phase values by an amount that varies in proportion to the frequency across the IF band. In the FX correlator this is easily done because the signals appear as an amplitude spectrum every FFT cycle, and the correction can be applied as required for each antenna, before the data are combined in antenna pairs. With a lag correlator there are two problems in this process. First, the transformation to a spectrum occurs after the data are combined for antenna pairs, so many more values require correction. Second, for long baselines the corrections required may occur more rapidly than the rate at which it is otherwise necessary for the cross-correlation values to be transformed to cross power spectra. Thus it may be possible to apply only a statistical correction rather than an exact one. See also Section 9.7 under *Discrete Delay Step Loss*.

Quantization Correction. The nonlinearity of the amplitude of the cross-correlation measured using coarsely quantized samples is seen in the Van Vleck relationship (Eq. 8.25) and the curves in Fig. 8.6. The true cross-correlation can be derived from the measured values by using these relationships. In the lag correlator this is a straightforward process because the cross-correlation values are directly calculated. To obtain the cross-correlation values in the FX correlator, the cross power spectrum data at the correlator output must be Fourier transformed from the frequency domain to the lag domain. After applying the correction, the data must then be transformed back to a frequency spectrum. Note that the correction is necessary only if the correlation of the total waveform (signal plus noise) is large for any pair of antennas. This condition implies observation of a source that is largely unresolved and sufficiently strong that the signal power in the receiver

is comparable to the noise, or greater. In the case of a spectral line observation, it is the power averaged over the receiver bandwidth that is important.

Editing of Invalid Data. Data editing to remove interference or other transient errors can be a serious problem in the FX correlator since bad or missing samples in a $2N$-sample sequence that is Fourier transformed may invalidate the whole sequence. The problem is more amenable to data editing in the case of the lag correlator since there is no particular length of input sequence that must be free from errors.

Adaptability. The FX design is somewhat more easily expanded or adapted to special requirements because more of the system is modularized per antenna rather than per baseline as in the lag correlator. Addition of an extra antenna to an FX correlator does not require such a complete restructuring as it does with a lag correlator.

Pulsar Observations. For pulsar observation, a gating system at the correlator output is required to separate data received during the pulsar-on period, so that the sensitivity is not degraded by noise received when the pulsar is off. For most pulsars, which have periods ≥ 0.1 s, time resolution of order 1 ms is adequate in the gating.[§] In the original lag correlator of the VLA, for example, a basic timing cycle requires that data readouts from the correlator occur at time intervals of 92.8 μs, which must be accommodated by the gating (Hankins 1999). This could be a limiting factor for very fast pulsars. With an FX correlator it is necessary to collect data in complete sequences of $2N$ samples, so the gating process has to accommodate data that arrive at time intervals of $\sim 2N\tau_s$. For example, with $N = 1000$ and a total bandwidth of 10 MHz, $2N\tau_s = 100$ μs. Again, this might restrict flexibility for the fastest pulsars. A nice feature of the FX correlator is that complete spectra are obtained during each $2N\tau_s$ interval in time. In the subsequent time averaging it is possible to process the frequency channels individually, and to vary the time of the gating pulse for each one so as to match the variation in pulse timing that results from dispersion in the interstellar medium.

Choice of Correlator Design. Because the relative advantages of the lag and FX schemes discussed above involve a number of different features, the best choice of architecture for any particular application may not be immediately obvious. Detailed design studies for different approaches, taking account of the precise requirements and the implementation of the VLSI circuits, are usually required. For further discussions of lag and FX correlators see D'Addario (1989) and Romney (1995).

[§]Many arrays can also be used in a phased-array mode, which provides one signal output per polarization. A specially designed pulsar processor can then provide measurements with high time resolution for study of the pulse profile and timing. In such cases the array is used only to provide a large collecting area for high sensitivity; see Section 9.9.

Hybrid Correlator

In designing a broadband correlator it may be advantageous to divide the analog signal from each antenna into n contiguous narrow sub-bands, where n is typically of order 10 or greater. A separate digital sampler is used for each such sub-band, and the correlator is designed as n sections operating in parallel to cover the full signal band. A system of this type that incorporates both analog filtering and digital frequency analysis is referred to as a *hybrid correlator*. If the digital part uses a lag design, then the rate of digital operations is reduced by a factor n relative to the rate for a lag correlator that processes the whole bandwidth without subdivision. This can be seen from Eq. (8.82), where for one sub-band the bandwidth is $\Delta\nu/n$, the number of channels required is N/n, but n such sections of digital processing are required. A hybrid correlator of this type is described by Weinreb (1984). However, if an FX implementation is used for the digital section, the hybrid scheme results in very little reduction in the number of operations, since in Eq. (8.83), N enters logarithmically. A general disadvantage of the hybrid correlator is that very careful calibration of the frequency responses of the sub-bands is required to avoid discontinuities in gain at the sub-band edges. In millimeter wavelength arrays IF bandwidths of order 10 GHz or more are practicable, and some analog filtering is necessary to subdivide such bandwidths down to a value for which Nyquist sampling is possible. In general it is advantageous to use the fastest samplers to minimize the analog filtering required.

Demultiplexing in Broadband Correlators

The bit rate for the very large scale (VLSI) integrated circuits used in large correlator systems is generally a few hundred Mbit s^{-1}, which is more than an order of magnitude slower than the digital samplers that are used with broadband correlators. Serial-to-parallel conversion at the sampler output, that is, demultiplexing in the time domain, allows use of optimum bit rates for the correlator. Consider a system in which each sampler output is demultiplexed into n streams, and assume for simplicity that there is one bit per sample; parallel architecture accommodates multiple bits. Any n contiguous samples all go to different streams. To obtain all the products required in a lag correlator for a pair of IF signals with this configuration of the data, it would be necessary to include cross-correlations between each stream of one signal with every stream of the other signal. To simplify the system, Escoffier (1997) has developed a scheme in which the n demultiplex bit streams from each signal are fed into a large random-access memory (RAM), and read out in reordered form. Each demultiplexed stream then contains a series of discontinuous blocks of $\sim 10^5$ samples. Each block contains data contiguous in time, as sampled. Cross-correlations are performed between data in corresponding blocks only. Thus for any pair of input signals, n cross-correlators running at the demultiplexed rate are required for each value of lag. Also each signal requires two RAM units so that one is filled as the other is read out. In Escoffier's system the sample rate is 4 Gbit s^{-1}, $n = 32$, and the length of a block of the demultiplexed data is approximately 1 ms. Since cross-correlations do not extend across the boundaries of any given block, there is a very small loss of efficiency which in

this case is about 0.2%. Another possible approach is based on demultiplexing in the frequency domain, as in the case of the hybrid correlator. It is then necessary only to cross-correlate corresponding frequency channels between each antenna, so the number of cross-correlators per signal pair is again equal to n for each lag. Carlson and Dewdney (2000) have described an all-digital development of the frequency demultiplexing principle used in the hybrid correlator. Broadband signals are digitized at full bandwidth, divided into frequency channels using digital filters, and resampled at the appropriate lower rate before cross-correlation. Thus the effect of small differences in the responses of analog filters is avoided. Both Escoffier's reordering scheme and demultiplexing in frequency provide approaches to the design of large broadband correlators. The latter requires fewer lags because the digital filters provide part of the spectral resolution.

For filtering sampled signals, digital filters of the FIR (finite impulse response) type can be used, in which the incoming sample stream is convolved with series of numbers, referred to as tap weights, the Fourier transform of which represents the filter response (Escoffier et al. 2000). The tap weights can be stored in a random-access memory and be readily changed as required. An advantage of digital filters is the high stability of the characteristics. However, it may be necessary to truncate the output data samples to match the number of bits per sample that can be handled by the correlator, and thus a further quantization loss may be incurred.

APPENDIX 8.1 EVALUATION OF $\sum_{q=1}^{\infty} r_{\infty}^2(q\tau_s)$

The periodic function $f(t)$ can be expressed as a Fourier series as follows:

$$f(t) = \frac{a_0}{2} + \sum_{q=1}^{\infty}\left[a_q \cos\left(\frac{2\pi qt}{\beta}\right) + b_q \sin\left(\frac{2\pi qt}{\beta}\right)\right], \tag{A8.1}$$

where β is the period and

$$\left.\begin{matrix} a_q \\ b_q \end{matrix}\right\} = \frac{2}{\beta}\int_0^{\beta} f(t)\left\{\begin{matrix}\cos \\ \sin\end{matrix}\right\}\left(\frac{2\pi qt}{\beta}\right) dt. \tag{A8.2}$$

Parseval's theorem for Eq. (A8.1) takes the form

$$\frac{2}{\beta}\int_0^{\beta} f^2(t)\,dt = \frac{a_0^2}{2} + \sum_{q=1}^{\infty}(a_q^2 + b_q^2). \tag{A8.3}$$

Now let $f(t)$ be a series of rectangular functions of unit height and width, one centered on $t = 0$ and the others centered on integral multiples of $\pm\beta$. Then one obtains

$$\begin{aligned} a_0 &= \frac{2}{\beta}, \qquad & a_q &= \frac{2}{\beta}\frac{\sin(\pi q/\beta)}{\pi q/\beta}, \\ b_q &= 0, \qquad & \int_0^{\beta} f^2(t)\,dt &= 1. \end{aligned} \tag{A8.4}$$

From Eqs. (A8.3) and (A8.4),

$$\sum_{q=1}^{\infty}\left[\frac{\sin(\pi q/\beta)}{\pi q/\beta}\right]^2 = \frac{\beta-1}{2}, \tag{A8.5}$$

which is the summation needed to evaluate Eq. (8.19).

APPENDIX 8.2 PROBABILITY INTEGRAL FOR TWO-LEVEL QUANTIZATION

The probability integration required in Eq. (8.21) can be performed as follows. The integral is

$$P_{11} = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}}\int_0^{\infty}\int_0^{\infty}\exp\left[\frac{-(x^2+y^2-2\rho xy)}{2\sigma^2(1-\rho^2)}\right]dx\,dy. \tag{A8.6}$$

Restore circular symmetry in the integral by the substitutions

$$z = \frac{y-\rho x}{\sqrt{1-\rho^2}}, \qquad dy = \sqrt{1-\rho^2}\,dz. \tag{A8.7}$$

Then

$$P_{11} = \frac{1}{2\pi\sigma^2}\int_0^{\infty} dx \int_{\frac{-\rho x}{\sqrt{1-\rho^2}}}^{\infty}\exp\left[\frac{-(x^2+z^2)}{2\sigma^2}\right]dz. \tag{A8.8}$$

Next substitute $x = r\cos\theta$ and $z = r\sin\theta$. The lower limit of the z integral in Eq. (A8.8) represents the line $z = -\rho x/\sqrt{1-\rho^2}$, which makes an angle θ with the x axis given by $\theta = -\sin^{-1}\rho$. The integral covers an area of the (x, z) plane between this line and the z axis ($\theta = \pi/2$). Thus

$$P_{11} = \frac{1}{2\pi\sigma^2}\int_0^{\infty} dr \int_{-\sin^{-1}\rho}^{\pi/2} r\exp\left(\frac{-r^2}{2\sigma^2}\right)d\theta. \tag{A8.9}$$

Finally, substitute $u = r^2/2\sigma^2$:

$$P_{11} = \frac{1}{2\pi}\int_0^{\infty} du \int_{-\sin^{-1}\rho}^{\pi/2} e^{-u}\,d\theta. \tag{A8.10}$$

Equation (A8.10) can be integrated directly to give

$$P_{11} = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho. \tag{A8.11}$$

TABLE A8.1 Optimal Thresholds and Efficiencies for Four-Level Quantization

n	v_0/σ	η_4
3	0.99568668	0.8811539496
3.3358750	0.98159883	0.8825181522
4	0.94232840	0.8795104597

APPENDIX 8.3 CORRECTION FOR FOUR-LEVEL QUANTIZATION

Schwab (1986) has investigated various aspects of the performance of correlators with four-level quantization. These include precise values for optimal thresholds and quantization efficiencies, and expressions for computation of the cross-correlation as a function of the correlator output. The threshold values and efficiencies are given Table A8.1.

The values of quantization efficiency η_4 for $n = 3$ and 4 are within 0.3% of the highest value, and are useful because nonintegral values of the weighting factor n would require more complicated implementation in a lag-type correlator.

Rational approximations for the cross-correlation $\tilde{\rho}$ are minimax solutions; that is, they minimize the maximum relative error. The variable r_N is the normalized correlator output, that is, the measured output divided by the corresponding output for $\rho = 1$. The first three approximations given below are valid for all $|r_N| \leq 1$.

For $n = 3$ and the corresponding value of v_0/σ in Table A8.1, the following approximation yields a maximum relative error of 1.51×10^{-4}:

$$\tilde{\rho}(r_N) = \frac{1.1347043 - 3.0971312r_N^2 + 2.9163894r_N^4 - 0.89047693r_N^6}{1 - 2.6892104r_N^2 + 2.4736683r_N^4 - 0.72098190r_N^6} r_N. \tag{A8.12}$$

For $n \approx 3.3359$ and the corresponding value of v_0/σ in Table A8.1, the following approximation yields a maximum relative error of 1.46×10^{-4}:

$$\tilde{\rho}(r_N) = \frac{1.1329552 - 3.1056902r_N^2 + 2.9296994r_N^4 - 0.90122460r_N^6}{1 - 2.7056559r_N^2 + 2.5012473r_N^4 - 0.73985978r_N^6} r_N. \tag{A8.13}$$

For $n = 4$ and the corresponding value of v_0/σ in Table A8.1, the following approximation yields a maximum relative error of 1.50×10^{-4}:

$$\tilde{\rho}(r_N) = \frac{1.1368256 - 3.0533973r_N^2 + 2.8171512r_N^4 - 0.85148929r_N^6}{1 - 2.6529114r_N^2 + 2.4027335r_N^4 - 0.70073934r_N^6} r_N. \tag{A8.14}$$

The following approximation also applies for $n = 4$ and the corresponding value of v_0/σ in Table A8.1, but is valid for only $|r_N| \leq 0.95$. It yields a maximum relative error of 2.77×10^{-5}:

$$\tilde{\rho}(r_N) = \frac{1.1369813 - 1.2487891r_N^2 + 4.5380174 \times 10^{-2}r_N^4 - 9.1448344 \times 10^{-3}r_N^6}{1 - 1.0617975r_N^2} r_N. \tag{A8.15}$$

REFERENCES

Abramowitz, M. and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1964, repr. by Dover, New York, 1965.

Ball, J. A., The Harvard Minicorrelator, *IEEE Trans. Instrum. Meas.*, **IM-22**, 193, 1973.

Benson, J. M. The VLBA Correlator, in *Very Long Baseline Interferometry and the VLA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron. Soc. Pacific Conf. Ser., **82**, 117–131, 1995.

Blackman, R. B. and J. W. Tukey, *The Measurement of Power Spectra*, Dover, New York, 1959.

Bowers, F. K. and R. J. Klingler, Quantization Noise of Correlation Spectrometers, *Astron. Astrophys. Suppl.* **15**, 373–380, 1974.

Bowers, F. K., D. A. Whyte, T. L. Landecker, and R. J. Klingler, A Digital Correlation Spectrometer Employing Multiple-Level Quantization, *Proc. IEEE*, **61**, 1339–1343, 1973.

Bracewell, R. N., *The Fourier Tramsform and Its Applications*, McGraw-Hill, New York, 2000 (earlier eds. 1965, 1978).

Burns, W. R. and S. S. Yao, Clipping Noise Loss in the One-Bit Autocorrelation Spectral Line Receiver, *Radio Sci.*, **4**, 431–436, 1969.

Carlson, B. R. and P. E. Dewdney, Efficient Wideband Digital Correlation, *Electronics Letters*, **36**, 987–988, 2000.

Chikada, Y., M. Ishiguro, H. Hirabayashi, M. Morimoto, K. Morita, K. Miyazawa, K. Nagane, K. Murata, A. Tojo, S. Inoue, T. Kanzawa, and H. Iwashita, A Digital FFT Spectro-Correlator for Radio Astronomy, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Univ. Press, Cambridge, UK, 1984, pp. 387–404.

Chikada, Y., M. Ishiguru, H. Hirabayashi, M. Morimoto, K. I. Morita, T. Kanzawa, H. Iwashita, K. Nakazimi, S. I. Ishiwaka, T. Takashi, K. Handa, T. Kasuga, S. Okumura, T. Miyazawa, T. Nakazuru, K. Miura, and S. Nagasawa, A 6 × 320-MHz 1024-Channel FFT Cross Spectrum Analyzer for Radio Astronomy, *Proc. IEEE*, **75**, 1203–1210, 1987.

Cole, T., Finite Sample Correlations of Quantized Gaussians, *Aust. J. Phys.*, **21**, 273–282, 1968.

Cooper, B. F. C., Correlators with Two-Bit Quantization, *Aust. J. Phys.*, **23**, 521–527, 1970.

D'Addario, L. R., Cross Correlators, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser., **6**, 59–82, 1989.

D'Addario, L. R., A. R. Thompson, F. R. Schwab, and J. Granlund, Complex Cross Correlators with Three-Level Quantization: Design Tolerances, *Radio Sci.*, **19**, 931–945, 1984.

Davis, W. F., Real-Time Compensation for Autocorrelation Clipper Bias, *Astron. Astrophys. Suppl.* **15**, 381–382, 1974.

Escoffier, R. P., *The MMA Correlator*, MMA Memo. 166, National Radio Astronomy Observatory, Socorro, New Mexico, 1997.

Escoffier, R. P., J. C. Webber, L. R. D'Addario, and C. M Broadwell, A Wideband Digital Filter using FPGAs, *Proc. SPIE*, **4015**, 106–113, 2000.

Goldstein, R. M., A Technique for the Measurement of the Power Spectra of Very Weak Signals, *IRE Trans. Space Electron. Telem.*, **8**, 170–173, 1962.

Granlund, J., *O'Sullivan's Zero-Padding*, VLBA Correlator Memo. 66, National Radio Astronomy Observatory, Charlottesville, VA, 1986.

Hagen, J. B. and D. T. Farley, Digital Correlation Techniques in Radio Science, *Radio Sci.*, **8**, 775–784, 1973.

Hankins T. H., Pulsar Observations at the VLA, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 613–624, 1999.

Harris, F. J., The Use of Windows for Harmonic Analysis with the Discrete Fourier Transform, *Proc. IEEE*, **66**, 51–83, 1978.

Jenet, F. A. and S. B. Anderson, The Effects of Digitization on Nonstationary Stochastic Signals with Applications to Pulsar Signal Baseband Recording, *Pub. Astron. Soc. Pacific*, **110**, 1467–1478, 1998.

Kulkarni, S. R. and C. Heiles, How to Obtain the True Correlation from a Three-Level Digital Correlator, *Astron. J.*, **85**, 1413–1420, 1980.

Lo, W. F., P. E. Dewdney, T. L. Landecker, D. Routledge, and J. F. Vaneldik, A Cross-Correlation Receiver for Radio Astronomy Employing Quadrature Channel Generation by Computed Hilbert Transform, *Radio Sci.*, **19**, 1413–421, 1984.

Moran, J. M., Very Long Baseline Interferometer Systems, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 174–197.

Napier, P. J., A. R. Thompson, and R. D. Ekers, The Very Large Array; Design and Performance of a Modern Synthesis Radio Telescope, *Proc. IEEE*, **71**, 1295–1320, 1983.

Nyquist, H., Certain Topics in Telegraph Transmission Theory, *Trans. Am. Inst. Electr. Eng.*, **47**, 617–644, 1928.

Okumura, S. K., M. Momose, N. Kawaguchi, T. Kansawa, T. Tsutsumi, A. Tanaka, T. Ichikawa, T. Suzuki, K. Ozeki, K. Natori, and T. Hashimoto, 1-GHz Bandwidth Digital Spectro-Correlator System for the Nobeyama Millimeter Array, *Publ. Astron. Soc. Japan*, **52**, 393–400, 2000.

Oppenheim, A. V. and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

O'Sullivan, J. D., *Efficient Digital Spectrometers—a Survey of Possibilities*, Note 375 Netherlands Foundation for Radio Astronomy, Dwingeloo, 1982.

Percival, B. D. and A. T. Walden, *Spectral Analysis for Physical Applications*, Cambridge Univ. Press, 1993, p. 289.

Price, R., A Useful Theorem for Nonlinear Devices Having Gaussian Inputs, *IRE Trans. Inf. Theory*, **IT-4**, 69–72, 1958.

Romney, J. D., Theory of Correlation in VLBI, in *Very Long Baseline Interferometry and the VLA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron. Soc. Pacific Conf. Ser., **82**, 17–37, 1995.

Schwab, F., *Two-Bit Correlators: Miscellaneous Results*, VLBA Correlator Memo. 75, Nat. Radio Astron. Obs., Charlotesville, VA, 1986.

Shannon, C. E., Communication in the Presence of Noise, *Proc. IRE*, **37**, 10–21, 1949.

Thompson, A. R., *Quantization Efficiency for Eight or more Sampling Levels*, MMA Memo 220, National Radio Astronomy Observatory, Socorro, NM, 1998.

Van Vleck, J. H. and D. Middleton, The Spectrum of Clipped Noise, *Proc. IEEE*, **54**, 2–19, 1966.

Weinreb, S., *A Digital Spectral Analysis Technique and Its Application to Radio Astronomy*, Technical Report 412, Research Lab for Electronics, MIT, Cambridge, MA, 1963.

Weinreb, S., Analog-Filter Digital-Correlator Hybrid Spectrometer, *IEEE Trans. Inst. Meas.*, **IM34**, 670–675, 1984.

Welch, P. D., The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short, Modified Periodograms, *IEEE Trans. Audio Electroacoust.*, **AU-15**, 70–73, 1967.

Willis A. G. and J. D. Bregman, Effects in Fourier Transformed Spectra, *User's Manual for Westerbork Synthesis Radio Telescope*, Netherlands Foundation for Radio Astronomy, Westerbork, Netherlands, 1981, Ch. 2, App. 2.

Yen, J. L., The Role of Fast Fourier Transform Computers in Astronony, *Astron. Astrophys. Suppl.*, **15**, 483–484, 1974.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

9 Very-Long-Baseline Interferometry

In 1967 a new technique of interferometry was developed in which the receiving elements were separated by such a large distance that it was expedient to operate them independently with no real-time communication link. This was accomplished by recording the data on magnetic tape for later cross-correlation at a central processing station. The technique was called very-long-baseline interferometry (VLBI), a term recalling the earlier long-baseline interferometers at Jodrell Bank Observatory, in which the elements were connected by microwave links that had reached 127 km in length. The principles involved in VLBI are fundamentally the same as those involved in interferometers with connected elements. The tape recorder can be considered as an IF delay line of limited capacity with an unusually long propagation time, weeks instead of microseconds. The use of tape recorders is motivated entirely by economics and places substantial limitations upon the system. Satellite links have been demonstrated (Yen et al. 1977), but their high cost discourages their use.

9.1 EARLY DEVELOPMENT

The motivation to develop VLBI came from the realization that many radio sources have structures that cannot be resolved by interferometers with baselines of a few hundred kilometers. By the mid-1960s it was well known that scintillation (discussed in Chapter 13) and time variability of the radiation from quasars implied angular sizes of < 0.01 arcsec. Maser emission from OH molecules at 18-cm wavelength was unresolved at 0.1 arcsec. Low-frequency burst radiation from Jupiter was believed to emanate from regions of small angular size. The aim of the first VLBI experiments was to measure the angular sizes of these radio sources. It is instructive to consider the operation of these early VLBI experiments in their most primitive form. Consider two telescopes with system temperatures T_{S1} and T_{S2}, which are pointed at a compact source giving antenna temperatures T_{A1} and T_{A2}. Each station records N data samples within the *coherence time*, that is, the interval during which the independent oscillators remain sufficiently stable that fringes can be averaged. In the subsequent processing these data streams are aligned, cross-correlated, and time-averaged after removing the quasisinusoidal fringes. The expected correlation for a point source is

$$\rho_0 \simeq \eta \sqrt{\frac{T_{A1} T_{A2}}{(T_{S1} + T_{A1})(T_{S2} + T_{A2})}}, \tag{9.1}$$

where η is a factor of value ~0.5 to account for losses due to quantization and processing (see Section 9.7). Here it is convenient to consider a normalized form of the visibility:

$$\mathcal{V}_N = \frac{\rho}{\rho_0} = \frac{\rho}{\eta} \sqrt{\frac{T_{S1} T_{S2}}{T_{A1} T_{A2}}}, \tag{9.2}$$

where ρ is the measured correlation, and we assume $T_A \ll T_S$. The rms noise level is

$$\Delta\rho \simeq \frac{1}{\sqrt{N}} \simeq \frac{1}{\sqrt{2\Delta\nu\tau_c}}, \tag{9.3}$$

where $\Delta\nu$ is the IF bandwidth and τ_c is the coherent integration time. Hence from Eqs. (9.1)–(9.3) the signal-to-noise ratio is

$$\frac{\rho}{\Delta\rho} = \eta \mathcal{V}_N \sqrt{\frac{T_{A1} T_{A2}}{T_{S1} T_{S2}} (2\Delta\nu\tau_c)}. \tag{9.4}$$

If the minimum useful signal-to-noise ratio is 4, the smallest detectable flux density is as follows, from Eqs. (1.3), (1.5), and (9.4):

$$S_{\min} \simeq \frac{8k}{\mathcal{V}_N \eta} \sqrt{\frac{T_{S1} T_{S2}}{A_1 A_2}} \frac{1}{\sqrt{2\Delta\nu\tau_c}}, \tag{9.5}$$

where k is Boltzmann's constant, and A_1 and A_2 are the antenna collecting areas. Typical parameters in 1967 were $A \simeq 250$ m^2 (25-m-diameter telescope), $T_S \simeq 100$ K, $\eta \simeq 0.5$, and $N = 1.4 \times 10^8$ bits (one bit per sample), the capacity of a tape at a standard density of 800 bpi (bits per inch) used in the NRAO Mark I system. For an unresolved source, $S_{\min} \simeq 2$ Jy. The development after three decades is indicated by the following parameter values: $A \simeq 1600$ m^2 (64-m-diameter telescope), $T_S \simeq 30$ K, and $N = 5 \times 10^{12}$ bits, the capacity of an instrumentation tape operated at 64 MHz bandwidth. For $\mathcal{V}_N = 1$, Eq. (9.5) gives $S_{\min} \simeq 0.6$ mJy. In both examples, the coherence time is assumed to be greater than the running time of the tape. The source size can be estimated from a single measurement of $\mathcal{V}_N$ by comparison with the visibility expected for a symmetric Gaussian model. Hence, as in Fig. 1.5, the full width at half maximum, a, is given by

$$a = \frac{2\sqrt{\ln 2}}{\pi u} \sqrt{-\ln \mathcal{V}_N}, \tag{9.6}$$

where u is the projected baseline (in wavelengths).

VLBI can be used only to study objects of exceedingly high intensity. Thus, the emission processes must normally be of nonthermal origin. To be detected on a baseline of length D, the source must be smaller than the fringe spacing. Since the flux density S is $2kT_B\Omega/\lambda^2$, where T_B is the brightness temperature, λ is the wavelength, and Ω is the source solid angle, the minimum detectable brightness temperature is

$$(T_B)_{\min} \simeq \frac{2}{\pi k} D^2 S_{\min}, \tag{9.7}$$

since $\Omega \simeq \pi(\lambda/2D)^2$. If $D = 10^3$ km and $S_{\min} = 2$ mJy, then $(T_B)_{\min} \simeq 10^6$ K. Therefore, observations of thermal phenomena occurring in molecular clouds, compact HII regions, and most stars are not possible. On the other hand, synchrotron sources such as supernova remnants, radio galaxies, and quasars, which are limited to 10^{12} K by Compton losses; masers in which $T_B \simeq 10^{15}$ K; and pulsars can be readily studied.

Three things were accomplished by early VLBI measurements:

1. Simple intensity distributions were derived by comparing measured visibilities with source models.
2. Masers were mapped by comparing fringe frequencies for different spectral features.
3. Source positions were measured to an accuracy of ~1 arcsec, and baselines to an accuracy of a few meters.

For a review of early techniques see Klemperer (1972). Since then the technique has moved steadily toward the mainstream of interferometry in terms of being able to produce reliable images of complex radio sources. The principal reason for this is the use of phase closure (see Section 10.3), which provides most of the phase information when a large enough number of antennas is available in the VLBI network.

9.2 DIFFERENCES BETWEEN VLBI AND CONVENTIONAL INTERFEROMETRY

In this section we briefly discuss the differences between VLBI and connected-element interferometry. Later sections in this chapter elaborate on these differences. Before beginning, we emphasize the theoretical unity of interferometry. The fundamental aim of all interferometry is to measure the coherence properties of the electromagnetic field. Thus the principles of connected-element interferometry and VLBI are basically identical. However, certain special techniques used in VLBI are needed because of the particular observational constraints. As the continuity of (u, v) coverage is improved, from a few meters to more than 10^5 km, with the largest spacing achieved by elements on distant satellites, and fiberoptic or other advanced communication systems make recording unnecessary, the concept of VLBI as a distinct technique will become a matter of history. Here

we deal with certain limitations that make classical VLBI practices somewhat distinct from those of connected-element interferometry.

Early VLBI experiments were conducted by organizing a diverse group of observatories that had been constructed for general radio astronomical research. Each telescope had its own limitations, calibration procedures, and management personnel. Various networks were formed to standardize procedures and automate the execution of VLBI experiments. Such ad hoc VLBI networks operate on an intermittent basis, and during observations the communication between elements to verify proper operation is limited. Small amounts of data from strong sources can be transmitted from the antennas to the correlator over telephone lines, and cross correlated to determine the instrumental delays and to check that the equipment is working properly. Later, arrays dedicated to VLBI were brought into operation [see, e.g., Napier et al. (1994)].

In VLBI one has less control over the system stability because independent frequency standards are used at each element. Frequency offsets in the standards can cause instrumental timing errors. These errors usually include an epoch error of a few microseconds and a drift of a few tenths of a microsecond per day (Section 9.5). Therefore, the correlation function of the received signals [with respect to time offset, τ, as defined in Eq. (3.27)] must be measured to determine and track the instrumental delay. In contrast, delay errors in connected-element interferometers, due mainly to baseline errors and atmospheric propagation delays, are usually less than 30 ps, corresponding to 1 cm of path length. These errors are negligible for bandwidths less than 1 GHz. Thus, the response in connected-element, delay-tracking interferometers is always centered on the white light fringe. Delay becomes important only when the field of view becomes too large for the bandwidth (see Sections 2.2 and 6.3) or when spectral line measurements are made by introducing time offsets. In VLBI it is necessary to search a range of delay values to find the correct time relationship that maximizes the correlation. Correlations for a number of delay offsets are usually formed simultaneously, so a VLBI correlator may resemble a digital spectral correlator, although the number of frequency channels may be less than generally used for spectral line observations. The frequency offsets in the standards, which cause drifts with time in the instrumental delay, also introduce offsets in the fringe frequency. Thus analysis of a VLBI experiment must begin with a two-dimensional search in delay and fringe frequency (delay rate) to find the peak of the correlation function. This process is referred to as fringe fitting.

The concept of coherence has different implications in VLBI and connected-element interferometry. In connected-element interferometry there is generally a suitable calibration source within a few degrees of the source of interest that can be observed every few minutes. Even if the instrumental phase drifts, there is no fundamental limit on integration time, and the concept of coherence time is replaced by that of the interval between calibrations. In VLBI, the short-term phase stability ($t < 10^3$ s) is worse. Atmospheric fluctuations above the stations are generally completely uncorrelated, and the frequency standards and frequency multipliers introduce phase noise in the fringes. Furthermore, a fundamental difference between connected-element interferometry and VLBI comes from the fact

that there are many fewer sources that are unresolved at VLBI spacings and can be used as calibrators. It is not always possible to find a calibrator close enough to the source under investigation to use as a phase reference. The time required to repoint the antennas and the decorrelation introduced by the atmosphere both increase with angular spacing. Thus, VLBI is subject to a fundamental coherence time that limits its sensitivity. For integration beyond the coherence time, it is necessary to average the fringe *amplitudes*, for which sensitivity improves only as the fourth root of the integration time (Section 9.3 under *Coherent and Incoherent Averaging*). It is also more difficult to calibrate phase in VLBI systems, although the situation improved during the late 1990s as enhanced sensitivity increased the number of sources that can be used as calibrators. Improved instrumental phase stability and more accurate modeling of the baselines, atmosphere, and similar factors have allowed the phase to be related to that of a calibrator several degrees away. Phase referencing in this manner is discussed in Section 12.2, and an example is shown in Fig. 12.2. Phase information can also be used in phase closure analysis. In measuring positions, fringe frequency and group delay (the delay pattern effect discussed in Sections 2.2 and 6.3) have also proved useful as measurement quantities.

Storage of the undetected signals before correlation presents VLBI with several problems. The average IF bandwidth is limited by the recording medium, which therefore limits the sensitivity of VLBI. The data must be stored as efficiently as possible, which requires a coarsely quantized representation of the signal, sampled at the Nyquist rate. With such a representation the basic operations of fringe rotation and delay tracking, when performed on the recorded data, introduce significant effects that must be allowed for in deriving the visibility (Section 9.7).

9.3 BASIC PERFORMANCE OF A VLBI SYSTEM

Time and Frequency Errors

A block diagram of a basic VLBI system and a possible processor configuration is shown in Fig. 9.1. The atomic frequency standards control the phases of the local oscillators and the sampling time for the tape recorders. In many VLBI applications, such as spectral line observations or astrometric programs, frequency-dependent effects must be accounted for precisely. To obtain a spectral analysis of the system, we consider the phase shifts encountered by a single frequency component. The signals received from a plane wave are $e^{j2\pi\nu t}$ at antenna 1, which we designate as the time-reference antenna, and $e^{j2\pi\nu(t-\tau_g)}$ at antenna 2, where τ_g is the geometric delay. The local oscillators have phases $2\pi\nu_{LO}t + \theta_1$ and $2\pi\nu_{LO}t + \theta_2$, where ν_{LO} is the local oscillator frequency, and θ_1 and θ_2 are the slowly varying terms that represent the phase noise due to the frequency standards. To start, we consider the upper-sideband response in Fig. 9.1, for which the local oscillator frequency is below the signal frequency. Thus, the phases after mixing are

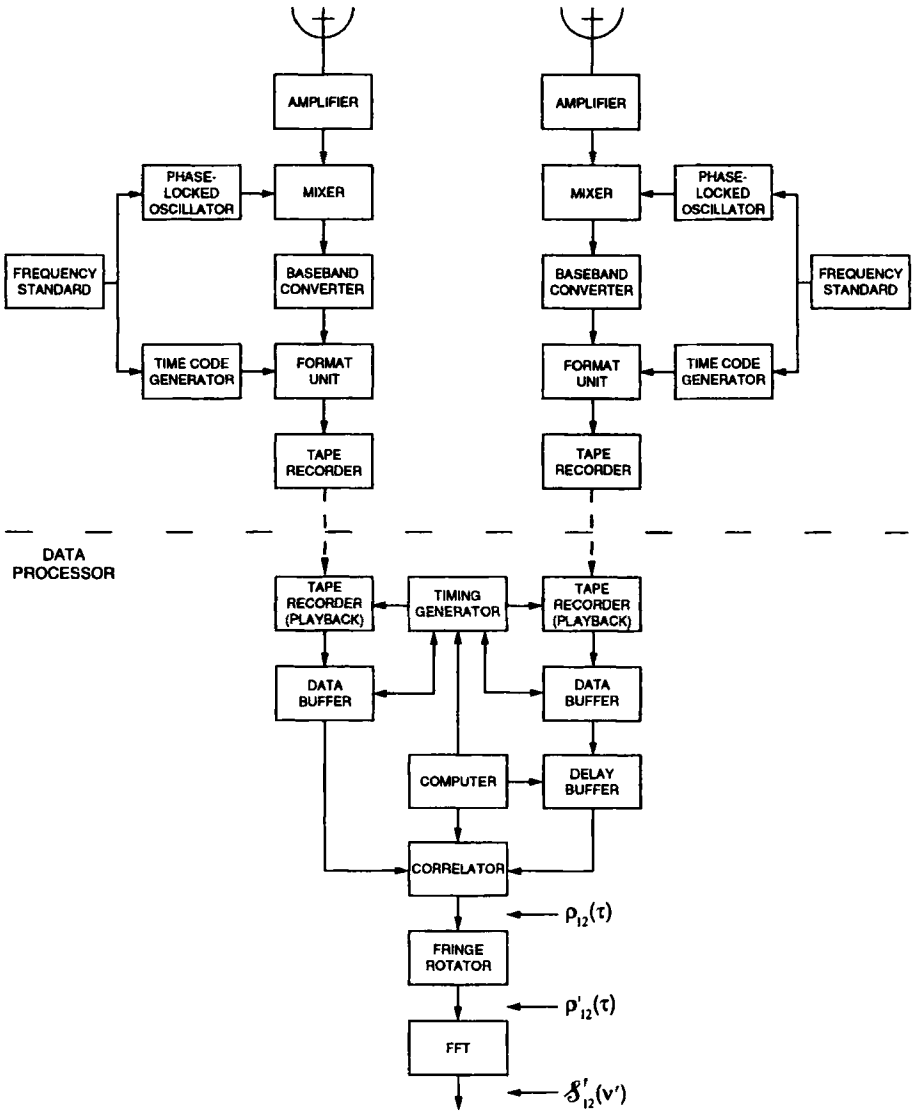


Figure 9.1 Block diagram of the essential elements of a VLBI system including data acquisition and processing. The system may pass the upper, lower, or both sidebands at the mixer inputs, depending on the passband of the amplifiers. For millimeter-wavelength observations the receiver input is often an SIS mixer, in which case both sidebands may be accepted. Quantization and sampling of the signals occur in the format units. The processor system shown illustrates the configuration described analytically by Eqs. (9.16)–(9.21). Major variations in the processing system relate to the position of the fringe rotator, which can also be located before the correlator (see Fig. 9.17).

$$\begin{aligned}\phi_1^{(1)} &= 2\pi(\nu - \nu_{\rm LO})t - \theta_1,\\ \phi_2^{(1)} &= 2\pi(\nu - \nu_{\rm LO})t - 2\pi\nu\tau_g - \theta_2.\end{aligned} \tag{9.8}$$

The recorded signals each have clock errors τ_1 and τ_2, so the phases of the recorded signals are

$$\begin{aligned}\phi_1^{(2)} &= 2\pi(\nu - \nu_{\rm LO})(t - \tau_1) - \theta_1,\\ \phi_2^{(2)} &= 2\pi(\nu - \nu_{\rm LO})(t - \tau_2) - 2\pi\nu\tau_g - \theta_2.\end{aligned} \tag{9.9}$$

During processing, the time series of signal samples from antenna 2 is advanced by τ_g', the estimate of τ_g, so

$$\phi_2^{(3)} = 2\pi(\nu - \nu_{\rm LO})(t - \tau_2 + \tau_g') - 2\pi\nu\tau_g - \theta_2. \tag{9.10}$$

The output of the multidelay correlator and Fourier transform processor is the cross power spectrum. The phase at the output of the processor for the signal component at frequency ν is

$$\begin{aligned}\phi_{12} &= \phi_1^{(2)} - \phi_2^{(3)}\\ &= 2\pi(\nu - \nu_{\rm LO})(\tau_2 - \tau_1) + 2\pi(\nu\Delta\tau_g + \nu_{\rm LO}\tau_g') + \theta_{21}\\ &= 2\pi(\nu - \nu_{\rm LO})(\tau_e + \Delta\tau_g) + 2\pi\nu_{\rm LO}\tau_g + \theta_{21},\end{aligned} \tag{9.11}$$

where $\Delta\tau_g = \tau_g - \tau_g'$ is the delay error, $\tau_e = \tau_2 - \tau_1$ is the clock error, and $\theta_{21} = \theta_2 - \theta_1$. Equation (9.11) applies to the upper-sideband frequency conversion in the mixers in Fig. 9.1, for which the intermediate frequency (IF) $(\nu - \nu_{\rm LO})$ is positive. For generality we also give the lower-sideband response, for which the IF is $(\nu_{\rm LO} - \nu)$. For the lower sideband

$$\phi_{12} = 2\pi(\nu_{\rm LO} - \nu)(\tau_e + \Delta\tau_g) - 2\pi\nu_{\rm LO}\tau_g - \theta_{21}. \tag{9.12}$$

Note that in the ideal case where $\tau_1 = \tau_2$, $\theta_1 = \theta_2$, and $\tau_g = \tau_g'$, Eqs. (9.11) and (9.12) reduce to $\phi_{12} = 2\pi\nu_{\rm LO}\tau_g$ for the upper sideband, and $\phi_{12} = -2\pi\nu_{\rm LO}\tau_g$ for the lower sideband.

The correlation function at the correlator output is real, but not even; thus, the cross power spectrum $\mathcal{S}_{12}$ for a source of continuum radiation has the property

$$\mathcal{S}_{12}(\nu') = \mathcal{S}_{12}^*(-\nu'), \tag{9.13}$$

where ν' is the intermediate frequency $(\nu - \nu_{\rm LO})$. We assume that the filters in the electronics have identical responses and therefore do not introduce any net phase shifts. The power response function of the instrumental filters is therefore real, and in terms of the voltage response, $H(\nu)$, of the filters for the two antennas, $\mathcal{S}(\nu') = H_1(\nu')H_2^*(\nu')$. By combining the phase from Eq. (9.11) and the magnitude of the power response, the cross power spectrum for the upper sideband can be written

$$\mathcal{S}_{12}(\nu') = \mathcal{S}(\nu')\exp\left\{j\left[2\pi\nu'(\tau_e + \Delta\tau_g) + 2\pi\nu_{\rm LO}\tau_g + \theta_{21}\right]\right\}. \tag{9.14}$$

The corresponding equation for the lower sideband can be obtained from Eq. (9.12). For the upper sideband the cross-correlation function can be calculated from Eqs. (9.13) and (9.14) as

$$\rho_{12}(\tau) = \int_{-\infty}^{\infty} \mathcal{S}_{12}(\nu') e^{j2\pi\nu'\tau}\, d\nu'. \tag{9.15}$$

For either sideband, integration includes both positive and negative frequencies, and since $\mathcal{S}_{12}$ is hermitian and $\mathcal{S}$ is purely real, we obtain

$$\rho_{12}(\tau) = 2F_1(\tau')\cos(2\pi\nu_{\mathrm{LO}}\tau_g + \theta_{21}) - 2F_2(\tau')\sin(2\pi\nu_{\mathrm{LO}}\tau_g + \theta_{21}), \tag{9.16}$$

where $\tau' = \tau + \tau_e + \Delta\tau_g$ and

$$\begin{aligned} F_1(\tau) &= \int_0^{\infty} \mathcal{S}(\nu')\cos(2\pi\nu'\tau)\, d\nu', \\ F_2(\tau) &= \int_0^{\infty} \mathcal{S}(\nu')\sin(2\pi\nu'\tau)\, d\nu'. \end{aligned} \tag{9.17}$$

If $\mathcal{S}(\nu')$ is a rectangular lowpass spectrum with bandwidth $\Delta\nu$, then

$$\begin{aligned} F_1(\tau) &= \Delta\nu \frac{\sin 2\pi\Delta\nu\tau}{2\pi\Delta\nu\tau}, \\ F_2(\tau) &= \Delta\nu \frac{\sin^2 \pi\Delta\nu\tau}{\pi\Delta\nu\tau}. \end{aligned} \tag{9.18}$$

These functions are shown in Fig. 9.2. By substituting Eq. (9.18) into Eq. (9.16), the cross-correlation function can be written

$$\rho_{12}(\tau) = 2\Delta\nu\cos(2\pi\nu_{\mathrm{LO}}\tau_g + \theta_{21} + \pi\Delta\nu\tau')\frac{\sin \pi\Delta\nu\tau'}{\pi\Delta\nu\tau'}. \tag{9.19}$$

A similar analysis is given by Rogers (1976).

The variation of τ_g with time results in fringe oscillations at the correlator output. The fringe frequency, $(1/2\pi)d\phi_{12}/dt$, is constant across the receiver bandwidth because the (instrumental) delay tracking removes the (geometric) delay-induced phase variation across the band. For the upper and lower sidebands, the rate of change of phase has opposite signs; note the term $2\pi\nu_{\mathrm{LO}}\tau_g$ in Eqs. (9.11) and (9.12). See also Fig. 6.5 and the related discussion. In VLBI the natural fringe frequency is fast enough that the fringes would be lost in the final averaging of the correlated data, so rotation of the phase to stop the fringes is applied at the correlator output in Fig. 9.1. In a double-sideband system, if the fringes are stopped for one sideband, the fringe frequency is doubled for the other sideband. However, it is possible to obtain the data from each sideband by processing the data twice with appropriate fringe offsets each time. In VLBI the source position and other

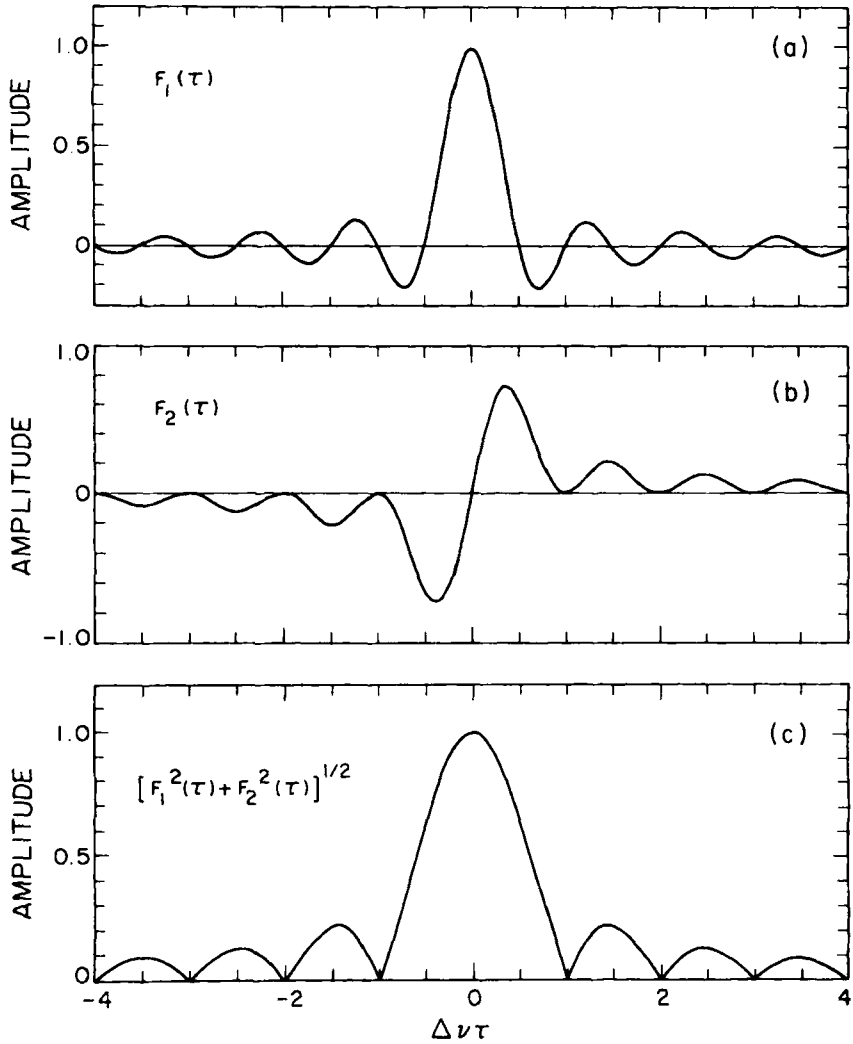


Figure 9.2 Functions $F_1(\tau)$ and $F_2(\tau)$, defined in Eq. (9.18), and the quantity $\sqrt{F_1^2(\tau) + F_2^2(\tau)}$.

parameters are not always known with sufficient accuracy when the observation is made, so in Fig. 9.1 the fringes are stopped after playback of the tapes to permit trial of different fringe rotation rates. This involves applying a phase shift to the quantized signals at the correlator input or output (see Section 9.7 under *Fringe Rotation Loss*). The effect on the cross-correlation function or the cross power spectrum can be described as multiplication by $e^{-j2\pi\nu_{LO}\tau_g'}$ for the upper sideband and filtering to select the low-frequency term. This process results in a complex correlation function:

$$\rho'_{12}(\tau) = \Delta\nu \exp\left[j\left(2\pi\nu_{\mathrm{LO}}\Delta\tau_g + \theta_{21} + \pi\Delta\nu\tau'\right)\right]\frac{\sin\pi\Delta\nu\tau'}{\pi\Delta\nu\tau'}. \tag{9.20}$$

Note that the principal fringe term, $2\pi\nu_{\mathrm{LO}}\tau_g$, has been eliminated, but residual fringes can result from terms in $\Delta\tau_g$ and $\Delta\nu$. The resulting cross power spectrum is

$$\mathcal{S}'_{12}(\nu') = \mathcal{S}(\nu')\exp\left\{j\left[2\pi\nu'(\tau_e + \Delta\tau_g) + 2\pi\nu_{\mathrm{LO}}\Delta\tau_g + \theta_{21}\right]\right\}. \tag{9.21}$$

This applies to the upper sideband, for which the fringes have been stopped, and the correlator output for the other sideband averages to zero.

An example of $\rho'_{12}(\tau)$ for eight values of τ is shown in Fig. 9.3. The waveforms represent the correlator output as a function of time for eight different delay offsets (lags) that differ sequentially by one Nyquist sample interval. Note that there is a phase shift of $\pi/2$ between adjacent delay steps. The fringe phase can be recovered by a proper interpolation (see Section 9.7 under *Discrete Delay Step Loss*) to the peak of the correlation function or from the phase of the cross power

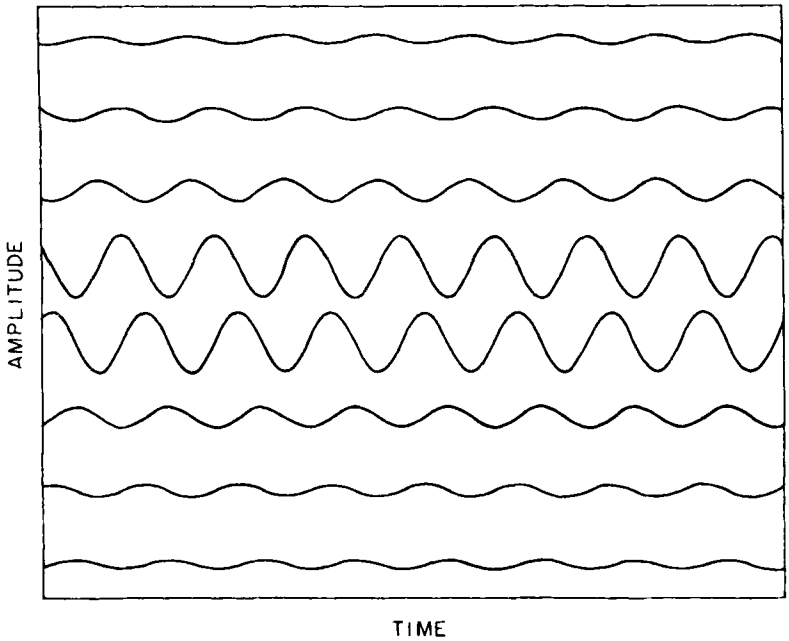


Figure 9.3 Each sinusoid represents the correlation function [the real part of Eq. (9.20)] versus time for a particular delay offset (from the top: $\frac{7}{2}$, $\frac{5}{2}$, $\frac{3}{2}$, $\frac{1}{2}$, $-\frac{1}{2}$, $-\frac{3}{2}$, $-\frac{5}{2}$, $-\frac{7}{2}$ times the Nyquist interval). The oscillations result from the residual fringe frequency, which includes any offsets in the frequency standards at the two antennas. Note the progressive phase shift of 90° between values of the correlation function at successive delay offsets.

spectrum at $\nu' = 0$. The group delay can be derived from the position of the correlation peak or the slope of the phase of the cross power spectrum. Note that the measured delay is $(1/2\pi)d\phi_{12}/d\nu$ and is therefore a group delay, not a phase delay.

The actual local oscillator frequencies may differ from the nominal value ν_{LO} due to an intentional offset from the nominal frequency or due to an offset error in the frequency standard. We can expand the phase terms θ_1 and θ_2 to include these frequency offsets, $\Delta\nu_1$ and $\Delta\nu_2$, and zero-mean phase components, θ_1' and θ_2':

$$\begin{aligned} \theta_1 &= 2\pi\Delta\nu_1 t + \theta_1', \\ \theta_2 &= 2\pi\Delta\nu_2 t + \theta_2'. \end{aligned} \tag{9.22}$$

Thus the fringe phase from Eq. (9.21) becomes

$$\phi_{12}(\nu') = 2\pi\left[\nu'(\tau_e + \Delta\tau_g) + \nu_{LO}\Delta\tau_g + \Delta\nu_{LO}t\right] + \theta_{21}', \tag{9.23}$$

where $\Delta\nu_{LO} = \Delta\nu_2 - \Delta\nu_1$, the difference in the local oscillator frequencies, and $\theta_{21}' = \theta_2' - \theta_1'$. The fringe frequency $(1/2\pi)d\phi_{12}/dt$ contains this local oscillator difference term. If $\Delta\nu_1$ is due to an offset in a frequency standard and is not zero, the measured fringe phase is actually more complicated than shown in Eq. (9.23). The clock error changes with time because of the frequency standard offset and is

$$\tau_1 = (\tau_1)_{t=0} + \frac{\Delta\nu_1}{\nu_{LO}}t. \tag{9.24}$$

The recovered time in the processor, based on the time of station 1, is related to the "true" time t by

$$t_1 = (\tau_1)_{t=0} + \left(1 + \frac{\Delta\nu_1}{\nu_{LO}}\right)t, \tag{9.25}$$

so that there is a slight shift in all measured frequencies and phases. Thus there is a fundamental asymmetry in the processing between the reference station from which time is derived and the other stations (Whitney et al. 1976).

For spectral line observations the quantity $\mathcal{S}(\nu')$ in Eq. (9.21) is the (temporal frequency) spectrum of the visibility of the source multiplied by the bandpass response of the interferometer. The bandpass response can be obtained by observation of the cross power spectrum of a continuum source with a flat spectrum. Alternatively, if the phase responses of the interferometer elements are identical, the bandpass response can be obtained from the geometric mean of the power spectra from the individual elements. These power spectra are obtained by observing a continuum source or blank sky, and measuring the autocorrelation of the waveform from each individual antenna. The frequency spectrum of the normalized visibility can be obtained by dividing the visibility spectrum by the geometric mean of the power spectra of the source as measured with each antenna.

Details of calibration procedures in VLBI spectral line observations are given by Moran (1973), Reid et al. (1980), Moran and Dhawan (1995), and Reid (1995, 1999).

Retarded Baselines

The estimate of delay τ_g must be accurate enough to ensure that the signal is within the delay and fringe-frequency ranges of the processor. The simplest approximation is

$$\tau_g = \mathbf{D} \cdot \frac{\mathbf{s}_0}{c}, \tag{9.26}$$

where $\mathbf{D} = \mathbf{r}_1 - \mathbf{r}_2$, $\mathbf{r}_1$ and $\mathbf{r}_2$ are vectors from the center of the earth to each station, and $\mathbf{s}_0$ is the unit vector to the center of the field. Account must be taken of the fact that the earth moves in the time between the arrival of a wave crest at one station and at another, since the earth is not an inertial reference. Therefore, in calculating the delay we should use not the instantaneous baseline, but the "retarded" baseline (Cohen and Shaffer 1971). A plane wave reaches the first station at time t_1 and the second station at a time t_2, which satisfies the equation

$$\mathbf{k} \cdot \mathbf{r}_1(t_1) - 2\pi\nu t_1 = \mathbf{k} \cdot \mathbf{r}_2(t_2) - 2\pi\nu t_2, \tag{9.27}$$

where $\mathbf{k} = (2\pi/\lambda)\mathbf{s}_0$. Now $t_2 - t_1 = \tau_g$, so

$$2\pi\nu\tau_g = \mathbf{k} \cdot [\mathbf{r}_2(t_1 + \tau_g) - \mathbf{r}_1(t_1)]. \tag{9.28}$$

Expansion of $\mathbf{r}_2$ in a Taylor series gives

$$\mathbf{r}_2(t_1 + \tau_g) \simeq \mathbf{r}_2(t_1) + \dot{\mathbf{r}}_2(t_1)\tau_g + \cdots \tag{9.29}$$

and

$$2\pi\nu\tau_g \simeq \mathbf{k} \cdot \left[\mathbf{D}(t_1) + \dot{\mathbf{r}}_2(t_1)\tau_g\right]. \tag{9.30}$$

Solving for τ_g yields

$$\tau_g = \frac{\mathbf{s}_0 \cdot \mathbf{D}}{c}\left[1 - \frac{\mathbf{s}_0 \cdot \dot{\mathbf{r}}_2}{c}\right]^{-1}, \tag{9.31}$$

where all quantities are evaluated at t_1. Since $\dot{\mathbf{r}} = \boldsymbol{\omega}_e \times \mathbf{r}$, where $\boldsymbol{\omega}_e$ is the angular velocity vector of the earth and $\times$ indicates the vector cross product, we can rewrite Eq. (9.31) as

$$\tau_g \simeq \frac{\mathbf{s}_0 \cdot \mathbf{D}}{c}\left[1 - \frac{\mathbf{s}_0 \cdot (\boldsymbol{\omega}_e \times \mathbf{r}_2)}{c}\right]^{-1}, \tag{9.32}$$

or

$$\tau_g \simeq \tau_{g0}(1 + \Delta), \tag{9.33}$$

where $1 + \Delta$ is the term in brackets on the right-hand side of Eq. (9.32). From the w term in Eq. (4.3),

$$\tau_{g0} = \frac{D}{c}\left[\sin d \sin\delta + \cos d \cos\delta \cos(H - h)\right]. \tag{9.34}$$

Here (H, δ) and (h, d) are the hour angle and declination coordinates of the source and baseline, respectively, the hour angles usually being specified with respect to the Greenwich meridian in VLBI practice. Also, we have

$$\Delta = \frac{\omega_e r_2}{c} \cos\mathcal{L}_2 \cos\delta \sin(h_2 - H), \tag{9.35}$$

where $\mathcal{L}_2$, h_2, and r_2 are the latitude, hour angle, and magnitude of $\mathbf{r}_2$. The function Δ has a maximum value of 1.5×10^{-6}, and τ_g can differ from τ_{g0} by a maximum of about 0.05 μs. Note that the appropriate coordinates in Eq. (9.34) are those that are uncorrected for refraction or diurnal aberration. An equivalent way of accounting for the retarded baseline is to use Eq. (9.26) for the delay but correct h and δ for the diurnal aberration at the remote site. The concept of retarded baselines does not apply if a heliocentric reference frame is used.

There are different ways to formulate VLBI observables. One system that may be described as station-oriented is to refer the measurements to the center of the earth, so that if tapes from two antennas are processed once and then interchanged and reprocessed, the phase obtained on the second pass will be the negative of that obtained on the first pass. This method presupposes an earth model, since the radius vectors must be known. For applications to astrometry or geodesy, a baseline-oriented system is usually preferred, in which the observables have no dependence on a priori values of earth parameters. A more precise discussion of VLBI observables can be found in Shapiro (1976) and Cannon (1978). For a full barycentric formulation, see Sovers, Fanselow, and Jacobs (1998).

Noise in VLBI Observations

We begin the discussion of noise by reviewing the statistical properties of fringe amplitude and phase, which were introduced in Section 6.2 [see also Moran (1976)]. The measured visibility is represented by a vector $\mathbf{Z} = \mathcal{V} + \boldsymbol{\varepsilon}$, where $\mathcal{V}$ and $\boldsymbol{\varepsilon}$ represent the true visibility (the signal) and noise components, respectively. We then select coordinates with x (real) and y (imaginary) so that $\mathcal{V}$ lies along the x axis, as shown in Fig. 6.8. The phase of the measured visibility resulting from the noise is a random variable denoted by ϕ. The components of $\boldsymbol{\varepsilon}$ have independent zero-mean Gaussian probability distributions in the x and y coordinates with

an rms deviation σ given by Eq. (6.50). In polar coordinates the amplitude of $\boldsymbol{\varepsilon}$ has a Rayleigh probability distribution, and the phase of $\boldsymbol{\varepsilon}$ has a uniform probability distribution [see, e.g., Papoulis (1965)]. $\mathbf{Z}$ is therefore a random variable whose x and y components, Z_x and Z_y, have a probability distribution given by

$$p(Z_x, Z_y) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(Z_x - |\mathcal{V}|)^2 + Z_y^2}{2\sigma^2}\right]. \tag{9.36}$$

It is often necessary to deal with the amplitude and phase of the visibility, denoted by Z and ϕ, respectively, whose probability distributions [Eqs. (6.63a) and (6.63b)] are

$$p(Z) = \frac{Z}{\sigma^2} \exp\left(-\frac{Z^2 + |\mathcal{V}|^2}{2\sigma^2}\right) I_0\left(f\frac{Z|\mathcal{V}|}{\sigma^2}\right), Z > 0 \tag{9.37}$$

where $Z = \sqrt{Z_x^2 + Z_y^2}$, and

$$\begin{aligned} p(\phi) = \frac{1}{2\pi} \exp\left(-\frac{|\mathcal{V}|^2}{2\sigma^2}\right) \Bigg\{1 + \sqrt{\frac{\pi}{2}} \frac{|\mathcal{V}| \cos\phi}{\sigma} \exp\left(\frac{|\mathcal{V}|^2 \cos^2\phi}{2\sigma^2}\right) \\ \times \left[1 + \operatorname{erf}\left(\frac{|\mathcal{V}| \cos\phi}{\sqrt{2}\sigma}\right)\right]\Bigg\}, \end{aligned} \tag{9.38}$$

where I_0 is the modified Bessel function of order zero, and erf is the error function. $p(Z)$ is known as the Rice distribution. Note that $\langle\phi\rangle = 0$, as expected, since the phase of $\mathcal{V}$ was set to zero. These probability distributions are plotted in Fig. 6.9. The expectations of Z, Z^2, and Z^4 are

$$\langle Z\rangle = \sqrt{\frac{\pi}{2}}\sigma \exp\left(\frac{-|\mathcal{V}|^2}{4\sigma^2}\right)\left[\left(1 + \frac{|\mathcal{V}|^2}{2\sigma^2}\right) I_0\left(\frac{|\mathcal{V}|^2}{4\sigma^2}\right) + \frac{|\mathcal{V}|^2}{2\sigma^2} I_1\left(\frac{|\mathcal{V}|^2}{4\sigma^2}\right)\right], \tag{9.39}$$

$$\langle Z^2\rangle = |\mathcal{V}|^2 + 2\sigma^2, \tag{9.40}$$

and

$$\langle Z^4\rangle = |\mathcal{V}|^4 + 8\sigma^2|\mathcal{V}|^2 + 8\sigma^4, \tag{9.41}$$

where I_1 is the modified Bessel function of order one. Higher even-order moments of Z can be readily calculated using the moment theorem for a Gaussian random distribution. When no signal is present, $I_0(0) = 1$, and the probability distributions of Z and ϕ are those of the noise, which are Rayleigh and uniform

distributions, respectively:

$$p(Z) = \frac{Z}{\sigma^2} \exp\left(-\frac{Z^2}{2\sigma^2}\right), \qquad Z > 0 \tag{9.42}$$

and

$$p(\phi) = \frac{1}{2\pi}, \quad 0 \le \phi < 2\pi. \tag{9.43}$$

For the no-signal case, $\langle Z \rangle = \sqrt{\pi/2}\sigma$, $\sigma_Z = \sqrt{\langle Z^2 \rangle - \langle Z \rangle^2} = \sigma\sqrt{2 - \pi/2}$, and $\sigma_\phi = \pi/\sqrt{3}$.

For the weak-signal case, defined as $|\mathcal{V}| \ll \sigma$, the probability distributions of Z and ϕ are

$$p(Z) \simeq \frac{Z}{\sigma^2} \exp\left(-\frac{Z^2}{2\sigma^2}\right)\left[1 - \frac{1}{2}\frac{|\mathcal{V}|^2}{\sigma^2} + \frac{1}{4}\left(\frac{Z|\mathcal{V}|}{\sigma^2}\right)^2\right] \tag{9.44}$$

and

$$p(\phi) \simeq \frac{1}{2\pi} + \frac{1}{\sqrt{8\pi}}\frac{|\mathcal{V}|}{\sigma}\cos\phi, \tag{9.45}$$

to first order in $|\mathcal{V}|/\sigma$. Thus,

$$\langle Z \rangle \simeq \sigma\sqrt{\frac{\pi}{2}}\left(1 + \frac{|\mathcal{V}|^2}{4\sigma^2}\right), \tag{9.46}$$

$$\sigma_Z \simeq \sigma\sqrt{2 - \frac{\pi}{2}}\left(1 + \frac{|\mathcal{V}|^2}{4\sigma^2}\right), \tag{9.47}$$

and

$$\sigma_\phi \simeq \frac{\pi}{\sqrt{3}}\left(1 - \sqrt{\frac{9}{2\pi^3}}\frac{|\mathcal{V}|}{\sigma}\right). \tag{9.48}$$

For the strong-signal case, $|\mathcal{V}| \gg \sigma$, the probability functions for Z and ϕ are approximately Gaussian distributions and are given by

$$p(Z) \simeq \frac{1}{\sqrt{2\pi}\sigma}\sqrt{\frac{Z}{|\mathcal{V}|}} \exp\left[-\frac{(Z - |\mathcal{V}|)^2}{2\sigma^2}\right] \tag{9.49}$$

and

$$p(\phi) \simeq \frac{1}{\sqrt{2\pi}}\frac{|\mathcal{V}|}{\sigma} \exp\left(-\frac{|\mathcal{V}|^2\phi^2}{2\sigma^2}\right). \tag{9.50}$$

For this case,

$$\langle Z \rangle \simeq |\mathcal{V}| \left(1 + \frac{\sigma^2}{2|\mathcal{V}|^2}\right), \tag{9.51}$$

$$\sigma_Z \simeq \sigma \left(1 - \frac{\sigma^2}{8|\mathcal{V}|^2}\right), \tag{9.52}$$

and

$$\sigma_\phi \simeq \frac{\sigma}{|\mathcal{V}|}. \tag{9.53}$$

Hence, in the strong-signal case, the statistics of Z are approximately Gaussian (see Fig. 6.9) and $\langle Z \rangle$ approaches $|\mathcal{V}|$. In this case, N samples of Z can be averaged and the signal-to-noise ratio improves with $\sqrt{N}$. In the weak-signal case the perturbation of the Rayleigh noise distribution by the signal is small and, as we shall discuss later in this section, it is difficult to improve the signal-to-noise ratio by averaging beyond the coherence time of the system.

Equations (9.46) and (9.51) show that $\langle Z \rangle$ is a biased estimate of $|\mathcal{V}|$. If only one measurement of Z is available, the most likely value of $|\mathcal{V}|$ is the one that maximizes $p(Z)$, given by Eq. (9.37). This maximum is closely approximated by the equation $Z_{\max} = \sqrt{|\mathcal{V}|^2 + \sigma^2}$, which is accurate to better than 8% for all values of $|\mathcal{V}|$ and to better than 1% for $|\mathcal{V}|/\sigma > 2$. Hence when one measurement of Z is available, the most likely value of $|\mathcal{V}|$ is approximately $\sqrt{Z^2 - \sigma^2}$.

Probability of Error in the Signal Search

When starting a new session of VLBI observations with an ad hoc array, the first task in the processing is to search for fringes. This is necessary because of the uncertainties in the station clocks and their drift rates, and means that the instrumental delay and fringe frequency must be found. This step is frequently unnecessary with a dedicated VLBI array, for which the values of fringe rate and delay are continuously updated from successive observations. A fringe search must be carried out on a large two-dimensional grid, as shown in Fig. 9.4. For example, consider an experiment where $\Delta\nu = 50$ MHz at an observing frequency of 10^{11} Hz. The delay increments are equal to the sampling interval of 0.01 μs. An instrumental delay uncertainty of ± 1 μs requires a search of 200 delay intervals. If the coherent integration time is 200 s and the frequency standards are only set to a fractional accuracy of 10^{-11}, then ± 1 Hz must be searched, which at an interval size of 0.005 Hz, is 400 discrete frequencies. The total number of cells to be searched is 80,000. If there is no signal present, then $p(Z)$ will be given by Eq. (9.42). The cumulative probability distribution (that is, the probability that Z is less than Z_0) in this case is the integral of Eq. (9.42) from zero to Z_0, or

$$P(Z_0) = 1 - \exp\left(-\frac{Z_0^2}{2\sigma^2}\right). \tag{9.54}$$

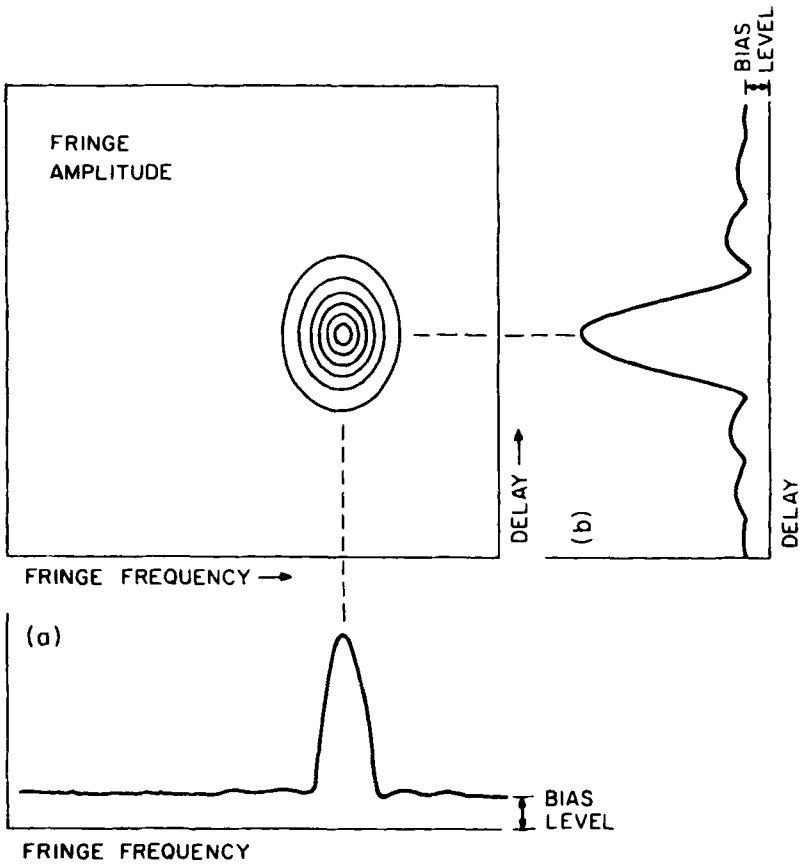


Figure 9.4 Fringe amplitude as a function of residual fringe frequency and delay. The one-dimensional plots are the peak fringe amplitude versus delay and fringe frequency. The probability distribution of the noise in these plots is given by Eq. (9.57) and the bias level by Eq. (9.58).

The cumulative probability distribution for the maximum of n independent samples $Z_m = \max\{Z_1, Z_2, \ldots, Z_n\}$ is

$$P(Z_m) = \left[1 - \exp\left(-\frac{Z_m^2}{2\sigma^2}\right)\right]^n. \tag{9.55}$$

Thus, the probability of one or more samples exceeding Z_m, which we call the probability of error, p_e, is

$$p_e = 1 - \left[1 - \exp\left(-\frac{Z_m^2}{2\sigma^2}\right)\right]^n. \tag{9.56}$$

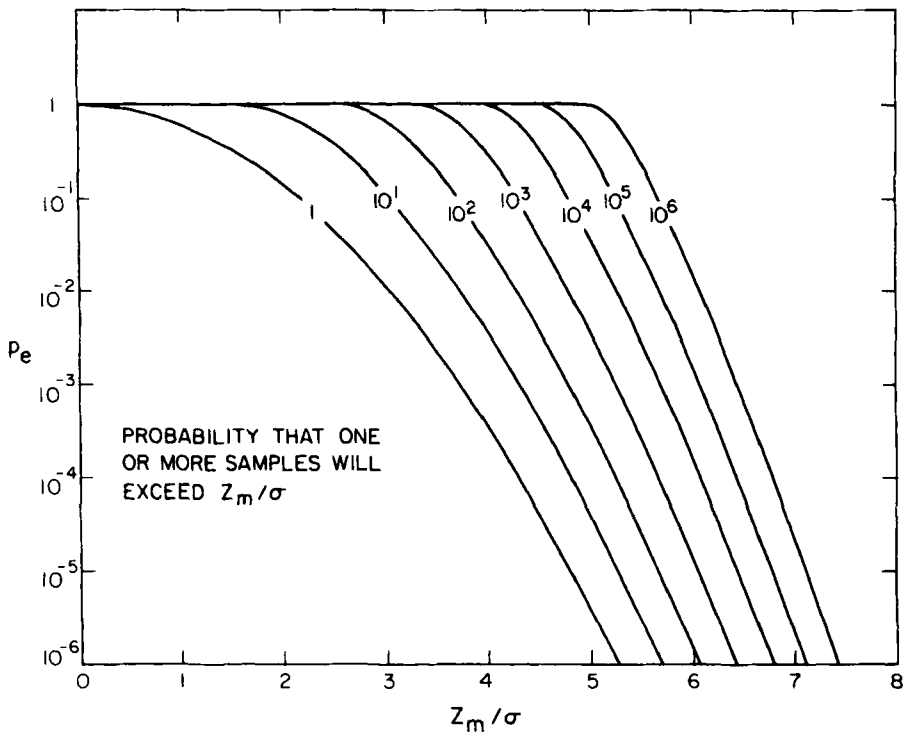


Figure 9.5 Probability that one or more samples of the fringe amplitude will exceed the value Z_m/σ in the absence of a signal, as given by Eq. (9.56). The curves are labeled by the number of samples measured.

This function is shown in Fig. 9.5. The probability distribution of Z_m is obtained by differentiating Eq. (9.55),

$$p(Z_m) = \frac{nZ_m}{\sigma^2} \exp\left(-\frac{Z_m^2}{2\sigma^2}\right)\left[1 - \exp\left(-\frac{Z_m^2}{2\sigma^2}\right)\right]^{n-1}. \tag{9.57}$$

For large n, this probability distribution is nearly Gaussian with mean value and standard deviation given by

$$\langle Z_m \rangle \simeq \sigma\sqrt{2\ln n}, \tag{9.58}$$

$$\sigma_m \simeq \frac{0.77\sigma}{\sqrt{\ln n}}. \tag{9.59}$$

Examples of $p(Z_m)$ for various values of n are shown in Fig. 9.6. It is frequently useful to reduce a two-dimensional function, such as the one shown in Fig. 9.4 of

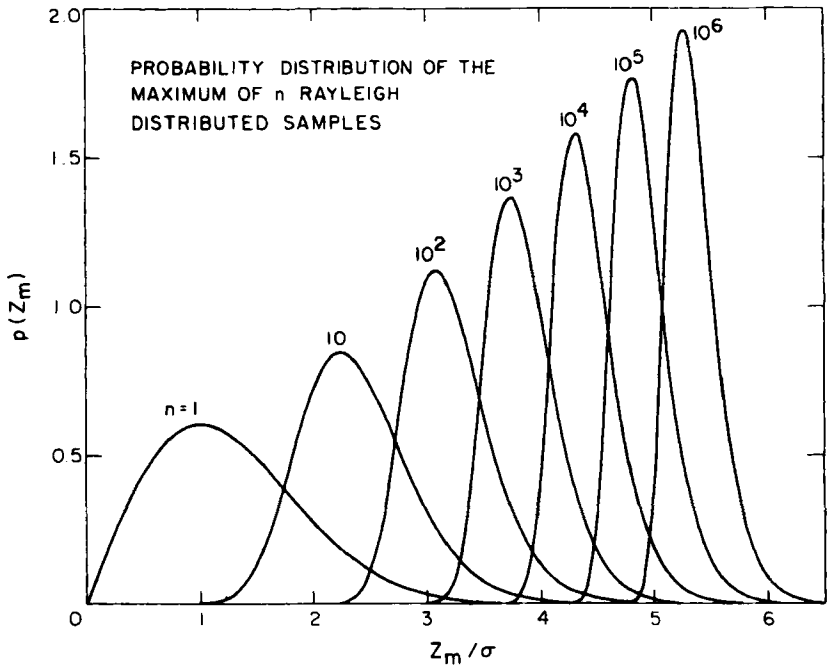


Figure 9.6 Probability distribution of the maximum of n random variables that have Rayleigh distributions, as given Eq. (9.57).

fringe amplitude versus fringe frequency and delay, to a one-dimensional function by searching for the maximum value of the function over one variable. This search process introduces a bias, equal to $\langle Z_m \rangle$, into the one-dimensional function. This bias increases with the number of samples and obscures weak signals.

We can also calculate the probability of misidentifying a signal. Suppose that we have measurements of fringe amplitude at two values of delay or fringe frequency with the signal present at one value. The probability that the amplitude in the channel with the signal (Z_1) is larger than the amplitude in the channel with only the noise (Z_2) is

$$p(Z_1 > Z_2) = \int_0^{\infty} p(Z_1) \left[\int_0^{Z_1} p(Z_2)\, dZ_2 \right] dZ_1. \tag{9.60}$$

$p(Z_1)$ is given by Eq. (9.37), and $p(Z_2)$ is given by Eq. (9.42). We can generalize this result for a search over n channels where the signal channel amplitude is Z_s. The probability that Z_s will exceed the values of Z in the other channels is, from Eqs. (9.54) and (9.60),

$$p(Z_s > Z_1, \ldots, Z_n) = \int_0^\infty p(Z)\left[1 - \exp\left(-\frac{Z^2}{2\sigma^2}\right)\right]^{n-1} dZ, \qquad (9.61)$$

where $p(Z)$ is given by Eq. (9.37). Thus, the probability of one or more samples exceeding the amplitude of the signal is

$$p_e' = 1 - \int_0^\infty p(Z)\left[1 - \exp\left(-\frac{Z^2}{2\sigma^2}\right)\right]^{n-1} dZ. \qquad (9.62)$$

p_e' is plotted in Fig. 9.7. For example, if the search is over 100 channels, a probability of misidentification of less than 0.1% requires $|\mathcal{V}|/\sigma > 6.5$.

Coherent and Incoherent Averaging

We wish to estimate the amplitude of a barely detectable signal. We examine a time series of correlator output values in which the phase, $\phi(t)$, represents the

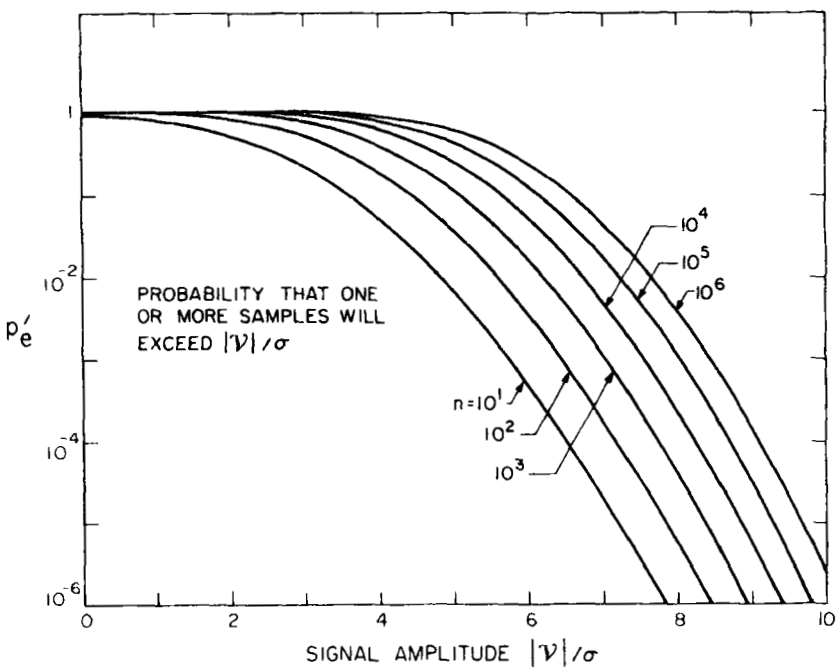


Figure 9.7 Probability that one or more samples of fringe amplitude among the samples with no signal will exceed the fringe amplitude of the sample with the signal, versus the signal amplitude, $|\mathcal{V}|$, as given in Eq. (9.62). The curves are labeled according to the total number of samples n. The asymptotic value of p_e' as $|\mathcal{V}|/\sigma$ goes to zero is $1 - 1/n$.

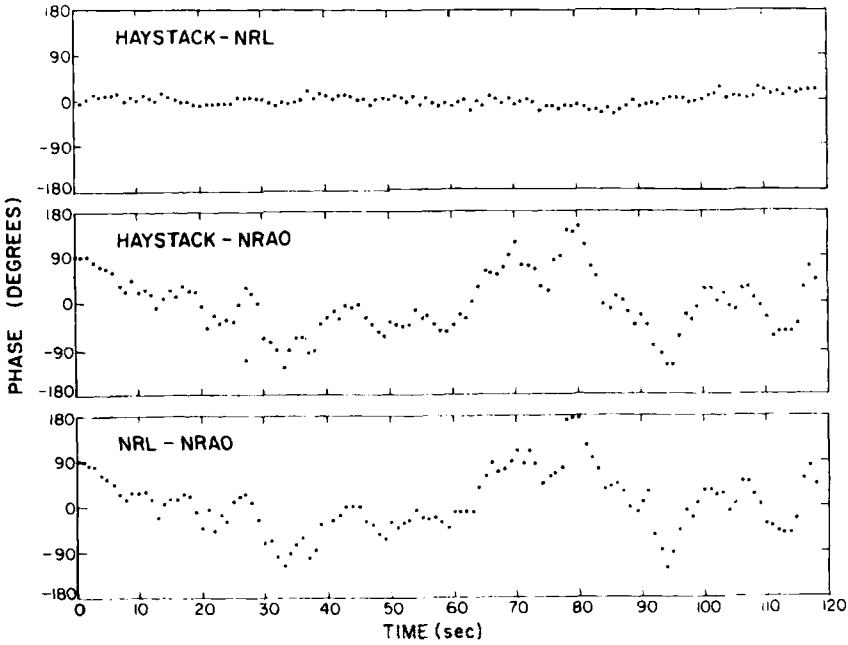


Figure 9.8 Fringe phase versus time from an observation of a strong source [the water vapor maser in W3 (OH)] on a three-baseline VLBI experiment at 22 GHz. Two of the stations, Haystack Observatory and the Naval Research Laboratory (Maryland Point Observatory), were equipped with hydrogen maser frequency standards, while the National Radio Astronomy Observatory used a rubidium vapor frequency standard. The phase noise in the top plot is dominated by contributions from the receivers and the atmosphere, while the phase noise in the bottom two plots is dominated by the phase noise in the rubidium frequency standard. These data were obtained in 1971 with the Mark I VLBI system.

effects of receiver noise, fluctuations in the frequency standards, or fluctuations in the atmospheric path. An example of phase versus time from a VLBI measurement is shown in Fig. 9.8. The correlator output is

$$r(t) = Z(t)e^{j\phi(t)}. \tag{9.63}$$

How do we estimate $|\mathcal{V}|$ when the time range of the data exceeds the coherence time? There are two useful procedures. First, note that $r(t)$ has a Fourier transform $R(\nu)$ for a time segment of length τ. Using Parseval's theorem,

$$\int_0^\tau r(t)r^*(t)\,dt = \int_{-\infty}^{\infty} R(\nu)R^*(\nu)\,d\nu, \tag{9.64}$$

and Eq. (9.40), we obtain an estimate $|\mathcal{V}|_e$ of the amplitude:

$$|\mathcal{V}|_e^2 = \frac{1}{\tau}\int_{-\infty}^{\infty} |R(\nu)|^d\nu - 2\sigma^2. \tag{9.65}$$

Equation (9.65) shows that one can take a time series of duration greater than τ_c and estimate the visibility amplitude from the fringe-frequency spectrum, $|R(\nu)|^2$. This procedure is a form of incoherent averaging because squares of spectral components in the fringe frequency domain are summed.

Another method of incoherent averaging involves the averaging of the squares of the time series samples. An unbiased estimate of the amplitude is

$$|\mathcal{V}|_e^2 = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i^2\right) - 2\sigma^2, \tag{9.66}$$

where

$$Z_i e^{j\phi_i} = \frac{1}{\tau_c}\int_{t_i}^{t_i+\tau_c} Z(t)e^{j\phi(t)}\,dt. \tag{9.67}$$

From Eqs. (9.40), (9.41), and (9.66) we have $\langle|\mathcal{V}|_e^2\rangle = |\mathcal{V}|^2$ and $\langle|\mathcal{V}|_e^4\rangle = |\mathcal{V}|^4 + 4\sigma^2(|\mathcal{V}|^2 + \sigma^2)/N$, so that the signal-to-noise ratio is

$$\mathcal{R}_{\rm sn} = \frac{\langle|\mathcal{V}|_e^2\rangle}{\sqrt{\langle|\mathcal{V}|_e^4\rangle - \langle|\mathcal{V}|_e^2\rangle^2}} = \frac{\sqrt{N}}{2\sigma^2}|\mathcal{V}|^2\frac{1}{\sqrt{1+|\mathcal{V}|^2/\sigma^2}}. \tag{9.68}$$

$|\mathcal{V}|/\sigma$ is equal to the signal-to-noise ratio at the output of a single-multiplier correlator, as given by Eqs. (6.48) and (6.49). For VLBI observations the quantization loss described in Section 8.3, η_Q, is replaced by the general loss factor η described in Section 9.7, and from Eq. (6.64) we obtain $|\mathcal{V}|/\sigma = (T_A\eta/T_S)\sqrt{2\Delta\nu\tau_c}$. Equation (9.68) then becomes

$$\mathcal{R}_{\rm sn} = \frac{T_A^2\eta^2}{T_S^2}\sqrt{\frac{\Delta\nu^2\tau\tau_c}{(1+2T_A^2\eta^2\Delta\nu\tau_c/T_S^2)}}, \tag{9.69}$$

where $\tau = N\tau_c$ is the total integrating time. The two limiting cases of Eq. (9.69) are

$$\mathcal{R}_{\rm sn} \simeq \frac{\eta}{\sqrt{2}}\frac{T_A}{T_S}\sqrt{\Delta\nu\tau}, \qquad T_A \gg \frac{T_S}{\sqrt{2\Delta\nu\tau_c}}, \tag{9.70}$$

$$\mathcal{R}_{\rm sn} \simeq \left(\frac{T_A\eta}{T_S}\right)^2\Delta\nu\sqrt{\tau\tau_c}, \qquad T_A \ll \frac{T_S}{\sqrt{2\Delta\nu\tau_c}}. \tag{9.71}$$

Note that in the strong-signal case incoherent averaging is not needed. When incoherent averaging is used, the coherent averaging time should be as long as possible without decreasing the fringe amplitude. If we assume that $\mathcal{R}_{\rm sn} = 4$ for detection, and recall that $\tau = N\tau_c$, then for the weak-signal case the minimum detectable antenna temperature can be found from Eq. (9.71) to be

$$(T_A)_{\rm min} = \frac{2T_S}{\eta N^{1/4}\sqrt{\Delta\nu\tau_c}}. \tag{9.72}$$

Thus, because of the $N^{1/4}$ dependence in Eq. (9.72), incoherent averaging is effective only if N is very large. If the coherence time is of the order of $1/\Delta\nu$, then the observing system reduces to a form of incoherent, or intensity, interferometer [see Section 16.1 and Clark (1968)]. For the weak-signal case, Eq. (9.71) then becomes

$$\mathcal{R}_{\rm sn} \simeq \left(\frac{T_A\eta}{T_S}\right)^2 \sqrt{\Delta\nu\tau}. \tag{9.73}$$

9.4 FRINGE FITTING FOR A MULTIELEMENT ARRAY

Global Fringe Fitting

In Section 9.3 we considered the problem of searching for fringes in the output from a single baseline. For VLBI, the basic requirement in fringe fitting is to determine the fringe phase (i.e., the phase of the visibility) and the rate of change of the fringe phase, with time and with frequency or delay. Fringe rate offsets result from errors in the positions of the source or antennas as well as antenna-related effects such as frequency offsets in local oscillators. Most of these can be specified as factors that relate to individual antennas, rather than to baselines. Because of this, data from all baselines can be used simultaneously to determine the fringe rate parameters. By simultaneously using all of the data from a multi-element VLBI array, it is possible to detect fringes that are too weak to be seen on a single baseline. This is particularly important for VLBI arrays with similar antennas and receivers; with an ad hoc array, a possible alternative is to use the data from the two most sensitive antennas to find the fringes and let this result constrain the solutions for other baselines.

A method of analysis that is based on simultaneous use of the complete data set from a multi-antenna observation was developed by Schwab and Cotton (1983) and is referred to as *global fringe fitting*. Let $Z_{mn}(t)$ be the correlator output, that is, the measured visibility, from the baseline for antennas m and n. The complex (voltage) gain for antenna n and the associated receiving system is $g_n(t_k, \nu_\ell)$, where t_k represents a (coherently) time-integrated sample of the correlator output for frequency channel ν_ℓ. Thus,

$$Z_{mn}(t_k, \nu_\ell) = g_m(t_k, \nu_\ell) g_n^*(t_k, \nu_\ell) \mathcal{V}_{mn}(t_k, \nu_\ell) + \epsilon_{mnk\ell}, \tag{9.74}$$

where $\mathcal{V}_{mn}$ is the true visibility for baseline mn and $\epsilon_{mnk\ell}$ represents the observational errors which result principally from noise. It should be remembered that the noise terms are present in all the measurements, but beyond this point they will usually be omitted from the equations. The gain terms can be written as

$$g_n(t_k, \nu_\ell) = |g_n| e^{j\psi_n(t_k, \nu_\ell)}. \tag{9.75}$$

To simplify the situation in Eq. (9.75), we assume that the gain terms and the amplitude of the source visibility are constant over the range of (t, ν) space covered

by the observation. To first order we can then write

$$
\begin{aligned}
Z_{mn}(t_k, \nu_\ell) = {} & |g_m||g_n||\mathcal{V}| \exp\left[j(\psi_m - \psi_n)(t_0, \nu_0)\right] \\
& \times \exp\left[j\left(\left.\frac{\partial(\psi_m - \psi_n + \phi_{mn})}{\partial t}\right|_{(t_0,\nu_0)} (t_k - t_0)\right.\right. \\
& \left.\left. + \left.\frac{\partial(\psi_m - \psi_n + \phi_{mn})}{\partial \nu}\right|_{(t_0,\nu_0)} (\nu_\ell - \nu_0)\right)\right],
\end{aligned}
\tag{9.76}
$$

where ϕ_{mn} is the phase of the (true) visibility $\mathcal{V}_{mn}$. The rates of change of the phase of the measured visibility with respect to time and frequency are the fringe rate

$$
r_{mn} = \left.\frac{\partial(\psi_m - \psi_n + \phi_{mn})}{\partial t}\right|_{(t_0,\nu_0)}, \tag{9.77}
$$

and the delay

$$
\tau_{mn} = \left.\frac{\partial(\psi_m - \psi_n + \phi_{mn})}{\partial \nu}\right|_{(t_0,\nu_0)}, \tag{9.78}
$$

for the baseline mn at time and frequency (t_0, ν_0). In terms of these quantities we can relate the measured visibility (correlator output) to the true visibility as follows:

$$
\begin{aligned}
Z_{mn}(t_k, \nu_\ell) = {} & |g_m||g_n|\mathcal{V}_{mn}(t_k, \nu_\ell) \exp\left\{j\left[(\psi_m - \psi_n)|_{t=t_0}\right.\right. \\
& \left.\left. + (r_m - r_n)(t_k - t_0) + (\tau_m - \tau_n)(\nu_\ell - \nu_0)\right]\right\}.
\end{aligned}
\tag{9.79}
$$

For each antenna there are four unknown parameters: the modulus of the gain, the phase of the gain, the fringe rate, and the delay. Since all of the data are in the form of relative phases of two antennas, it is necessary to designate one antenna as the reference. For this antenna the phase, fringe rate, and delay are usually taken to be zero, leaving $4n_a - 3$ parameters to be determined. However, it is possible to simplify further and consider only the phase terms in the fringe fitting. The amplitudes of the antenna gains are subsequently calibrated separately. The number of parameters to be determined is thereby reduced to $3(n_a - 1)$. Then to obtain the global fringe solution, the source visibility $\mathcal{V}_{mn}$ is represented by a model of the source, and a least-squares fit of the parameters in Eq. (9.79) to the visibility measurements is made. For details on a method for the least-squares solution, see Schwab and Cotton (1983). The source model, which is a "first guess" of the true structure, could in some cases be as simple as a point source.

Another method of using the data for several baselines simultaneously in fringe fitting is an extension of the method described earlier for single baselines. The measured visibility data are required to be specified in terms of fringe frequency and delay, which can be obtained, for example, by a time-to-frequency Fourier

transformation of the data from a lag correlator. Then for each antenna pair there is a matrix of values of the interferometer response at incremental steps in the delay and fringe rate. The maximum amplitude indicates the solution for delay and fringe rate for the corresponding baseline, as illustrated in Fig. 9.4. However, the method can be extended to include the responses from a number of baselines by using the closure phase principle, which is discussed in more detail in Section 10.3. Because we are considering fringe fitting in phase only, the measured data are represented by ϕ_{mn}. Since ψ_{mn}, the instrumental phase for baseline mn, is equal to the difference between the measured and true visibility phases, we can write

$$\psi_{mn} = \psi_m - \psi_n = \tilde{\phi}_{mn} - \phi_{mn}, \tag{9.80}$$

where the ψ terms represent the instrumental phases, the ϕ terms represent the visibility phases, and the tilde ($\tilde{}$) indicates measured visibility phases. Now consider including a third antenna, designated p. For this combination we can write

$$\psi_{mpn} = \psi_{mp} + \psi_{pn} = (\psi_m - \psi_p) + (\psi_p - \psi_n) = \psi_m - \psi_n. \tag{9.81}$$

Thus ψ_{mpn} provides another measured value of ψ_{mn}, equal to

$$\psi_{mp} + \psi_{pn} = (\tilde{\phi}_{mp} - \phi_{mp}) + (\tilde{\phi}_{pn} - \phi_{pn}). \tag{9.82}$$

Similarly, for four antennas

$$\psi_{mpqn} = \psi_m - \psi_n = (\tilde{\phi}_{mp} + \tilde{\phi}_{pq} + \tilde{\phi}_{qn}) - (\phi_{mp} + \phi_{pq} + \phi_{qn}). \tag{9.83}$$

Thus estimated values of ψ_{mn} can be obtained from the measurements from loops of antenna pairs starting with antenna m and ending with antenna n. Combinations of more than three baselines (four antennas) can be expressed as combinations of smaller numbers of antennas, and the noise in such larger combinations is not independent. Loops of three and four antennas provide additional information that contributes to the sensitivity and accuracy of the fringe fitting for antennas m and n. Note, however, that the model visibilities are also required.

Of the two techniques, the least-squares fitting is better with respect to uniform combination of the data, but it requires a good starting estimate if it is to converge efficiently. Schwab and Cotton (1983) used the second of the two methods to provide a starting point for the full least-squares solution. This procedure has subsequently become the basis of standard reduction programs for VLBI data (Walker 1989a,b).

Although global fringe fitting provides sensitivity superior to that of baseline-based fitting, in practice some experience is needed to determine when use of the global method is appropriate. If the source under study has complicated structure, with large variations in the visibility amplitude, it will probably not be well represented by the model visibility required in the global fitting method. In such a case it may be better to start with a smaller number of antennas in the fringe fitting or, if the source is sufficiently strong, to consider baselines separately. On

the other hand, if the source contains a strong unresolved component, it may be adequate to consider smaller groups of antennas separately and thus reduce the overall computing load.

Relative Performance of Fringe Detection Methods

In the regime where the phase noise limits the sensitivity, careful investigation of detection techniques is warranted. The most important of these have been examined by Rogers, Doeleman, and Moran (1995) to determine their relative performance. We assume in all cases that the visibility data from the correlator outputs have been averaged for a time equal to the coherence time, τ_c, discussed earlier. We have seen in Eq. (9.72) that incoherent averaging of N time segments of data reduces the level at which a signal is detectable by an amount proportional to $N^{-1/4}$. Rogers et al. show that for a detection threshold for which the probability of a false detection is <0.01% in a search of 10^6 values, the threshold of detection is lower than that without incoherent averaging (in effect, $N = 1$) by a factor $0.53N^{-1/4}$. This result is accurate only for large N, and they find empirically that for smaller N the detection threshold decreases in proportion to $N^{-0.36}$; that is, the improvement with increasing N is greater when N is small. Table 9.1 includes the improvement factor $0.53N^{-1/4}$, together with other results that are discussed below. The fourth column of Table 9.1 gives numerical examples of relative sensitivity for $N = 200$ time segments and $n_a = 10$ antennas. Note that for lines 1–5 of Table 9.1, the criterion for detection is a probability of error of less than 1% in a search of 10^6 values of delay and fringe rate for each of $n_a - 1$ elements of the array, the values for the reference antenna being taken to be zero. For line 6 the search spans only the two dimensions of right ascension and declination.

TABLE 9.1[a] Relative Thresholds for Various Detection[b] Methods

	Method	Threshold (Relative Flux Density)	
1	One baseline, coherent averaging	1	1
2	One baseline, incoherent averaging	$0.53N^{-1/4}$	0.14 ($N = 200$)
3	3-baseline triple product	$\left(\frac{4}{N}\right)^{1/6}$	0.52 ($N = 200$)
4	Array of n_a elements, coherent global search	$\left(\frac{2}{n_a}\right)^{1/2}$	0.45 ($n_a = 10$)
5	Global search with incoherent averaging	$0.53\left(\frac{4}{Nn_a^2}\right)^{1/4}$	0.05 ($N = 200$, $n_a = 10$)
6	Incoherent averaging over time segments and baselines	$0.53\left(\frac{2}{Nn_a(n_a-1)}\right)^{1/4}$	0.05 ($N = 200$, $n_a = 10$)

[a] From Rogers, Doeleman, and Moran (1995).
[b] See text for detection criterion.

Triple Product, or Bispectrum

Another form of the output of a multielement array that can be considered is the triple product, or bispectrum, which is the product of the complex outputs for three baselines that form a triangle. The triple product is given by the product of measured visibilities

$$P_3 = |Z_{12}||Z_{23}||Z_{31}|e^{j(\tilde{\phi}_{12}+\tilde{\phi}_{23}+\tilde{\phi}_{31})} = |Z_{12}||Z_{23}||Z_{31}|e^{j\phi_c}, \tag{9.84}$$

where ϕ_c represents the closure phase (Section 10.3), which is zero if the source is unresolved. We assume here that the amplitude of the measured visibility, Z, is calibrated separately, so that the moduli of the gain factors g_m and g_n in Eq. (9.74) are unity. Each of the measured visibility terms includes noise of power $2\sigma^2$, that is, the noise power in the output of a complex correlator. For the low-signal case, the noise determines the variance of the triple product, which is

$$\langle |P_3|^2 \rangle = \langle |Z_{12}|^2 |Z_{23}|^2 |Z_{31}|^2 \rangle = 8\sigma^6. \tag{9.85}$$

For a point source the signal is real and is equal to $\langle (\mathcal{R}e P_3)^2 \rangle = \langle |P_3|^2 \rangle / 2$, where $\mathcal{R}e$ indicates the real part. The ratio of this triple product signal term to the noise in the real output of the correlator is $\mathcal{V}^3/2\sigma^3$. Rogers, Doeleman, and Moran (1995) also give an expression for the signal-to-noise ratio that is not restricted to the small-signal case, and Kulkarni (1989) gives a general expression in a detailed analysis of the subject.

Now consider the incoherent average of N values of the triple product for three antennas, each of which represents an average of the correlator output over the coherence interval, τ_c. We represent this average of triple products by

$$\overline{P}_3 = \frac{1}{N} \sum_N |Z_{12}||Z_{23}||Z_{31}|e^{j\phi_c}. \tag{9.86}$$

If the signal amplitudes are equal, the expectation of the real part of $\overline{P}_3$ is

$$\langle \mathcal{R}e \overline{P}_3 \rangle = \mathcal{V}^3, \tag{9.87}$$

and the second moment of $\mathcal{R}e \overline{P}_3$ is

$$\langle (\mathcal{R}e \overline{P}_3)^2 \rangle = \frac{1}{N} \langle |P_3|^2 \rangle \langle \cos^2 \phi_c \rangle. \tag{9.88}$$

In the small-signal case, in which the value of $\langle |P_3|^2 \rangle$ results mainly from noise, the expectation of the second moment is, from Eq. (9.85), $4\sigma^6/N$. The signal-to-noise ratio is equal to the expectation of $\overline{P}_3$ divided by the square root of the expectation of the second moment

$$\mathcal{R}_{sn} = \frac{\sqrt{N}\mathcal{V}^3}{2\sigma^3}. \tag{9.89}$$

from which

$$\mathcal{V} = (2\mathcal{R}_{sn})^{1/3}\sigma N^{-1/6}. \tag{9.90}$$

Line 3 of Table 9.1 gives the signal strength for a value of $\mathcal{R}_{sn}$ that allows detection at a level corresponding to the specified error criterion.

Fringe Searching with a Multielement Array

With an array of n_a VLBI antennas the amount of information gathered in a given time is greater than that with a single antenna pair by a factor $n_a(n_a - 1)/2$. One might thus expect that the array would offer an increase in sensitivity $\simeq [n_a(n_a - 1)/2]^{1/2}$. However, the larger number of antennas also introduces a very large increase in the parameter space to be searched. Thus, the probability of encountering high-noise amplitudes within this parameter space is correspondingly greater. It is therefore necessary to increase the signal level used as a detection threshold in order to avoid increasing the probability of false detection.

Consider a two-element array in which the number of data points to be searched in the parameter space (frequency × delay) is n_d. If a third antenna is then introduced, and correlation is measured for all baselines, the number of data points to be searched becomes n_d^2. For n_a antennas, it becomes $n_d^{(n_a-1)}$. The probability distribution of the maximum of n Rayleigh-distributed values of the signal plus noise, Z_m, is given in Eq. (9.57) and for large n has a mean value of $\sigma(2\ln n)^{1/2}$; see Eq. (9.58). Thus for a given probability of occurrence, increasing the number of points to be searched from n_d to $n_d^{(n_a-1)}$ increases the level Z_m from $\sigma(2\ln n_d)^{1/2}$ to $\sigma[2(n_a - 1)\ln n_d]^{1/2}$; that is, the probability of finding a level $(n_a - 1)^{1/2}Z_m$ in a search of $n_d^{(n_a-1)}$ points is the same as that of finding a level Z_m in a search of n_d points. By increasing the number of antennas from 2 to n_a, the overall rms uncertainty in the signal level is reduced by a factor $[n_a(n_a - 1)/2]^{1/2}$, but since the detection threshold has increased by $(n_a - 1)^{1/2}$, the effective gain in sensitivity for detection of sources is increased by only $(n_a/2)^{1/2}$. Rogers (1991) and Rogers, Doeleman, and Moran (1995) consider other factors in deriving this result and show that the sensitivity increase $(n_a/2)^{1/2}$ should be multiplied by a factor which lies between 0.94 and 1. This factor is not included in Table 9.1.

Multielement Array with Incoherent Averaging

In Table 9.1 the last two lines are concerned with incoherent averaging of data taken with a multielement array. The method on line 5 involves data that have been averaged over the coherence time and subsequently averaged incoherently before the application of a global fringe search. The relative threshold value is the product of the threshold on line 4 for a multielement global search with that on line 2 for incoherent averaging over a single baseline. The method in line 6 involves incoherent averaging over both time segments (equal to the coherence time) and baselines. The relative threshold is obtained from that in line 2 by in-

creasing the number of data from N (the number of time segments per baseline) to N multiplied by the number of baselines.

9.5 PHASE STABILITY AND ATOMIC FREQUENCY STANDARDS

Precision oscillators have been steadily improved since the 1920s, when the invention of the crystal-controlled (quartz) oscillator had immediate application to the problem of precise timekeeping. In the early 1950s cesium-beam clocks allowed better timekeeping than could be obtained from astronomical observations. This development led to an atomic definition of time that differs from the astronomical one, and to the establishment of the definition of the second of time based on a particular transition frequency of cesium.

The mathematical theory of the interpretation of measurements of oscillator phase was systematized by an IEEE committee (Barnes et al. 1971). This paper helped standardize the approach to handling low-frequency divergence in the noise of oscillators. The physical theory of noise in oscillators was treated by Edson (1960). In this section we develop relevant aspects of the theory, and describe the operation of atomic frequency standards with particular emphasis on the hydrogen maser. The theory and analysis of phase fluctuations are discussed in more detail by Blair (1974) and Rutman (1978).

Analysis of Phase Fluctuations

The desired signal from an oscillator is a pure sine wave:

$$V(t) = V_0 \cos 2\pi \nu_0 t. \tag{9.91}$$

This is unobtainable since all devices have some phase noise. A more realistic model is given by

$$V(t) = V_0 \cos [2\pi \nu_0 t + \phi(t)], \tag{9.92}$$

where $\phi(t)$ is a random process characterizing the phase departure from a pure sine wave. We ignore amplitude fluctuations since they do not directly affect performance in VLBI applications. The instantaneous frequency $\nu(t)$ is the derivative of the argument of Eq. (9.92) divided by 2π, that is,

$$\nu(t) = \nu_0 + \delta\nu(t), \tag{9.93}$$

where

$$\delta\nu(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt}. \tag{9.94}$$

The instantaneous fractional frequency deviation is defined as

$$y(t) = \frac{\delta\nu(t)}{\nu_0} = \frac{1}{2\pi\nu_0}\frac{d\phi}{dt}. \tag{9.95}$$

This definition allows the performance of oscillators at different frequencies to be compared.

We assume that the random processes $\phi(t)$ and $y(t)$ are statistically stationary, so that correlation functions can be defined. This assumption is not always valid and can cause difficulty (Rutman 1978). The autocorrelation function of $y(t)$ is

$$R_y(\tau) = \langle y(t)y(t+\tau)\rangle. \tag{9.96}$$

$R_y(\tau)$ is a real and even function, so $\mathcal{S}'_y(f)$, the power spectrum of $y(t)$, is a real and even function of frequency f. In order to prevent confusion between $\nu(t)$ and its frequency components, we use the symbol f for the frequency variable in the following spectral analysis. Following the somewhat nonstandard convention that is used in most of the literature on phase stability (Barnes et al. 1971), we replace the double-sided spectrum $\mathcal{S}'_y(f)$ with a single-sided spectrum $\mathcal{S}_y(f)$, where $\mathcal{S}_y(f) = 2\mathcal{S}'_y(f)$ for $f \geq 0$, and $\mathcal{S}_y(f) = 0$ for $f < 0$. Since $\mathcal{S}'_y(f)$ is even, no information is lost in this procedure. Thus, the Fourier transform relation $R_y(\tau) \rightleftharpoons \mathcal{S}'_y(f)$, can also be written as

$$\begin{aligned} \mathcal{S}_y(f) &= 4\int_0^\infty R_y(\tau)\cos(2\pi f\tau)\,d\tau, \\ R_y(\tau) &= \int_0^\infty \mathcal{S}_y(f)\cos(2\pi f\tau)\,df. \end{aligned} \tag{9.97}$$

Similarly, the autocorrelation function of the phase is

$$R_\phi(\tau) = \langle \phi(t)\phi(t+\tau)\rangle. \tag{9.98}$$

$\mathcal{S}_\phi(f)$, the power spectrum of ϕ, and $R_\phi(\tau)$ are related by a Fourier transform. From the derivative property of Fourier transforms, the relationship between $\mathcal{S}_y(f)$ and $\mathcal{S}_\phi(f)$ can be shown to be

$$\mathcal{S}_y(f) = \frac{f^2}{\nu_0^2}\mathcal{S}_\phi(f). \tag{9.99}$$

$\mathcal{S}_y(f)$ and $\mathcal{S}_\phi(f)$ serve as primary measures of frequency stability. They both have the dimensions of Hz^{-1}. Another commonly used specification of oscillator performance is $\mathcal{L}(f)$, which is defined as the power in 1 Hz bandwidth at frequency f in one sideband of a double-sided spectrum, expressed as a fraction of the total power of the oscillator. When the phase deviation is small compared with one radian, $\mathcal{L}(f) \simeq \mathcal{S}_\phi(f)/2$.

A second approach to frequency stability is based on time domain measurements. The average fractional frequency deviation is

$$\overline{y}_k = \frac{1}{\tau}\int_{t_k}^{t_k+\tau} y(t)\,dt, \tag{9.100}$$

which, from Eq. (9.95), becomes

$$\bar{y}_k = \frac{\phi(t_k + \tau) - \phi(t_k)}{2\pi \nu_0 \tau}, \tag{9.101}$$

where the measurements of $\bar{y}_k$ are made with a repetition interval T ($T \geq \tau$) such that $t_{k+1} = t_k + T$ (see Fig. 9.9a). Measurements of $\bar{y}_k$ are directly obtainable with conventional frequency counters. The measure of frequency stability is the sample variance of $\bar{y}_k$, given by

$$\langle \sigma_y^2(N, T, \tau) \rangle = \frac{1}{N-1} \left\langle \sum_{n=1}^{N} \left(\bar{y}_n - \frac{1}{N} \sum_{k=1}^{N} \bar{y}_k \right)^2 \right\rangle, \tag{9.102}$$

where N is the number of samples in a single estimate of σ_y^2. In the limit as $N \to \infty$ the quantity presented above is the true variance, which we represent as $I^2(\tau)$. However, in many cases Eq. (9.102) does not converge because of the low-frequency behavior of $\mathcal{S}_y(f)$, and $I^2(\tau)$ is then not defined. To avoid some of the convergence problems, a particular case of Eq. (9.102), the two-sample or Allan variance, $\sigma_y^2(\tau)$, has gained wide acceptance (Allan 1966). The Allan variance,

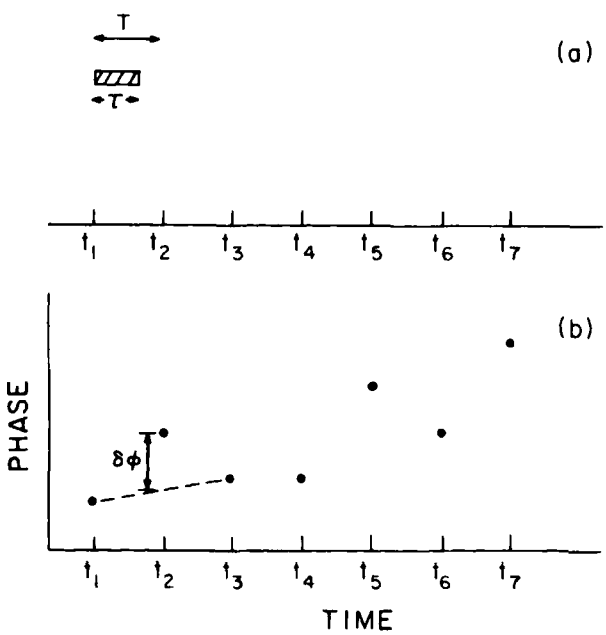


Figure 9.9 (a) Time intervals involved in the measurement of $\bar{y}_k$ as defined in Eq. (9.101). (b) Plot of a series of phase samples versus time. The Allan variance, defined in Eq. (9.103), is the average of the square of the deviation, $(\delta\phi)^2$, of each sample from the mean of its two adjacent samples.

for which $T = \tau$ (no dead time between measurements) and $N = 2$, is defined as follows:

$$\sigma_y^2(\tau) = \frac{\langle(\overline{y}_{k+1} - \overline{y}_k)^2\rangle}{2}, \tag{9.103}$$

or, from Eq. (9.101):

$$\sigma_y^2(\tau) = \frac{\langle[\phi(t+2\tau) - 2\phi(t+\tau) + \phi(t)]^2\rangle}{8\pi^2\nu_0^2\tau^2}. \tag{9.104}$$

The procedure for estimating the Allan variance can be understood as follows. Take a series of phase measurements at interval T, as shown in Fig. 9.9b. For each set of three independent points, draw a straight line between the outer two and determine the deviation of the center point from the line. With m samples of $\overline{y}$, the average of the squared deviations divided by $(2\pi\nu_0\tau)^2$ is an estimate of $\sigma_y^2(\tau)$, denoted $\sigma_{ye}^2(\tau)$, where

$$\sigma_{ye}^2(\tau) = \frac{1}{2(m-1)}\sum_{k=1}^{m-1}(\overline{y}_{k+1} - \overline{y}_k)^2. \tag{9.105}$$

The accuracy of this estimate is (Lesage and Audoin 1979)

$$\sigma(\sigma_{ye}) \simeq \frac{K}{\sqrt{m}}\sigma_y, \tag{9.106}$$

where K is a constant of order unity, whose exact value depends on the power spectrum of y.

We can now relate the true variance and the Allan variance to the power spectrum of y or ϕ. From Eq. (9.101) the true variance is $I^2(\tau) = \langle\overline{y}_k^2\rangle$, given by

$$I^2(\tau) = \frac{1}{(2\pi\nu_0\tau)^2}\left[\langle\phi^2(t+\tau)\rangle - 2\langle\phi(t+\tau)\phi(t)\rangle + \langle\phi^2(t)\rangle\right], \tag{9.107}$$

which, from Eq. (9.98), is

$$I^2(\tau) = \frac{1}{2(\pi\nu_0\tau)^2}\left[R_\phi(0) - R_\phi(\tau)\right]. \tag{9.108}$$

Then, since $R_\phi(\tau)$ is the Fourier transform of $\mathcal{S}_\phi(f)$, by using Eq. (9.99), we obtain from Eq. (9.108) the result

$$I^2(\tau) = \int_0^\infty \mathcal{S}_y(f)\left(\frac{\sin\pi f\tau}{\pi f\tau}\right)^2 df. \tag{9.109}$$

Similarly, from Eq. (9.104), we obtain

$$\sigma_y^2(\tau) = \frac{1}{(2\pi\nu_0\tau)^2}\left[3R_\phi(0) - 4R_\phi(\tau) + R_\phi(2\tau)\right] \tag{9.110}$$

and therefore,

$$\sigma_y^2(\tau) = 2\int_0^\infty \mathcal{S}_y(f)\left[\frac{\sin^4 \pi f\tau}{(\pi f\tau)^2}\right] df. \tag{9.111}$$

$I^2(\tau)$ and $\sigma_y^2(\tau)$ are dimensionless quantities, measured in rad^2, but we can think of them as the power obtained after filtering $y(t)$ with two different frequency responses, $H_I^2(f)$ and $H_A^2(f)$, respectively. These are

$$H_I^2(f) = \left(\frac{\sin \pi f\tau}{\pi f\tau}\right)^2 \tag{9.112}$$

and

$$H_A^2(f) = \frac{2\sin^4 \pi f\tau}{(\pi f\tau)^2}. \tag{9.113}$$

The functions $H_I^2(f)$ and $H_A^2(f)$ and the corresponding impulse responses $h_I(t)$ and $h_A(t)$ are shown in Fig. 9.10. Note that $I^2(\tau)$ can be estimated from a series of measurements $\overline{y}_k$ as the average of the square of $h_I(t_k) * \overline{y}_k$, where the asterisk indicates convolution. Similarly, $\sigma_y^2(\tau)$ can be estimated as the average of the square of $h_A(t_k) * \overline{y}_k$. Other transfer functions could be chosen. In time-domain measurements, additional filtering with high- and low-frequency cutoffs can be performed. For example, removing a long-term trend from the frequency data is a form of highpass filtering. Clearly, measurements of $\mathcal{S}_y(f)$ are preferable to those of $\sigma_y^2(\tau)$, because σ_y^2 can be calculated from $\mathcal{S}_y$ using Eq. (9.111), but $\mathcal{S}_y$ cannot be calculated from σ_y^2. However, in many cases of interest, as in the power-law spectra discussed below, the form of σ_y^2 is indicative of the behavior of $\mathcal{S}_y$. Traditionally, it has been easier to make time-domain measurements, and most published results are given in terms of the Allan variance σ_y^2.

The effect of local oscillator noise on the measured coherence of signals received at two antennas is given by Eq. (7.34) in terms of the rms deviation of the phase of the oscillator at one antenna relative to that at the other. For VLBI this rms deviation is equal to the square root of the sum of the true variances of the local oscillators at the two antennas. In the case of a connected-element array, low-frequency components of the phase noise of the master oscillator cause similar effects in the local oscillator phase at each antenna, and therefore their contributions to the relative phase at different antennas tend to cancel. For exact cancellation the time delay in the path of the reference signal from the master oscillator to each antenna, plus the time delay of the IF signal from the corresponding mixer to the correlator input (including the variable delay that compen-

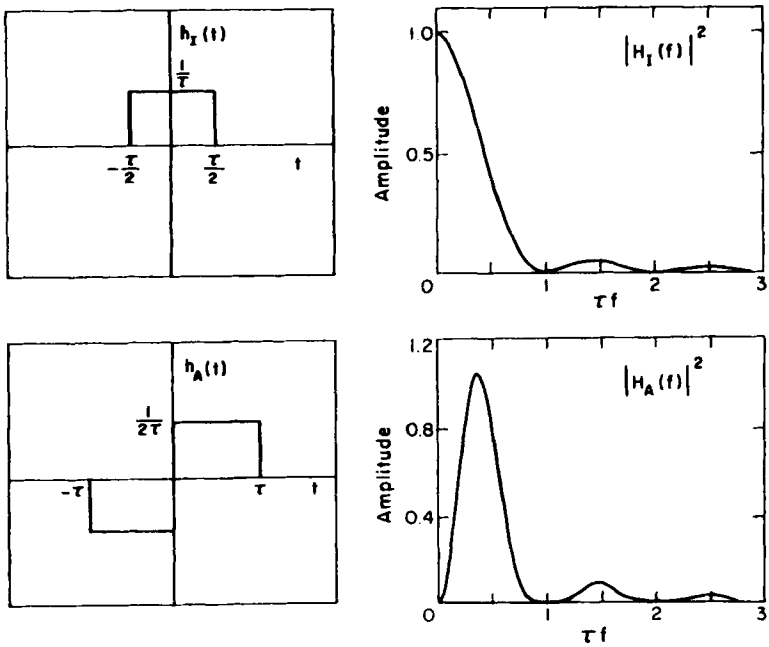


Figure 9.10 (*Top*) The impulse function $h_I(t)$ and the square of its Fourier transform, $|H_I(f)|^2$, given by Eq. (9.112), which is used to relate the power spectrum $\mathcal{S}_y(f)$ to the true variance $I^2(\tau)$, as defined in Eq. (9.109). (*Bottom*) The impulse response $h_A(t)$ and the square of its Fourier transform, $|H_A(f)|^2$, given by Eq. (9.113), which is used to relate the power spectrum $\mathcal{S}_y(f)$ to the Allan variance $\sigma_y^2(\tau)$, as defined in Eq. (9.111). Note that the sensitivity of the Allan variance decreases rapidly with decreasing frequency for $f < 0.3/\tau$.

sates for the geometric delay), should be equal for each antenna. It is generally impractical to preserve this equality. The bandwidths of phase-locked loops in the local oscillator signals at the antennas can also limit the frequency range over which phase noise in the master oscillator is canceled. In practice, cancellation of phase noise resulting from the master oscillator should generally be effective up to a frequency f in the range of a few hundred hertz to a few hundred kilohertz, depending on the parameters of the particular system.

Laboratory measurements show that $\mathcal{S}_y(f)$ is often a combination of power-law components. A useful model, shown in Fig. 9.11, is

$$\mathcal{S}_y(f) = \sum_{\alpha=-2}^{2} h_\alpha f^\alpha, \qquad 0 < f < f_h, \tag{9.114}$$

where α is a power-law exponent with integer values between -2 and 2, and f_h is the cutoff frequency of a lowpass filter. An equation similar to Eq. (9.114) can be written for $\mathcal{S}_\phi(f)$ using Eq. (9.99). Each term in Eq. (9.114) or the equivalent

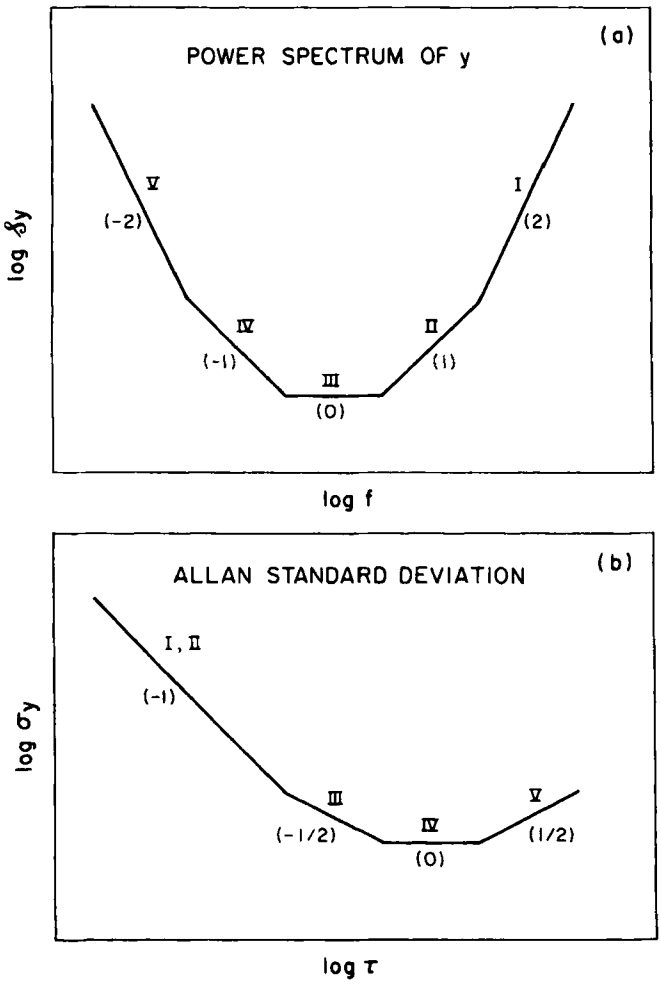


Figure 9.11 (a) The idealized power spectrum $\mathcal{S}_y(f)$ of the fractional frequency deviation $y(t)$ [see Eq. (9.114)]. The various spectral regimes are marked by Roman numerals, and the power-law coefficients are given in parentheses. The regimes are I, white-phase noise; II, flicker-phase noise; III, white-frequency noise; IV, flicker-frequency noise; and V, random-walk-of-frequency noise. (b) Two-point rms deviation, or Allan standard deviation, versus the time between samples. The spectral regimes are marked by the Roman numerals, and the power-law coefficients are given in parentheses.

equation for $\mathcal{S}_\phi(f)$ has a name based on traditional terminology (see Table 9.2). Noise with a power-law dependence f^0, independent of frequency, is called "white noise"; f^{-1} is called "flicker noise," or colloquially, "one-over-f noise"; and f^{-2} is called "random-walk noise." There are well-known origins for some of these processes, which we discuss briefly [see also Vessot (1976)]. The frequency dependence given in parentheses below is for $\mathcal{S}_y$.

TABLE 9.2 Characteristics of Noise in Oscillators[a]

Noise Type	$\mathcal{S}_y(f)$	$\mathcal{S}_\phi(f)$	$\sigma_y^2(\tau)$	μ^b	$I^2(\tau)$
White phase[c]	$h_2 f^2$	$\nu_0^2 h_2$	$\frac{3h_2 f_h}{4\pi^2\tau^2}$	-2	$\frac{h_2 f_h}{2\pi^2\tau^2}$
Flicker phase	$h_1 f$	$\nu_0^2 h_1 f^{-1}$	$\frac{3h_1}{4\pi^2\tau^2}\ln(2\pi f_h\tau)$	~ -2	—
White frequency or random walk of phase	h_0	$\nu_0^2 h_0 f^{-2}$	$\frac{h_0}{2\tau}$	-1	$\frac{h_0}{2\tau}$
Flicker frequency	$h_{-1} f^{-1}$	$\nu_0^2 h_{-1} f^{-3}$	$(2\ln 2)h_{-1}$	0	—
Random walk of frequency	$h_{-2} f^{-2}$	$\nu_0^2 h_{-2} f^{-4}$	$\frac{2\pi^2\tau}{3}h_{-2}$	1	—

[a] Adapted from Barnes et al. (1971).
[b] Power-law exponent of Allan variance: $\sigma_y^2(\tau) \propto \tau^\mu$.
[c] For $\sigma_y^2(\tau)$, $2\pi f_h \tau \gg 1$.

1. *White-phase noise* (f^2) is usually due to additive noise outside the oscillator, for example, noise introduced by amplifiers. This process dominates at large values of f, corresponding to short averaging times.
2. *Flicker-phase noise* (f^1) is seen in transistors and may be due to diffusion processes across junctions.
3. *White-frequency* or *random-walk-of-phase noise* (f^0) is due to internal additive noise within the oscillator, such as the thermal noise inside the resonant cavity. Shot noise also has this spectral dependence.
4. *Flicker-frequency noise* (f^{-1}) and *random-walk-of-frequency noise* (f^{-2}) are the processes that limit the long-term stability of oscillators. They are due to random changes in temperature, pressure, and magnetic field in the oscillator environment. This noise is associated with long-term drift. There is a large body of literature on flicker-frequency noise, which is encountered in many situations [see Keshner (1982) for a general discussion, Dutta and Horn (1981) for applications in solid-state physics, and Press (1978) for applications in astrophysics].

The variances $I^2(\tau)$ and $\sigma_y^2(\tau)$ can be calculated for the various types of noise described above. For $\alpha = 1$ and 2, the variances converge only if a high-frequency cutoff f_h is specified. With this restriction, σ_y^2 converges for all cases. $I^2(\tau)$ converges only for $\alpha \geq 0$. These functions are listed in Table 9.2. Except for the logarithmic dependence in flicker-phase noise, each noise component maps into a component of Allan variance of the form τ^μ. From Table 9.2 we can write the total Allan variance as

$$\sigma_y^2(\tau) = [K_2^2 + K_1^2 \ln(2\pi f_h \tau)]\tau^{-2} + K_0^2\tau^{-1} + K_{-1}^2 + K_{-2}^2\tau, \tag{9.115}$$

where the K values are constants. The subscripts correspond to the subscripts of h (see Table 9.2). White-phase and flicker-phase noise both result in $\mu \simeq -2$, but these two processes can be distinguished by varying f_h. Note that for white-phase and white-frequency noise the following relations hold [see Eqs. (9.109) and (9.111)]:

$$\sigma_y^2(\tau) = \tfrac{3}{2} I^2(\tau), \qquad \alpha = 2, \tag{9.116}$$

$$\sigma_y^2(\tau) = I^2(\tau), \qquad \alpha = 0. \tag{9.117}$$

In general, when $I^2(\tau)$ is defined, we see from Eqs. (9.108) and (9.110) that

$$\sigma_y^2(\tau) = 2[I^2(\tau) - I^2(2\tau)]. \tag{9.118}$$

Oscillator Coherence Time

A quantity of special interest in VLBI is the coherence time. The approximate coherence time is that time τ_c for which the rms phase error is 1 radian:

$$2\pi \nu_0 \tau_c \sigma_y(\tau_c) \simeq 1. \tag{9.119}$$

Rogers and Moran (1981) calculated a more exact expression for the coherence time that they defined in terms of the coherence function

$$C(T) = \left| \frac{1}{T} \int_0^T e^{j\phi(t)}\, dt \right|, \tag{9.120}$$

where $\phi(t)$ is the component of fringe phase of instrumental origin and T is an arbitrary integration time. $\phi(t)$ includes effects that cause the fringe phase to wander, such as atmospheric irregularities and noise in frequency standards. The rms value of $C(T)$ is a monotonically decreasing function of time with the range 1–0. The coherence time is defined as the value of T for which $\langle C^2(T)\rangle$ drops to some specified value, say, 0.5. The mean-square value of C is

$$\langle C^2(T)\rangle = \frac{1}{T^2} \int_0^T \int_0^T \langle \exp\left\{ j\left[\phi(t) - \phi(t')\right] \right\} \rangle\, dt\, dt'. \tag{9.121}$$

If ϕ is a Gaussian random variable, then

$$\langle C^2(T)\rangle = \frac{1}{T^2} \int_0^T \int_0^T \exp\left[-\frac{\sigma^2(t,t')}{2} \right] dt\, dt', \tag{9.122}$$

where $\sigma^2(t,t')$ is the variance $\langle [\phi(t) - \phi(t')]^2 \rangle$, which we assume depends only on $\tau = t' - t$. Then from Eq. (9.98),

$$\begin{aligned}\sigma^2(t', t) &= \sigma^2(\tau) \\ &= \langle[\phi(t) - \phi(t')]^2\rangle = 2[R_\phi(0) - R_\phi(\tau)].\end{aligned} \tag{9.123}$$

Note that $\sigma^2(\tau)$ is the structure function of phase and is related to $I^2(\tau)$ by Eq. (9.108):

$$\sigma^2(\tau) = 4\pi^2\tau^2\nu_0^2 I^2(\tau). \tag{9.124}$$

The integral in Eq. (9.122) can be simplified by noting that the integrand is constant along diagonal lines in (t, t') space for which $t' - t = \tau$. These lines have length $\sqrt{2}(T - \tau)$ so that

$$\langle C^2(T)\rangle = \frac{2}{T}\int_0^T \left(1 - \frac{\tau}{T}\right)\exp\left[-\frac{\sigma^2(\tau)}{2}\right] d\tau. \tag{9.125}$$

Thus, from Eqs. (9.109) and (9.124),

$$\langle C^2(T)\rangle = \frac{2}{T}\int_0^T \left(1 - \frac{\tau}{T}\right)\exp\left[-2(\pi\nu_0\tau)^2\int_0^\infty \mathcal{S}_y(f)H_I^2(f)\,df\right] d\tau, \tag{9.126}$$

where $H_I^2(f)$ is defined in Eq. (9.112). Since $\mathcal{S}_y(f)$ is often not available, it is useful to relate $\langle C^2(T)\rangle$ to $\sigma_y^2(\tau)$. We can solve Eq. (9.118) for $I^2(\tau)$ by series expansion, obtaining

$$2I^2(\tau) = \sigma_y^2(\tau) + \sigma_y^2(2\tau) + \sigma_y^2(4\tau) + \sigma_y^2(8\tau) + \cdots, \tag{9.127}$$

provided that the series converges. Therefore, from Eqs. (9.124), (9.125), and (9.127),

$$\langle C^2(T)\rangle = \frac{2}{T}\int_0^T \left(1 - \frac{\tau}{T}\right)\exp\left\{-\pi^2\nu_0^2\tau^2\left[\sigma_y^2(\tau) + \sigma_y^2(2\tau) + \cdots\right]\right\} d\tau. \tag{9.128}$$

This integral is readily calculable for the cases where $I^2(\tau)$ is defined.

We now consider white-phase noise and white-frequency noise, which are important processes in frequency standards on short time scales. For the case of white-phase noise, $\sigma_y^2 = K_2^2\tau^{-2}$, where $K_2^2 = 3h_2 f_h/4\pi^2$ is the Allan variance in 1 s (Table 9.2), and the coherence function can be evaluated from Eq. (9.126) or Eq. (9.128):

$$\langle C^2(T)\rangle = \exp\left(\frac{-4\pi^2\nu_0^2K_2^2}{3}\right) = \exp(-h_2 f_h \nu_0^2). \tag{9.129}$$

For white-frequency noise, $\sigma_y^2 = K_0^2 \tau^{-1}$, where $K_0^2 = h_0/2$, and we obtain

$$\langle C^2(T) \rangle = \frac{2(e^{-aT} + aT - 1)}{a^2 T^2}. \tag{9.130}$$

Here, $a = 2\pi^2 \nu_0^2 K_0^2 = \pi^2 h_0 \nu_0^2$. The limiting cases for white-frequency noise are

$$\begin{aligned} \langle C^2(T) \rangle &= 1 - \frac{2\pi^2 \nu_0^2 K_0^2 T}{3}, & 2\pi^2 \nu_0^2 K_0^2 T \ll 1, \\ &= \frac{1}{\pi^2 \nu_0^2 K_0^2 T}, & 2\pi^2 \nu_0^2 K_0^2 T \gg 1. \end{aligned} \tag{9.131}$$

The approximate relation for coherence time in Eq. (9.119) corresponds to rms values of the coherence function of 0.85 and 0.92 for white-phase noise and white-frequency noise, respectively. These calculations assume that one station has a perfect frequency standard. In practice, the effective Allan variance is the sum of the Allan variances of the two oscillators:

$$\sigma_y^2 = \sigma_{y1}^2 + \sigma_{y2}^2. \tag{9.132}$$

Thus, if two stations have similar standards, the coherence loss is doubled if the loss is small. If the short-term stability is dominated by white-phase noise, which is usually the case for hydrogen masers, the coherence function is independent of time. This means that there is a maximum frequency above which a particular standard will not be usable for VLBI, regardless of the integration time. This frequency is approximately $1/(2\pi K_2)$ Hz, which for a hydrogen maser is about 1000 GHz.

In practice, the coherence $C(T)$ is measured at the peak amplitude of the correlator output, which varies as a function of fringe frequency. This operation is equivalent to removing a constant frequency drift from the phase data and can be considered as highpass filtering of the data with a cutoff frequency of $1/T$. Modeling this operation as the response of a single-pole, highpass filter, one can show that it ensures the convergence of Eq. (9.128) for all processes for which the Allan variance exponent $\mu < 1$. To compare the various representations of frequency stability, we show in Figs. 9.12 and 9.13 an example of the performance of a hydrogen maser given by the functions σ_y^2, $\mathcal{S}_y(f)$, and $\langle C^2(T) \rangle^{1/2}$.

Precise Frequency Standards

Precise frequency standards of interest for VLBI include crystal oscillators and atomic frequency standards such as rubidium vapor cells, cesium-beam resonators, and hydrogen masers (Lewis 1991). Atomic frequency standards incorporate crystal oscillators that are phase-locked or frequency-locked to the atomic process, using loops with time constants in the range 0.1–1 s, so that short-term

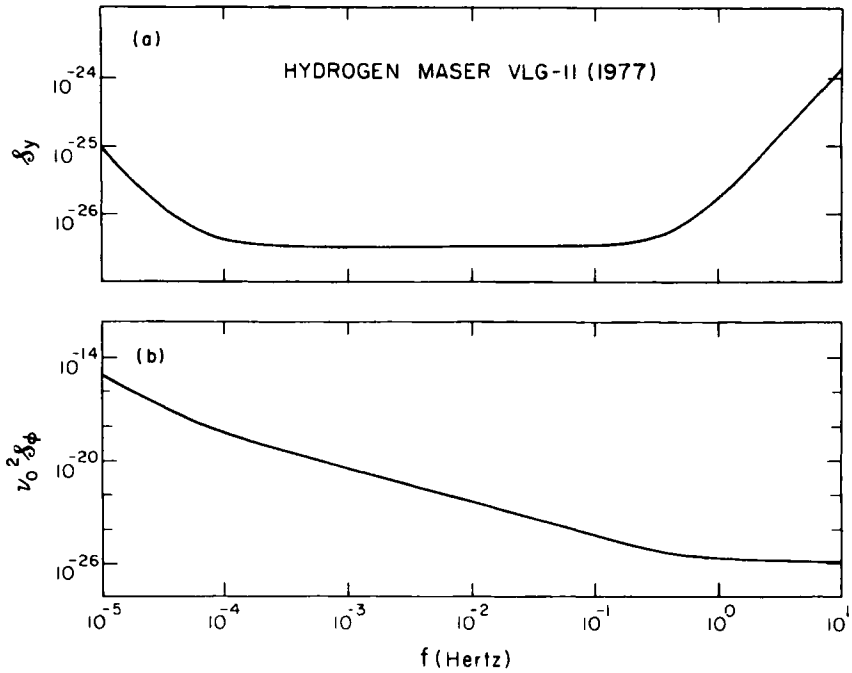


Figure 9.12 (a) Power spectrum of the fractional frequency deviation $\mathcal{S}_y(f)$ for a hydrogen maser frequency standard, and (b) the normalized power spectrum of the phase noise $\nu_0^2 \mathcal{S}_\phi(f)$. $\mathcal{S}_y(f)$ and $\mathcal{S}_\phi(f)$ are related by Eq. (9.99). For frequencies above 10 Hz, $\mathcal{S}_\phi(f)$ approaches the spectrum of the crystal oscillator to which the maser is locked, which declines as f^{-3}. The data were adapted from the measurements of Vessot (1979).

performance becomes that of the crystal oscillator. Details of how these loops are implemented are given by Vanier, Têtu, and Bernier (1979). The performance of the crystal oscillator is very important because unless it has high spectral purity, the phase-locked loops involved in generating the local oscillator signal from the frequency standard will not operate properly (Vessot 1976).

We first consider a frequency standard as a "black box" that puts out a stable sinusoid at a convenient frequency such as 5 MHz, or some higher frequency, at which the crystal oscillator is locked to the atomic process. The performance of various devices is shown in Fig. 9.14. These somewhat idealized plots show that the Allan variances of the standards have three regions: short-term noise dominated by either white-phase or white-frequency noise; flicker-frequency noise, which gives the lowest value of Allan variance and is therefore referred to as the "flicker floor"; and finally, for long periods, random-walk-of-frequency noise. Two other parameters can be specified, a drift rate and an accuracy. The drift rate is the linear change in frequency per unit time interval. Note that if the standard drives a clock, then a constant drift rate results in a clock error that accumulates

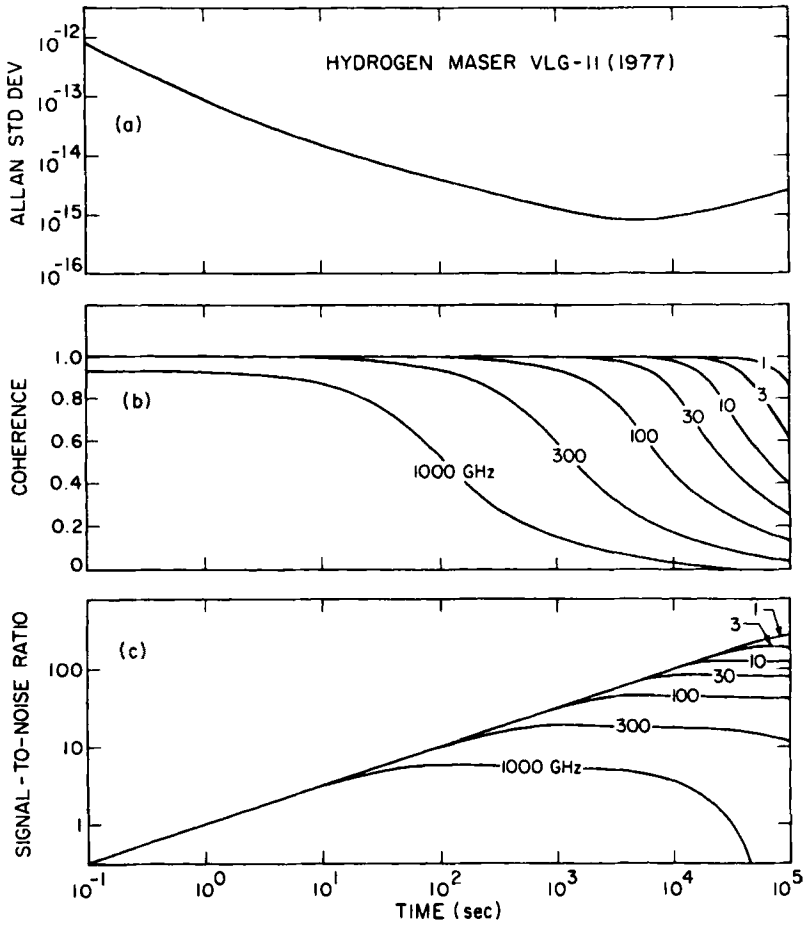


Figure 9.13 (a) Allan standard deviation versus sample time for a hydrogen maser frequency standard. Data from Vessot (1979). (b) Coherence $\sqrt{\langle C^2(T)\rangle}$, defined by Eq. (9.125), for various radio frequencies based on two frequency standards with Allan standard deviations given in (a). (c) Signal-to-noise ratio, normalized to unity at one second, of the measured visibility versus integration time for various frequencies. In a VLBI system the coherence and signal-to-noise ratios will be further reduced by atmospheric fluctuations.

as time squared. The accuracy refers to how well the standard can be set to its nominal frequency. The performance parameters are summarized in Table 9.3.

Atomic frequency standards are based on the detection of an atomic or molecular resonance. There are three parts to any frequency standard [e.g., Kartashoff and Barnes (1972)]. These are (1) particle preparation, (2) particle confinement, and (3) particle interrogation. Particle preparation involves enhancing the population difference in the desired transition. This is necessary for radio transitions in a gas with temperature T_g for which $h\nu/kT_g \ll 1$, so that the level populations

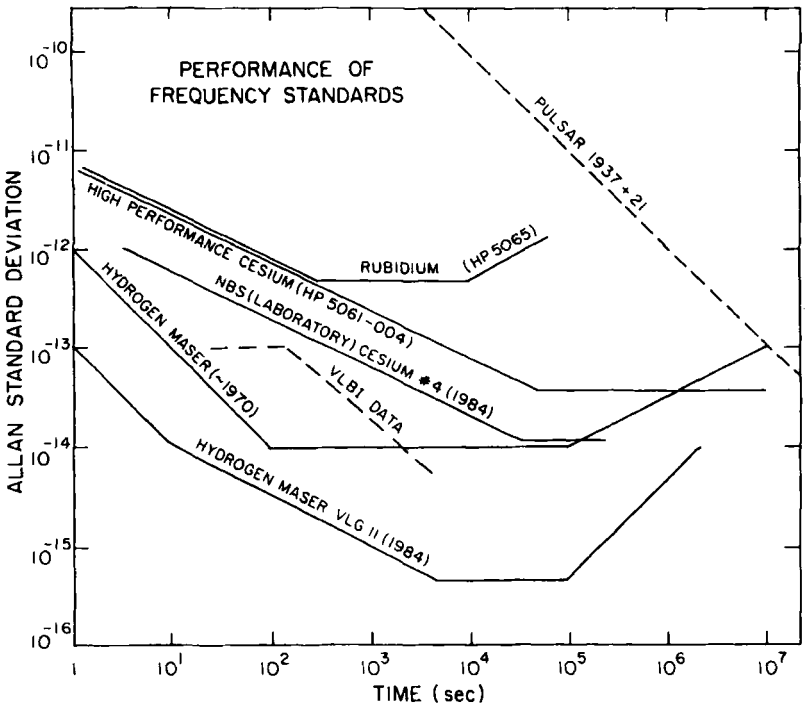


Figure 9.14 Idealized performance of various frequency standards and other systems. Pulsar data are from Davis et al. (1985). VLBI data, which show the effect of path length stability through the atmosphere in approximately average conditions, are from Rogers and Moran (1981).

TABLE 9.3 Typical Performance[a] Data on Available Frequency Standards[b]

Type	K_2 (10^{-12} s)	K_0 (10^{-12} $s^{1/2}$)	K_{-1} (10^{-15})	K_{-2} (10^{-17} $s^{-1/2}$)	Drift Rate[c] (10^{-15})	Fractional Accuracy (10^{-12})
H(active)	0.1	0.03	0.4	0.1	< 1	1
Cs	–	50	100	3	1	5
Cs[d]	–	7	40	3	1	2
Rb	–	7	500	300	10^2	10^2
Crystal	1	–	500	300	10^3	–

[a]Two-point Allan standard deviation; coefficient defined by Eq. (9.115).
[b]Updated from Hellwig (1979).
[c]Fractional frequency change per day.
[d]High performance Cs.

are nearly equal. Preparation is usually done either by state selection in a beam passing through a magnetic or electric field, or by optical pumping. Particle confinement makes it possible to obtain narrow resonance lines from long interaction times, since according to the Heisenberg uncertainty principle, the linewidth is equal to the reciprocal of the interaction time. Particles can be confined in beams or storage cells. Storage cells either contain a buffer gas or have specially coated walls so that particle collisions do not result in phase changes. Finally, particle interrogation is the process of sensing the interaction of particles and radiation fields. Frequency standards can be either active or passive. An example of an active standard is a maser oscillator. Passive standards require an external radiation field, and transitions are observed by (1) absorption, (2) re-emission, (3) detection of particles having made the transition, or (4) indirectly by detection of a quantity such as a variation in the rate of optical pumping. To show how some principles are implemented in practice, we give brief descriptions of the operation of several types of standards. Other types of frequency standards are under development [Drullinger, Rolston, and Itano (1996), Berkeland et al. (1998)].

Rubidium and Cesium Standards

Rubidium is an alkali metal with a single valence electron and thus a hydrogen-like spectrum. The electronic ground state is split into two levels, with a transition frequency of 6835 MHz. These levels correspond to the spin of the unpaired electron being parallel or antiparallel to the nuclear spin vector. A schematic diagram of the oscillator system is shown in Fig. 9.15. An RF plasma discharge in a tube containing ^{87}Rb excites the gas to an electronic level about 0.8 μm above the ground state. The light from this discharge passes through a filter that removes the components involving the $F = 2$ level and passes the light at 0.7948 μm. This filter consists of a cell of ^{85}Rb atoms whose energy levels are slightly shifted from those of the ^{87}Rb atoms, such that both gases have transitions near 0.7800 μm. The filtered light passes through another cell of ^{87}Rb gas inside a microwave cavity resonant at the transition frequency between the $F = 2$ and $F = 1$ levels. With no RF signal applied to the cavity, the gas is nearly transparent and the discharge beam is unattenuated as it reaches the photodetector. The application of an RF signal at 6835 MHz stimulates transitions from the $F = 2$ to $F = 1$ level. The atoms reaching the lower level are then pumped to the excited state by the light from the filtered ^{87}Rb lamp. The ^{87}Rb light therefore suffers absorption. A buffer gas, consisting of inert atoms that collide elastically with the ^{87}Rb atoms in the resonance cell, extends the interaction time to about 10^{-2} s, the mean collision time with the cell walls, and gives an absorption resonance with a linewidth of about 10^2 Hz. The cavity is magnetically shielded to minimize external fields. A weak homogeneous field is applied so that only $\Delta M_F = 0$ transitions, which have zero first-order Doppler shift, are obtained. The absorption resonance has a width of 10^2–10^3 Hz. The shot noise of individual arriving photons leads to white-frequency noise.

The radio frequency signal is frequency- or phase-modulated so that the resonance line is continuously scanned. A control voltage is generated by comparing

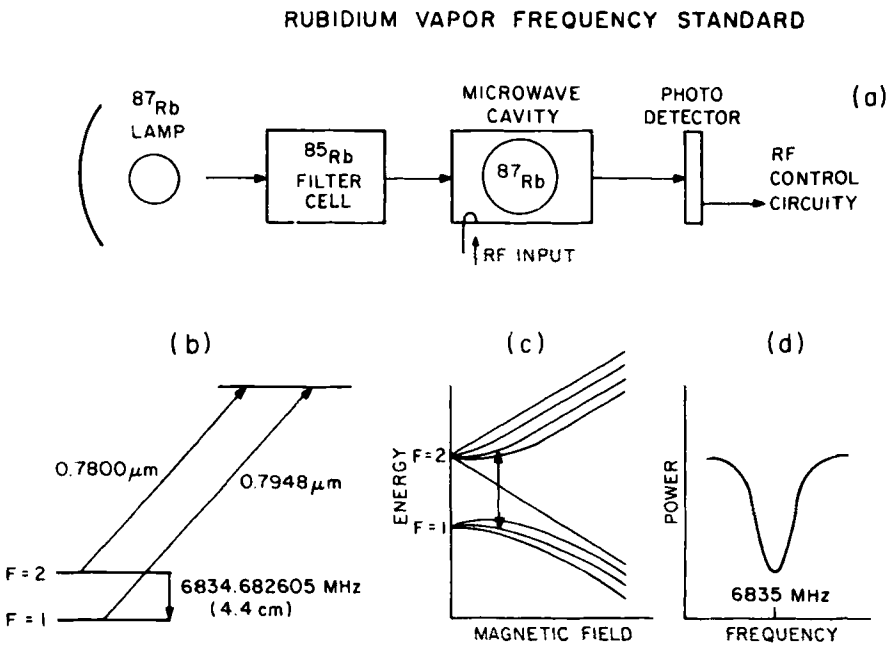


Figure 9.15 (a) Schematic diagram of a rubidium gas-cell frequency standard; (b) pump and microwave transitions; (c) magnetic sublevels of microwave transition versus magnetic field; (d) absorption of ^{87}Rb light versus microwave frequency. Adapted from Vessot (1976).

the modulation signal and the detector signal, and is fed back to the slave oscillator driving the cavity to correct its frequency to the peak of the resonance.

Rubidium standards have the advantage of being small, inexpensive, and readily transportable. They are sometimes used in VLBI below 1 GHz, where the ionosphere dominates system stability. At higher frequencies the use of rubidium standards results in degraded performance. They are useful as a backup for a primary standard, and can also be used in OVLBI spacecraft to reduce the uncertainty in the timing when the radio link from the ground station is interrupted.

Cesium, like rubidium, is an alkali metal with a single valence electron. The cesium standard is important because it is used to define the standard of atomic time. The frequency of the ground-state, spin-flip transition is exactly 9192.631770 MHz, by definition of the second of atomic time. A ribbon-shaped beam of cesium gas is passed through a state-selector magnet that passes the atoms in the $F = 3$ level into a resonator. Cesium frequency standards are larger and substantially more expensive than rubidium standards. Because of their low signal-to-noise ratio, their short-term stability is poor. Thus, they are not used in VLBI for controlling local oscillators. However, they provide excellent long-term stability and are used to monitor time. They have also been used to verify the capability of transferring time via VLBI (Clark et al. 1979). The historical development of the cesium-beam resonator is described by Forman (1985).

Hydrogen Maser Frequency Standard

The hydrogen maser is the usual VLBI standard, and we discuss its operating principles in some detail. The quantum mechanical analysis of the hydrogen maser is presented in a classic paper by Kleppner, Goldenberg, and Ramsey (1962). Fundamental principles of masers are given by Shimoda, Wang, and Townes (1956), and details of maser construction are given by Kleppner et al. (1965) and Vessot et al. (1976).

The hydrogen maser oscillator uses the ground-state, spin-flip transition at 1420.405 MHz, the well-known 21-cm line in radio astronomy. A schematic diagram of the oscillator is shown in Fig. 9.16. The hydrogen for the maser comes from a tank of molecular hydrogen gas that is dissociated in an RF discharge. The gas in the discharge is ionized and emits the reddish glow of the Balmer lines as the hydrogen atoms recombine and cascade to the ground state. The atomic gas flows out of the dissociator through a hexapole-magnet state selector. The inhomogeneous magnetic field separates the two upper states, $F = 1$, $M_F = 1$ and $F = 1$, $M_F = 0$, from the lower states, $F = 1$, $M_F = -1$ and $F = 0$, $M_F = 0$. The beam of atoms in the two upper states is directed into the storage bulb that is located inside a microwave cavity resonant in the TE_{011} or TE_{111} mode at 1420.405 MHz. The atoms bounce around the inside of the bulb about 10^5 times before escaping through the entrance hole. The spent atoms are evacuated from the system, which operates at low pressure, by an ion pump. The cavity is surrounded by several layers of material with high magnetic permeability that shield it from ambient magnetic fields. Inside the shield is a solenoid that creates a weak homogenous field. This field allows the ($F = 1$, $M_F = 0$)-to-($F = 0$, $M_F = 0$) transition to radiate and minimizes transitions from the $F = 1$, $M_F = 1$ level. There is no first-order Zeeman effect for the $\Delta M_F = 0$ transition (see Fig. 9.16). The maser will oscillate if the cavity is tuned close to the transition frequency and the losses are small enough. In the active maser, the 1420-MHz signal is picked up by a cavity probe and used to phase-lock a crystal oscillator from which a signal at the hydrogen line frequency has been synthesized.

The interaction lifetime of an atom in the bulb can be described by an exponential probability function

$$f(t) = \gamma e^{-\gamma t}, \tag{9.133}$$

where γ is the total relaxation rate. The line has an approximately Lorentzian profile with a linewidth (full width at half maximum) $\Delta\nu_0$ of γ/π. The most important contribution to γ is the rate at which atoms escape through the entrance hole. This rate is

$$\gamma_e = \frac{v_0 A_h}{6V}, \tag{9.134}$$

where $v_0 = \sqrt{8kT_g/m}$ is the average particle speed, T_g is the gas temperature, m is the mass of a hydrogen atom, A_h is the area of the entrance hole, and V is the volume of the bulb. γ_e is about 1 s^{-1}. The atoms lose coherence after many

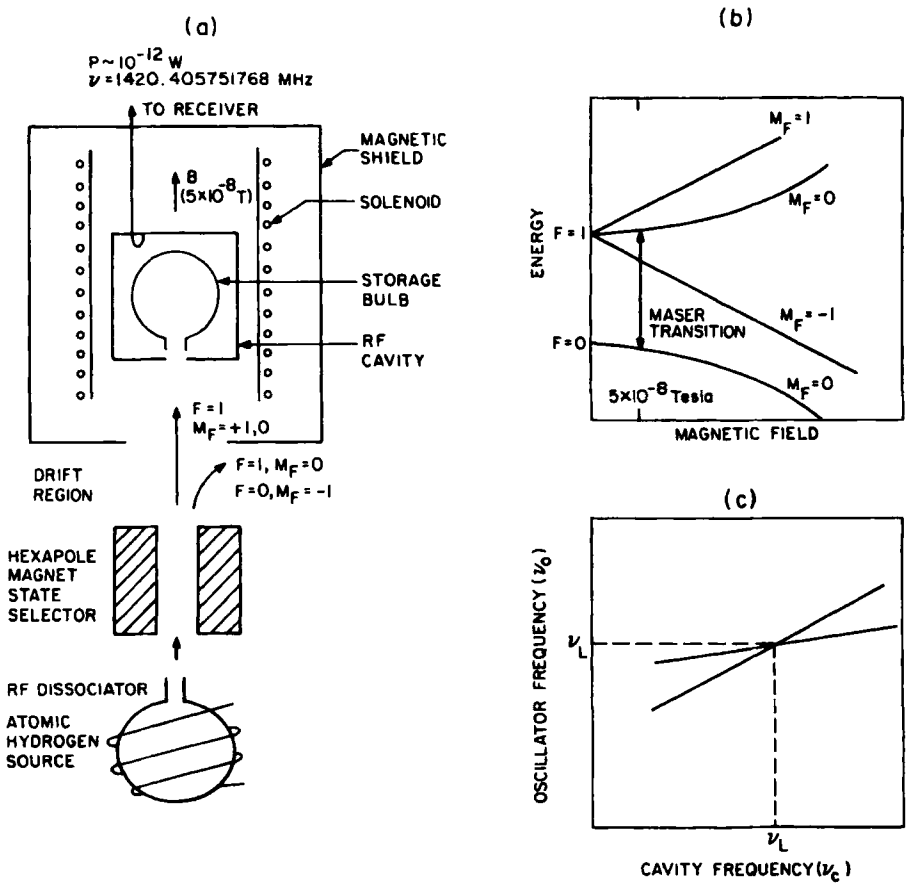


Figure 9.16 (a) Schematic diagram of a hydrogen maser frequency standard. The line frequency shown is the rest frequency of the transition in free space from Hellwig et al. (1970). The actual frequency will differ typically by ~0.1 Hz because of cavity pulling, second-order Doppler, and the wall shift. (b) Energies of magnetic sublevels versus magnetic field for the 21-cm transition. Adapted from Vessot (1976). (c) Curves of resonance frequency ν_0 versus cavity frequency ν_C for two values of linewidth [see Eq. (9.138)]. The intersection of the curves, which can be found empirically, gives the best operating frequency.

wall collisions, and this leads to a loss rate $\gamma_w \simeq 10^{-4}\ \mathrm{s}^{-1}$. Collisions between hydrogen atoms cause spin-exchange relaxation at a rate γ_{se} that is proportional to the gas density and to v_0. The net relaxation rate is approximately the sum of the three most important terms:

$$\gamma = \gamma_e + \gamma_w + \gamma_{se} = \pi \Delta\nu_0. \tag{9.135}$$

All three terms are proportional to v_0 and thus also to $\sqrt{T_g}$. Note that the random thermal motions of the atoms do not give rise to a first-order Doppler broadening

of the line, because the interaction between the atoms and the RF field takes place in a resonant cavity [see Kleppner, Goldenberg, and Ramsey (1962)].

The maser oscillator has two resonant frequencies, the line frequency ν_L and the electromagnetic cavity resonance frequency ν_C, defined by the cavity's dimensions. In classical oscillators the frequency is the mean of these two, weighted by the respective Q factors, Q_L for the line and Q_C for the cavity:

$$\nu_0 = \frac{\nu_L Q_L + \nu_C Q_C}{Q_L + Q_C}. \tag{9.136}$$

The Q factor is defined as π times the reciprocal of the fractional loss in energy per cycle of the resonant frequency. Hence, from Eq. (9.133), Q_L is given by [see, e.g., Siegman (1971)]

$$Q_L \simeq \frac{\pi \nu_0}{\gamma} = \frac{\nu_0}{\Delta \nu_0}. \tag{9.137}$$

A typical value of Q_L is about 10^9. The practical value of Q_C for a silver-plated cavity is about 5×10^4. Since $Q_L \gg Q_C$, the resonance frequency is approximately

$$\nu_0 \simeq \nu_L + \frac{Q_C}{Q_L}(\nu_C - \nu_L). \tag{9.138}$$

Equation (9.138) describes the effect of "cavity pulling" on the resonance frequency. Temperature changes cause the size, and thus the resonant frequency, of the cavity to change. Hence, a fractional frequency stability of 10^{-15} for the maser requires a fractional mechanical stability of about 5×10^{-10} for the cavity. The cavity dimensions therefore must be stable to about 10^{-8} cm. The cavity must be made from material with a small thermal expansion coefficient or the temperature must be carefully controlled. Extreme mechanical stability is also required so that atmospheric pressure changes do not affect the frequency. The TE_{011} cavity is a cylinder about 27 cm in length and diameter, appreciably larger than the free space wavelength because of the loading by the storage bulb. Coarse tuning is accomplished by moving the end plate of the cavity and fine tuning by a varactor diode. From Eq. (9.138) it is clear that the maser frequency is most stable when ν_C is set to ν_L so that ν_0 equals ν_L regardless of the values of Q_C and Q_L. This optimal tuning point of the maser can be found by making a plot of ν_0 versus ν_C, which is a straight line with slope Q_C/Q_L, according to Eq. (9.138). By varying Q_L (for example, by varying the gas pressure and thereby changing γ), a family of straight lines can be generated that intersect at the desired frequency $\nu_0 = \nu_L = \nu_C$ (see Fig. 9.16c). Servomechanisms are used in some systems to keep the maser cavity continuously tuned.

The performance of hydrogen masers is shown in Figs. 9.13 and 9.14. For periods less than 10^3 s the performance is limited by two fundamental processes: (1) white-frequency noise due to thermal noise generated inside the cavity and (2) white-phase noise due to thermal noise in the external amplifier. The thermal noise generated inside the cavity produces a fractional frequency variance (Allan

variance) of

$$\sigma_{yf}^2 = \frac{1}{Q_L^2}\frac{kT_g}{P_0\tau}, \tag{9.139}$$

where P_0 is the power delivered by the atoms (Edson 1960; Kleppner, Goldenberg, and Ramsey 1962). There is also shot noise in the cavity due to the discrete radiation of photons. However, this process, described by the Allan variance σ_{ys}^2, is smaller than σ_{yf}^2 by the ratio $h\nu/kT_g$, which is 2×10^{-4} at room temperature. Spontaneous emission also contributes a small amount of noise, equivalent to increasing T_g by $h\nu/k \simeq 0.07$ K. Finally, the maser receiver adds a noise power $kT_R\Delta\nu$ to the signal coupled out of the cavity, where T_R is the receiver noise temperature and $\Delta\nu$ is the receiver bandwidth. This noise causes an Allan variance of (Cutler and Searle 1966)

$$\sigma_{yR}^2 = \frac{1}{(2\pi\nu_0\tau)^2}\frac{kT_R\Delta\nu}{P_0}. \tag{9.140}$$

These two processes are independent, so the net Allan variance is $\sigma_y^2 = \sigma_{yf}^2 + \sigma_{yR}^2$. The effects of both processes are clearly evident in the data in Fig. 9.14. Note that a flicker floor is not reached because of long-term drifts. The short-term performance can be improved by increasing the atomic flux level, which increases P_0. However, increasing the flux increases the spin-exchange rate, which decreases Q_L, thereby making the oscillator more susceptible to the long-term effects of cavity pulling.

The frequency of a maser is not exactly equal to the atomic transition frequency because of several effects. These effects limit the accuracy to which the frequency can be set, and because most of them are temperature dependent, they probably contribute to flicker-frequency and random-walk-of-frequency noise. Cavity pulling, which has been described already, is an important effect, and to minimize it the cavity must be tuned carefully. The collision-induced spin-exchange process gives a frequency shift that varies with Q_L in the same way as the cavity pulling. Thus, the cavity-tuning procedure also eliminates this shift. Collisions with the cavity walls produce an effect called the "wall shift," which is difficult to predict and may be the ultimate limiting factor in the absolute precision of the maser frequency (Vessot and Levine 1970). This shift depends on the temperature and wall coating material. Its fractional value is about 10^{-11}. The first-order Doppler effect cancels, but the second-order Doppler effect does not, because of its v^2/c^2 dependence [see Kleppner, Goldenberg, and Ramsey (1962)]. The fractional frequency shift is about equal to $-1.4 \times 10^{-13}T_g$. Finally, there is no first-order Zeeman effect in the $(F = 1, M_F = 0)$-to-$(F = 0, M_F = 0)$ transition. However, the second-order Zeeman fractional-frequency shift is $2.0 \times 10^2 B^2$, where B is the magnetic field in tesla.

Local Oscillator Stability

Local oscillator signals are generated by multiplying a signal from the locked oscillator of the frequency standard. The multipliers must have exceptional stability,

as discussed in Section 7.2, to avoid the introduction of additional noise and drift. Imperfect multipliers are sensitive to vibration and temperature and may have modulation at harmonics of the power line frequency. In an ideal multiplier, a signal of the form of Eq. (9.92) is converted to

$$V(t) = \cos[2\pi M\nu_0 t + M\phi(t)], \tag{9.141}$$

where M is the multiplication factor, ν_0 is the fundamental frequency, and ϕ is the random phase noise of the frequency standard. If the phase noise is small, $M\phi(t) \ll 1$, then the single-sided power spectrum of $V(t)$ is given by

$$\mathcal{S}_v(\nu) = \delta(\nu - M\nu_0) + M^2\mathcal{S}_\phi(\nu - M\nu_0), \tag{9.142}$$

where δ is a delta function representing the desired signal and $\mathcal{S}_\phi$ is the power spectrum of the phase noise. Thus, the noise power increases as the square of the multiplication factor. In the general case, $\mathcal{S}_v$ can be written (Lindsey and Chie 1978)

$$\mathcal{S}_v(\nu) = \delta(\nu - M\nu_0) + \sum_{n=1}^{\infty} \frac{M^{2n}}{n!}\left[\mathcal{S}_\phi(\nu - M\nu_0) * \mathcal{S}_\phi(\nu - M\nu_0) * \cdots\right]. \tag{9.143}$$

where the term in brackets contains n replications of the same function convolved together. When only the leading term in the summation is retained Eq. (9.143) reduces to Eq. (9.142). The higher-order terms in Eq. (9.143) represent a series of approximately Gaussian components because of the repeated convolutions. The rms phase deviation of the multiplier output frequency $M\nu_0$ is proportional to the rms voltage of the noise in the output bandwidth, that is, to the square root of the noise power. Thus for the case represented by Eq. (9.142), the rms phase fluctuation is proportional to M.

Phase Calibration System

One way to check the integrity of an entire VLBI system is to inject into the front end of the receiver an RF signal that is independently derived from the frequency standard. The RF test signal can be derived by driving a step-recovery diode with, say, a 1-MHz signal from the frequency standard so as to generate a pulse train with 1-μs period. Such a signal has harmonics at 1-MHz intervals throughout the microwave region, all of which have the same phase at the reference intervals. When the RF band is mixed down to baseband, one of the injected harmonics can be made to appear at a convenient frequency of order 10 kHz. This is then compared with a reference signal from the frequency standard. The phase calibration signal can be continuously injected during VLBI recording since a low enough level can be used that it can only be detected by very narrowband filtering in the processor ($\sim$10-Hz bandwidth). The calibration allows one to compensate for variations such as those caused by thermal effects in cables (Whitney et al.

1976, Thompson and Bagri 1991, Thompson 1995). Similar methods are used in some linked-element interferometers.

Time Synchronization

The clocks at VLBI stations must be synchronized accurately enough to avoid time-consuming searches for interference fringes. Until around 1980, Loran C was widely used to monitor time at VLBI stations. Loran, an acronym for *Long Range Navigation*, is a system originally developed during World War II for ocean navigation (Pierce, McKenzie, and Woodward 1948). The transmission frequency is 100 kHz. The relative time of arrival of signals from three stations defines the observer's location on the earth's surface. For a detailed discussion of Loran C, see Frank (1983). Accuracies from a few hundred nanoseconds to a few tens of microseconds are possible, depending upon the accuracy of the estimate of propagation time.

The Global Positioning System (GPS) provides higher accuracy than Loran and has been used in almost all VLBI systems since the early 1980s. In the GPS system the user receives signals at 1.23 or 1.57 GHz from a number of satellites whose positions are known and whose clocks are synchronized to Coordinated Universal Time (UTC; see Section 12.3). If timing measurements from four satellites are made, and corrected for propagation effects in the atmosphere, users can determine their positions in three coordinates and their clock errors. The accuracies available to civilian users have improved over about a decade from 100 ns in time (Parkinson and Gilbert 1983, Lewandowski and Thomas 1991) to ~7 ns, and further improvement is expected (Lewandowski, Azoubib, and Klepczynski 1999). An analysis of the time transfer problem, including relativistic effects, is given by Ashby and Allan (1979). For general information on GPS usage see, for example, Leick (1995).

For time scales of a year, the accuracy of timing from pulsar observations approaches 1 part in 10^{14} (Davis et al. 1985). Ultimately, the best time transfer may be obtainable from the processed VLBI data (Clark et al. 1979).

9.6 RECORDING SYSTEMS

The basic consideration for any recording system is the representation of the signal and the method of incorporating the time information. Recording can be either analog or digital, and various data storage technologies are available. Here we discuss only digital recording on magnetic tape since the technologies involved are well suited to VLBI and are widely used.

A basic parameter of a recording system is its data rate, ν_b (bits s^{-1}). This parameter limits the number of bits that can be recorded in a given time, and thus also the sensitivity of continuum observations in which the potential IF bandwidth is larger than $\nu_b/2N_b$, where N_b is the number of bits per sample. The signal is represented by samples having Q quantization levels taken at β times the Nyquist rate. For N samples there are Q^N possible data configurations, which require a minimum of $N \log_2 Q$ bits. Therefore, as noted in Section 8.3 under *Comparison*

of Quantization Schemes, the maximum RF bandwidth is

$$\Delta\nu = \frac{\nu_b}{2\beta N_b} = \frac{\nu_b}{2\beta \log_2 Q}. \tag{9.144}$$

The signal-to-noise ratio obtained in time τ is proportional to $\eta_Q\sqrt{\Delta\nu\tau}$, where η_Q is the quantization efficiency introduced in Section 8.3. From Eq. (9.144),

$$\eta_Q\sqrt{\Delta\nu\tau} = \eta_Q\sqrt{\frac{\nu_b\tau}{2\beta N_b}}. \tag{9.145}$$

If τ is the recording time for the tape, $\nu_b\tau$ is equal to the number of recorded bits on the tape. The quantity $\eta_Q/\sqrt{\beta N_b}$ thus provides an indication of the performance per bit, which it is desirable to maximize. For two- and four-level sampling, the obvious encoding schemes are one bit and two bits per sample, respectively. For three-level sampling a problem arises since encoding one sample (one of three possible states) in two data bits (representing four possible states) is inefficient. Putting three samples into five bits or five samples into eight bits gives data rates of 1.67 and 1.60 bits per sample, respectively, compared to the theoretical optimum value of $\log_2 3 = 1.585$. The values of $\eta_Q/\sqrt{\beta N_b}$ for various values of Q and β, and several encoding schemes, are listed in Table 9.4. The

TABLE 9.4 Performance of Various Signal Representations as a Function of Number of Quantization Levels, Sampling Rate, and Encoding Format[a]

Signal Representation		η_Q	N_b	$\frac{\eta_Q}{\sqrt{\beta N_b}}$
Sampling at Nyquist Rate ($\beta = 1$)				
Two-level		0.637	1.0	0.637
Three-level	"Ideal" encoding[b]	0.810	1.585	0.643
	5 samples /8 bit	0.810	1.60	0.640
	3 samples/5 bit	0.810	1.667	0.627
	1 sample/2 bit	0.810	2.0	0.573
Four-level	All products	0.881	2.0	0.623
	Low-level omitted	0.87	2.0	0.61
Sampling at 2× *Nyquist Rate* ($\beta = 2$)				
Two-level		0.74	1.0	0.52
Three-level	"Ideal" encoding[b]	0.89	1.585	0.50
	5 samples/8 bit	0.89	1.60	0.50
	3 samples/5 bit	0.89	1.667	0.49
	1 sample/2 bit	0.89	2.0	0.45
Four-level	All products	0.94	2.0	0.47

[a] η_Q = quantization efficiency; N_b = number of bits per sample; β = oversampling factor.
[b] N samples encoded in $N \log_2 3$ bits.

highest signal-to-noise ratio is achieved with three-level sampling at the Nyquist rate, although two- and four-level sampling give almost the same performance.

In addition to the encoding schemes discussed above, in which the number of bits required for a given number of samples is constant, one can also envisage a scheme in which the number of bits depends on the sample values, that is, a variable-length code. For example, D'Addario (1984) has suggested encoding the $+1$, 0, and -1 values in three-level quantization as the binary numbers 11, 0, and 10, respectively. It is possible to decode such a data string uniquely, since all one-bit representations begin with 0 and all two-bit representations with 1. The average number of bits per sample depends on the amplitude probability distribution of the signal waveform and the threshold level settings. For a given number of bits, the threshold settings that maximize the signal-to-noise ratio are generally not the same as those derived in Section 8.3, which are optimum for a given number of bits per sample. With D'Addario's encoding scheme, the best performance is achieved with the threshold set such that $\eta_Q = 0.769$ and $N_b = 1.370$ bits per sample, giving a performance factor $\eta_Q/\sqrt{\beta N_b}$ equal to 0.657. Thus, an increase in sensitivity of about 3% compared with the use of the scheme with 1.6 bits per sample could be achieved. However, the effects of bit errors or interfering signals that change the amplitude distribution could be more serious. Finally, the data could be encoded statistically in large blocks that would allow a theoretically optimal value of N_b of 1.317 bits per sample, which, with η_Q of 0.769, would give a performance factor of 0.670 (D'Addario 1984).

In practice, the desirability of a simple encoding scheme and other design considerations have usually resulted in the choice of two-level quantization. All five VLBI systems developed in the United States during the period 1968–1997 (Mark I, Mark II, Mark III, VLBA, and Mark IV) use two-level sampling, but for the last two of these four-level sampling is also an option. For spectral line observations, where the bandwidth of the signal is small with respect to the bandwidth of the recording system, multilevel sampling is advantageous. Note that multilevel sampling is a more effective way of using recording capacity than sampling faster than the Nyquist rate (Table 9.4).

Each data sample must have either an implicit or explicit time tag. Although an error rate of 10^{-3} in decoding the data bits is acceptable, a one-bit shift in the time axis can be a serious defect and is not acceptable. In virtually all recording systems the data are blocked into records. Each new record begins at a precise time so that the temporal registration of the data stream can be recovered if it is lost during the previous record. These record lengths are: Mark I, 0.2 s (144,000 bits); Mark II, 16.7 ms (66,600 bits); and Mark III, 5 ms (20,000 bits). In the Mark I system, which used standard computer tape format, the accuracy of recording was very high, and the time of any bit was obtained by counting bits from the beginning of the record and counting records from the beginning of the tape. In the Mark II system, which uses video cassette recorders (VCRs), the data are recorded with a self-clocking code, while in the Mark III system, which uses instrumentation recorders, the data transitions themselves serve as the clock. The characteristics of several systems are given in Table 9.5. In all of these the recording is in digital form, except for the Canadian system used during 1971–1883. Wietfeldt and D'Addario (1991) discuss the compatibility of some of these systems.

TABLE 9.5 Characteristics of Some VLBI Tape-Recording Systems

System	Period of Use	Basic Description	Tape Recorder	Sample Rate[a] (10^6 s^{-1})	Tape Time (min)	References
NRAO Mark I[b]	1967–78	IBM computer-compatible format	Ampex TM-12	0.72	3.2	Bare et al. (1967)
NRAO Mark II(A)	1971–78	Digital recording on TV recorder	Ampex VR660C	4	190	Clark (1973)
NRAO Mark II(B)	1976–82	Digital recording on TV recorder	IVC 800	4	64	
NRAO Mark II(C)	1979–	Video cassette recorder	RCA VCT 500	4	246	
Canadian	1971–83	Analog recording on TV recorder	IVC 800	8	64	Broten et al. (1967), Moran (1976)
MIT/NASA Mark III	1977–	Instrumentation recorder	Honeywell 96	112[c]	13.6	Rogers et al. (1983)
MIT/NASA Mark III(A)	1984–	Instrumentation recorder	Honeywell 96[d]	112[c]	164	Clark et al. (1985)
NRAO VLBA	1990–	Instrumentation recorder	Honeywell 96[d]	128[e]	720[f]	Hinteregger et al. (1991), Rogers (1995)
MIT/NASA Mark IV	1997–	Instrumentation recorder	Honeywell 96[d]	1024	90	Whitney (1993), Rogers (1995)
S2 (Canada)	1992–	8 Video cassette recorders		128	256	Wietfeldt et al. (1996) Cannon et al. (1997)
K-4 (Japan)	1990–	Video cassette recorder	Sony DIR-1000	256	63	Kawaguchi (1991)

[a] Two-level quantization can be used with all of these systems, and four-level can also be used with the VLBA and Mark IV systems.
[b] A similar system was developed in the former Soviet Union (Kogan and Chesalin 1981).
[c] Used with 14 baseband converters of bandwidth 2 MHz or narrower and outputs for both upper and lower sidebands.
[d] Trackwidth is 40 μm. Other models equivalent to Honeywell 96 also used.
[e] Average rates. Eight baseband converters of 16 MHz bandwidth or narrower, outputs for both upper and lower sidebands, and one- or two-bit quantization provide data rates up to 1024 Mbits s^{-1}.
[f] Minimum specification.

9.7 PROCESSING SYSTEMS AND ALGORITHMS

A VLBI processor has two main functions: (1) reproduction of smooth data streams and (2) cross-correlation analysis of the data streams. The data stream from a tape recorder can be expected to have time-base irregularities of up to 100 μs, caused by jitter in the mechanical playback system, and to be subject to dropouts because of tape imperfections. The processor must derive the true time base either from the encoded clock transitions in the case of a self-clocking code, or from the data transitions themselves when a bit synchronizer is used. There must be enough buffer storage to handle at least the mechanical jitter. The geometric delay can be corrected with minimal buffer space by shifting the playback time, thereby retaining the data on the tape until they are needed by the correlator. If the data are read in synchronism from the tapes, a buffer memory of sample capacity about 5×10^4 times the clock rate in megahertz is needed for geometric delay compensation.

The major differences between the design of the correlation part of the processor for VLBI and for a conventional interferometer are related to the fact that in VLBI, fringe rotation and delay compensation are usually performed on the quantized and sampled signal. This leads to special problems, which we discuss here. Digitization of the signals introduces several signal-to-noise loss factors: η_Q, the loss factor associated with amplitude quantization of the recorded signals, discussed in Section 8.3; η_R, the loss factor incurred by quantizing the phase of the fringe rotation waveform; η_S, the loss factor incurred by inadequate sideband rejection as a result of the limited number of delays in the correlator; and η_D, the loss caused by compensating the geometric delay in discrete steps.

Fringe rotation and delay compensation can be done on the analog signals at the telescope before recording. For example, the fringe rotation can be done at the telescopes by offsetting the local oscillators as described in Section 6.1 under *Delay Tracking and Fringe Rotation* for a connected-element array. The advantage of this arrangement is that only a real correlation function (with both positive and negative delays) needs to be calculated (see Sections 8.7 and 9.1). Hence only half the correlator circuits are required. Also, the sensitivity loss from a digital fringe rotator is not incurred. A disadvantage is that the output of the correlator must be averaged over a short enough interval to accommodate the residual fringe frequency of a source anywhere in the primary beams of the antennas. The maximum residual fringe frequency of a source at the half-power point of the primary beam is $\Delta\nu_f \simeq D\omega_e/d$ [see Eq. (12.21)], where D is the baseline length, d is the antenna diameter, and ω_e is the angular velocity of the earth in radians per second. Hence, the averaging time of the correlator output must be less than $1/(2\Delta\nu_f)$; for example, it should not exceed 30 ms for a baseline equal to the earth's diameter and $d = 25$ m. The correlation functions can be averaged further after they have been passed through a fringe rotator, which removes the residual fringe frequency. Also, the unit at the telescope that continually changes the local oscillator frequency must be carefully designed so that full phase accountability is provided for astrometric work. Further information on VLBI systems and processing algorithms can be found in Thomas (1981) and Herring (1983).

Fringe Rotation Loss (η_R)

Fringe rotation is used to reduce to near zero the frequency of the fringe component of the correlated signals (see Section 6.1 under *Delay Tracking and Fringe Rotation*). Here we consider the fringe frequency to include the effect of offsets in the frequency standards. Fringe rotation in the processor can be implemented in a number of ways, as shown in Fig. 9.17. If the fringe rotator is placed after the correlator (Fig. 9.17a), then the correlation function from the correlator must be averaged over an interval short with respect to the fringe period. If the local oscillators at the antennas are offset to slow the fringes, so that only a little further adjustment is required after the correlator, then this scheme is convenient. Otherwise the short averaging time required and the resulting high data rate from the correlator make this arrangement unattractive. Alternatively, before correlation, one of the data streams can be passed through a digital single-sideband mixer that shifts the Fourier components of the signal by the appropriate fringe frequency as shown in Fig. 9.17b. The 90° phase shift in this mixer is difficult to implement without introducing spectral distortion, so this type of fringe rotator is rarely used (see also Section 8.6). The fringe rotation scheme shown in Fig. 9.17c is commonly used, but application of fringe rotation to the quantized signal introduces two complications. First, the fringe function with which the signal is multiplied must be coarsely quantized so as not to increase the number of bits per sample going to the correlator—this also applies to scheme (b). Second, the multiplication introduces an unwanted noise sideband, which is described below under *Fringe Sideband Rejection Loss*. We now consider the first of these effects.

The data stream is multiplied by a complex function $\mathcal{F}$ whose real and imaginary parts, $\mathcal{F}_R$ and $\mathcal{F}_I$, approximate $\cos\phi$ and $\sin\phi$, where ϕ is the desired phase function. In the simplest approximation these functions are square waves with the appropriate frequency and phases. Thus, as shown in Fig. 9.18, the quantized signal is multiplied by a fringe rotation function whose amplitude is constant but whose phase steps by 90° every quarter cycle instead of smoothly progressing. The resulting visibility function then has a phase component with a 90° sawtooth modulation at the fringe frequency. This resembles phase noise in which the phase is uniformly distributed between ±45°. Therefore, the average signal amplitude is degraded by $\sin(\pi/4)/(\pi/4) = 0.900$. Another approach to calculating the loss in signal-to-noise ratio is to calculate the harmonics in the fringe rotation function. The first harmonic of $\mathcal{F}_R$ or $\mathcal{F}_I$ has an amplitude of $4/\pi = 1.273$. Only the signal mixed with the first harmonic appears in the processor output, since the other harmonics are removed by time averaging. Thus part of the signal is scattered out of the fringe passband. The fraction retained is the square root of the ratio of the power in the first harmonic to the total power of the fringe rotation function, which is $\sqrt{8}/\pi = 0.900$. This represents the loss in signal-to-noise ratio. There is also a scale-factor change since the fringe amplitudes are increased by the action of the fringe rotator. Thus the fringe amplitudes must be divided by $4/\pi$, the relative amplitude of the first harmonic of $\mathcal{F}_R$.

A better fringe rotation function is the three-level approximation of a sine wave (Clark, Weimer, and Weinreb 1972) shown in Fig. 9.18b. When the fringe rotation

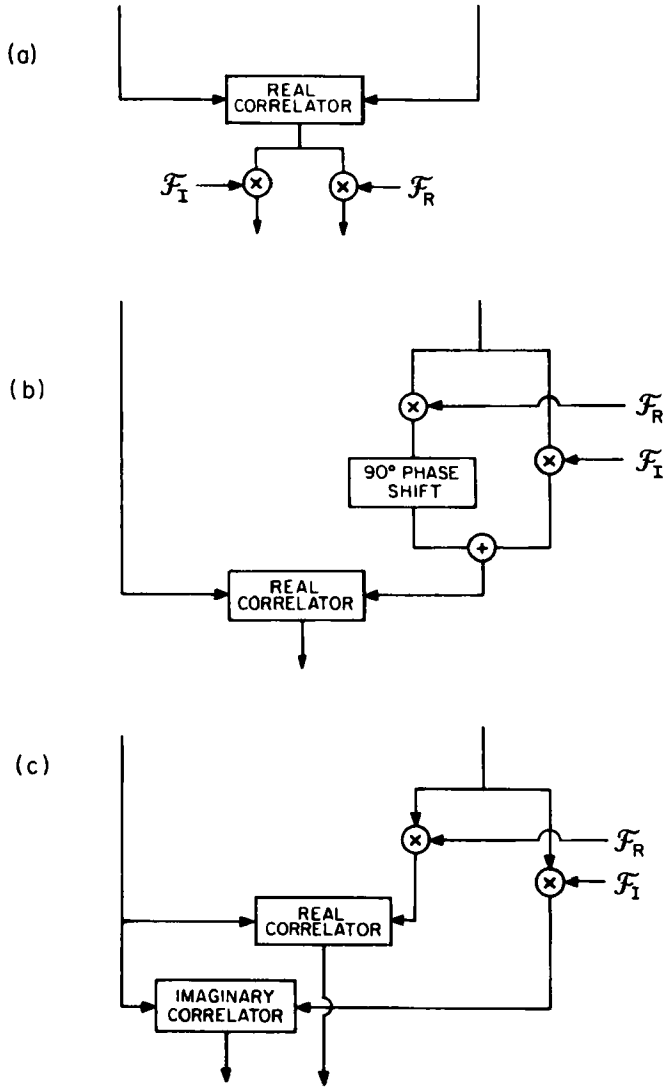


Figure 9.17 Various processor configurations showing possible locations of fringe rotator. $\mathcal{F}_R$ and $\mathcal{F}_I$ are cosine and sine representations of the fringe function. See text for discussion of relative merits.

function is zero, the correlator is inhibited. Since the real and imaginary parts of $\mathcal{F}$ are never zero simultaneously, all data bits are used at least once. This fringe rotation function can be thought of as a phasor whose tip traces out a square such that it has phase jumps in 45° increments and its amplitude alternates between $\sqrt{2}$ and 1. The resulting jitter in phase is uniformly distributed between $\pm 22.5°$ and results in a loss of signal amplitude of $\sin(\pi/8)/(\pi/8) = 0.974$. Also, the

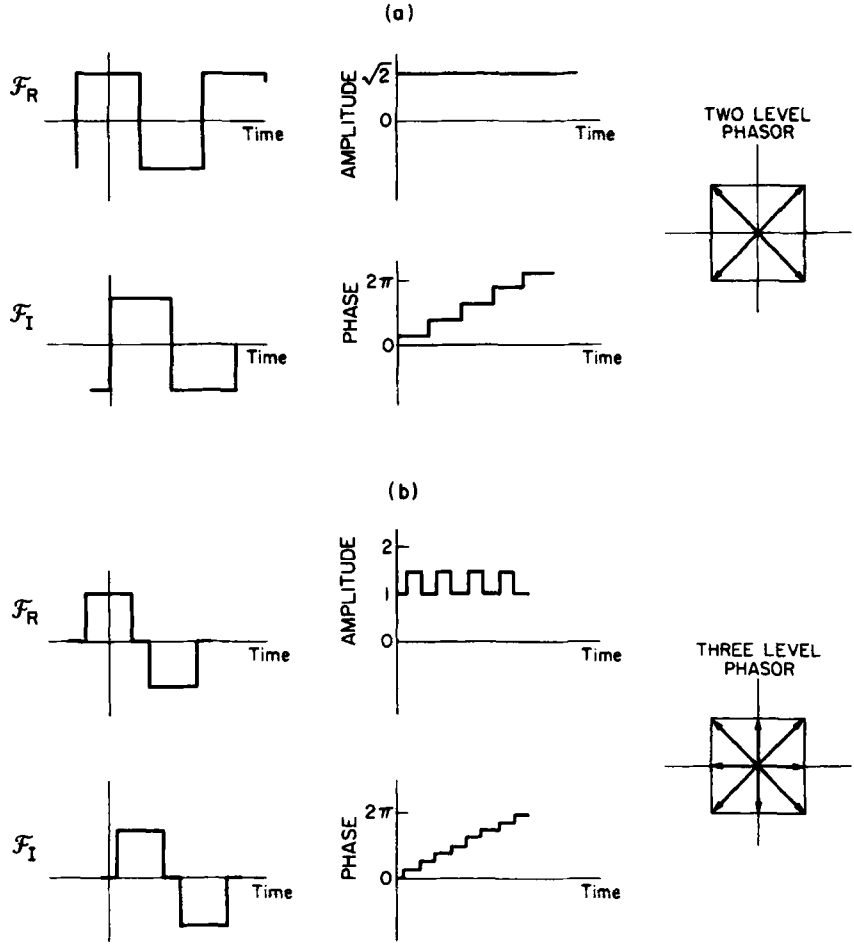


Figure 9.18 (a) Mathematical model of two-level fringe rotator showing $\mathcal{F}_R$ and $\mathcal{F}_I$, functions that approximate $\cos\phi$ and $\sin\phi$ (left), the amplitude and phase representation of $\mathcal{F}$ (center), and the phasor plot of $\mathcal{F}$ (right). (b) Same plots for a three-level fringe rotator.

variation in the amplitude of the phasor introduces a nonuniform weighting of the signal samples. This reduces the signal-to-noise ratio by a further factor equal to $(1 + \sqrt{2})/\sqrt{6} = 0.986$. The net loss in signal-to-noise ratio is 0.960. The reduction in signal-to-noise ratio is also equal to the square root of the ratio of the power in the first harmonic to the total power in $\mathcal{F}_R$. The first harmonic of $\mathcal{F}_R$ is $(4/\pi)\cos(\pi/8) = 1.18$, which is the scale factor correction for the visibility. The three-level fringe function considered here is used in many VLBI processors. The fringe period is divided into 16 parts to generate $\mathcal{F}$. The transitions in $\mathcal{F}$, which then occur at integral multiples of $1/16$ of the fringe period, are not optimally located, but this approximation results in no more than 0.1% additional loss. Note

that an FX correlator can be made to accept input data with more than one or two bits per sample rather more easily than a lag correlator. With more data bits per signal sample, more accurate representations of sine and cosine functions can be used.

Fringe Sideband Rejection Loss (η_S)

The digital fringe rotator shown in Fig. 9.17c is not a single-sideband mixer. Thus, as well as the wanted output, shifted in frequency by the fringe frequency, an unwanted component of noise corresponding to the image response of a mixer also appears. To understand the effect of this noise, consider the cross power spectrum of the correlator output. Recall that ν' is the intermediate frequency defined following Eq. (9.13), and note that in the output of a spectral correlator $\nu' > 0$ and $\nu' < 0$ refer to the upper and lower sidebands, respectively. For upper-sideband operation, the cross power spectrum of the signal is given by Eq. (9.21), which is nonzero only for the upper sideband. However, there will be noise at both positive and negative frequencies. Thus, the cross power spectrum of the correlator output is

$$\mathcal{S}'_{12}(\nu') = \begin{cases} \mathcal{S}(\nu')e^{j\Phi(\nu')} + n_u(\nu'), & \nu' > 0 \\ n_\ell(\nu'), & \nu' < 0, \end{cases} \tag{9.146}$$

where $\mathcal{S}(\nu')$ is the instrumental response defined in Eq. (9.14), $j\Phi(\nu')$ is the exponent in Eq. (9.21), and n_u and n_ℓ are the noise spectra for the upper- and lower-sideband responses. For observations in which a spectral line correlator is used, $\mathcal{S}'_{12}(\nu')$ is computed and the noise at $\nu' < 0$ is simply ignored. For continuum observations using a correlator with only a small number of channels (lags), the noise at $\nu' < 0$ contributes excess noise in the correlation function and must be removed. A straightforward way to remove the noise at $\nu' < 0$ is to compute $\mathcal{S}'_{12}(\nu')$ and multiply it by the filtering function

$$H_F(\nu') = \begin{cases} 1, & 0 < \nu' < \Delta\nu \\ 0, & \text{elsewhere}\,. \end{cases} \tag{9.147}$$

The resulting function, $\mathcal{S}'_{12}(\nu')H_F(\nu')$, can be Fourier transformed back into a correlation function. Alternatively, the filtering can be applied by convolving the correlation function at the output of the correlator with the Fourier transform of $H_F(\nu')$, which is

$$h_F(\tau) = \Delta\nu e^{j\pi\Delta\nu\tau}\left(\frac{\sin\pi\Delta\nu\tau}{\pi\Delta\nu\tau}\right), \tag{9.148}$$

or

$$h_F(\tau) = F_1(\tau) + jF_2(\tau), \tag{9.149}$$

where F_1 and F_2 are as defined in Eq. (9.18). The convolution leaves the desired signal unchanged but removes the negative (lower) sideband noise. Thus,

the resulting correlation function still has the form of Eq. (9.20), plus the positive (upper) sideband noise that cannot be removed.

The role of $h_F(\tau)$ can be understood in a different way. The correlation function at the output of the correlator is computed at discrete delays at intervals of $(2\Delta\nu)^{-1}$. Therefore the correlation function in Eq. (9.20) has a full width at half maximum of about three delay steps. In order to estimate the amplitude and phase of the correlation function, one would like to do more than just take these values from the peak of $\rho'_{12}(\tau)$. Rather, one would like to use all the information provided by the correlation function at various delays. $h_F(\tau)$ is the appropriate interpolation function that properly weights the correlation function, gathering up the power at different delays to provide an optimal estimate of the fringe amplitude, phase, and delay. Note that $h_F(\tau)$ and $\rho'_{12}(\tau)$ are identical forms except for the unknown amplitude, phase, and delay. These unknown quantities can be estimated by the usual procedure of matched filtering or, equivalently, least-mean-squares analysis in which the correlation function is convolved with $h_F(\tau)$. However, $\rho'_{12}(\tau)$ is measured only over a finite number of delay steps, and some information is lost, so the signal-to-noise ratio is reduced. Assume that the system lowpass response is rectangular and the delay errors $\Delta\tau_g$ and τ_e are zero, so that the correlation function is centered in the delay range of the correlator. Let M be the number of delay steps (lags) in the correlator. The loss factor η_S is the signal-to-noise ratio when M values of the correlation function are available, divided by the signal-to noise ratio when the entire function is available:

$$\eta_S = \sqrt{\frac{\sum_{k=-M'}^{M'} |h_F(\tau_k)|^2}{\sum_{k=-\infty}^{\infty} |h_F(\tau_k)|^2}}, \tag{9.150}$$

where $\tau_k = k/2\Delta\nu$, $M' = (M-1)/2$, and M is an odd integer. The denominator in Eq. (9.150) equals $2\Delta\nu^2$ [e.g., see Eq. (A8.5)], so

$$\eta_S = \sqrt{\frac{1}{2} + \sum_{k=1}^{M'} \left[\frac{\sin(\frac{\pi k}{2})}{\frac{\pi k}{2}}\right]^2}. \tag{9.151}$$

For $M = 1$, $\eta_S = 1/\sqrt{2}$, which corresponds to the case of no image rejection. M must be at least 3 to ensure that the peak of the correlation function can be determined; $M \simeq 7$, for which $\eta_S = 0.975$, is adequate for most purposes. Note that because we assumed the correlation function was exactly centered, its value will be zero at delay steps $2, 4, 6, 8, \ldots$, and so on. This suggests, for example, that a 9-delay correlator ($M' = 4$) is no better than a 7-delay correlator ($M' = 3$). In practice, the 9-delay correlator is better because the correlation function is rarely aligned perfectly in the correlator. In general, η_S is slightly smaller than given in Eq. (9.151) if the correlation function is not perfectly aligned (Herring 1983).

Discrete Delay Step Loss (η_D)

The delay introduced to align the bit streams is quantized at the sampling rate, which we assume to be the Nyquist rate. Thus there is a periodic sawtooth delay error with a peak-to-peak amplitude equal to the sampling period. This effect is also known as the *fractional bit-shift error*. The delay error gives rise to a periodic phase shift that is a function of the baseband frequency, as shown in Fig. 9.19. The phase error has a peak-to-peak value of

$$\phi_{pp} = \frac{\pi \nu'}{\Delta \nu}, \tag{9.152}$$

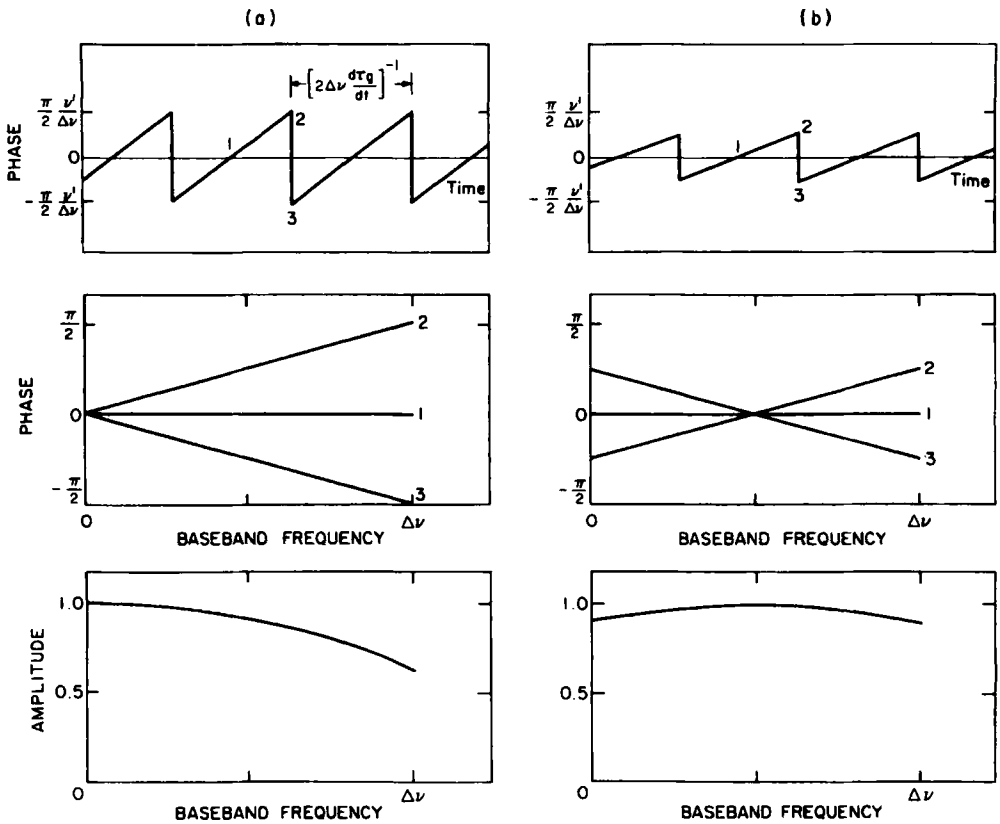


Figure 9.19 Discrete delay step effect. Case (a) applies when the fringe rotator corrects the phase for zero baseband frequency, and case (b) applies when the fringe rotator also inserts a $\pi/2$ phase shift when the delay changes by one Nyquist sample. The top plots show the phase versus time at baseband frequency ν'. The middle plots show the phase across the baseband at three different times denoted by 1, 2, and 3. The bottom plots show the average amplitude across the baseband.

and the sawtooth frequency is proportional to the fringe frequency and has a maximum value of

$$\nu_{\mathrm{ds(max)}} = \frac{2\Delta\nu D\omega_e}{c} \text{ (delay steps per second)}, \tag{9.153}$$

where D is the baseline length and ω_e is the angular velocity of the earth's rotation in radians per second. If nothing is done to correct for this effect and the fringe amplitude is averaged over many times $1/\nu_{\mathrm{ds}}$, then the phase at any frequency ν' is uniformly distributed over ϕ_{pp}. The amplitude loss as a function of baseband frequency is

$$L(\nu') = \frac{\int_0^{\phi_{\mathrm{pp}}/2} \cos(\phi_{\mathrm{pp}}/2)d\phi}{\int_0^{\phi_{\mathrm{pp}}/2} d\phi} = \frac{\sin(\phi_{\mathrm{pp}}/2)}{\phi_{\mathrm{pp}}/2}, \tag{9.154}$$

and the net signal-to-noise reduction over a baseband response of width $\Delta\nu$ is, using Eqs. (9.152) and (9.154),

$$\eta_D = \frac{1}{\Delta\nu}\int_0^{\Delta\nu} \frac{\sin(\pi\nu'/2\Delta\nu)}{\pi\nu'/2\Delta\nu} d\nu' = 0.873. \tag{9.155}$$

Unless the fringe amplitude averaging is done over an integral number of fringe periods there is also a residual phase error, the amplitude of which decreases with the number of periods. When the fringe frequency is near zero this phase error can be significant.

The effect of the discrete delay step can be compensated, and no sensitivity loss need occur. The delay error caused by delay quantization is a known quantity that introduces a phase slope in the cross power spectrum. Therefore, if the cross power spectra are calculated on a period short with respect to $1/\nu_{\mathrm{ds}}$, which can be as small as 20 ms on a 5000-km baseline with $\Delta\nu = 20$ MHz [see Eq. (9.153)], then the effect of the discrete delay step can be removed by adjusting the slope of the phase of the cross power spectrum. This correction is easily done in spectral line work where spectra are calculated anyway. Note that if this correction is not made, the sensitivity loss factor is 0.64 at the high-frequency edge of the band, as given by Eq. (9.154). In this case, the amplitude response should be compensated by dividing the cross power spectra by $L(\nu')$. In continuum work, the correction is sometimes omitted because of the need to Fourier transform to the frequency domain and then back to cross-correlation.

A way to compensate partially for the effect of discrete delay steps is to move the frequency at which the phase is unperturbed from zero to $\Delta\nu/2$, the baseband center. The phase of the fringe rotator is increased by $\pi\Delta\nu\Delta\tau_s$, where $\Delta\tau_s$ is the delay error. Thus, when the delay changes by one sampling interval, a phase jump of $\pi/2$ is inserted in the fringe rotator. The resulting loss at the band edges is then only 0.90. The average loss over the band is given by an equation similar to Eq. (9.155), but with the upper limit of integration changed to $\Delta\nu/2$, and

equals 0.966. Also, for a symmetrical bandpass response, the residual phase error is zero because the net phase shift over the band at any instant is zero.

Summary of Processing Losses

The loss factors that we have considered are all multiplicative, so the total loss is given by the equation

$$\eta = \eta_Q \eta_R \eta_S \eta_D, \tag{9.156}$$

where η_Q = quantization loss, η_R = fringe rotation loss, η_S = fringe sideband rejection loss, and η_D = discrete delay step loss.

If there are fringe rotators in each signal path to the correlator, the fringe rotation loss will be η_R^2 because the fringe rotator phases will be uncorrelated. A summary of the loss factors is given in Table 9.6. As an example, a processor might have two-level sampling ($\eta_Q = 0.637$), three-level fringe rotators in each signal path ($\eta_R = 0.922$), 11-channel correlation function ($\eta_S = 0.983$), and band-center delay compensation ($\eta_D = 0.966$), giving a net loss of 0.558. Thus the sensitivity is worse than that of an ideal analog system with the same bandwidth by a factor of about 2.

There are other loss factors that we have not discussed here. The passband will not in reality be perfectly flat, or the response zero for frequencies above half the Nyquist sampling frequency. These imperfections introduce loss, which for an ideal nine-pole Butterworth filter amounts to 2% (Rogers 1980). The frequency

TABLE 9.6 Signal-to-Noise Loss Factors

1. *Quantization Loss* (η_Q)[a]	
(a) Two-level	0.637
(b) Three-level	0.810
(c) Four-level, all products	0.881
2. *Fringe Rotation Loss* (η_R)	
(a) Two-level, one path	0.900
(b) Three-level, one path	0.960
(c) Two-level, both paths	0.810
(d) Three-level, both paths	0.922
3. *Fringe Sideband Rejection Loss* (η_S)	
(a) 1 channel	0.707
(b) 3 channels	0.952
(c) 7 channels	0.975
(d) 11 channels	0.983
4. *Discrete Delay Step Loss* (η_D)	
(a) Spectral correction	1.000
(b) Baseband center correction	0.966
(c) No correction	0.873

[a] See Section 8.3.

TABLE 9.7 Normalization Factors[a]

1. *Quantization*[b]	
(a) Two-level	1.57
(b) Three-level	1.23
(c) Four-level	1.13
2. *Fringe Rotation*	
(a) Two-level, one path	0.786
(b) Three-level, one path	0.850
(c) Two-level, both paths	0.617
(d) Three-level, both paths	0.723

[a] Multiply correlator output by listed value to obtain normalized correlation function.
[b] See Section 8.3.

responses will not be perfectly matched for different antennas (see Section 7.3). The phase settings of the fringe rotator may be calculated exactly at convenient intervals and extrapolated by Taylor series; this approximation will introduce periodic phase jumps. The local oscillators may have power-line harmonic and noise sidebands that put some fringe power outside the usual fringe filter passband. Empirical values of η typical of the first decade of VLBI development were about 0.4 (Cohen 1973).

The η values refer to loss in signal-to-noise ratio. The fringe amplitudes must also be corrected for scale changes due to signal quantization and fringe rotation. We summarize the multiplicative normalization factors to be applied to the fringe amplitudes in Table 9.7.

9.8 BANDWIDTH SYNTHESIS

For geodetic and astrometric purposes it is useful to measure the geometric group delay

$$\tau_g = \frac{1}{2\pi}\frac{\partial\phi}{\partial\nu} \tag{9.157}$$

as accurately as possible. With a single RF band, the delay can be found by fitting a straight line to the phase versus frequency of the cross power spectrum. The uncertainty in this delay, from the usual application of least-mean-squares analysis, is

$$\sigma_\tau = \frac{\sigma_\phi}{2\pi\,\Delta\nu_{\rm rms}}, \tag{9.158}$$

where σ_ϕ is the rms phase noise for a bandwidth $\Delta\nu$ and $\Delta\nu_{\rm rms}$ is the rms bandwidth, which for a single band of width $\Delta\nu$ is equal to $\Delta\nu/(2\sqrt{3})$ (see Appendix 12.1). σ_ϕ can be obtained from Eq. (6.64), and if processing losses are neglected,

Eq. (9.158) becomes

$$\sigma_\tau = \frac{T_S}{\zeta T_A \sqrt{\Delta\nu_{\text{rms}}^3 \tau}}, \tag{9.159}$$

where ζ is a constant equal to $\pi(768)^{1/4} \simeq 16.5$ [see derivation of Eq. (A12.33)], and T_S and T_A are the geometric mean system and antenna temperatures. A much higher value of $\Delta\nu_{\text{rms}}$ can be realized by observing at several different radio frequencies. This can be accomplished by switching the local oscillator of a signal-band system sequentially in time among N frequencies, or by dividing up the recorded signal into N simultaneous RF bands (channels), which are spread over a wide frequency interval. The temporal switching method has the disadvantage that phase changes during the switching cycle degrade or bias the delay estimate. These methods are commonly referred to as bandwidth synthesis (Rogers 1970, 1976).

In a practical system, signals from a small number of RF bands ($\sim$10) are recorded. The problem of determining the optimum distribution of these bands in frequency is similar to the problem of finding a minimum-redundancy distribution of antenna spacings in a linear array, as discussed in Section 5.5. However, here we do not need to have all multiples of the unit (frequency) spacing up to the maximum value, and some gaps are not necessarily detrimental. From the spectral point of view, we wish to have the bands placed in some geometric sequence of increasing separation so that phase can be extrapolated from one band to the next, as shown in Fig. 9.20, without having any 2π ambiguities in the phase connection process. The rms bandwidth depends critically on the unit spacing, which depends on the minimum signal-to-noise ratio. The delay accuracy for a multiband system is obtained from Eq. (9.158) in the same way as for Eq. (9.159) but without the condition $\Delta\nu_{\text{rms}} = \Delta\nu/(2\sqrt{3})$. Thus we obtain

$$\sigma_\tau = \frac{T_S}{2\sqrt{2}\pi T_A \sqrt{\Delta\nu\tau}\,\Delta\nu_{\text{rms}}}, \tag{9.160}$$

where $\Delta\nu_{\text{rms}}$ for a typical bandwidth synthesis system is approximately 40% of the total frequency interval spanned, $\Delta\nu$ is the total bandwidth, and τ is the integration time for each band. To avoid explicitly the problem of phase connection, we can form an equivalent delay function from the cross power spectra [see Eq. (9.21)] of the various bands observed:

$$D_R(\tau) = \sum_{i=1}^{N} \int_0^{\Delta\nu} \mathcal{S}_{12i}(\nu - \nu_i) e^{j2\pi\nu\tau}\, d\nu, \tag{9.161}$$

where the ν_i are the local oscillator frequencies relative to the lowest one, and $\nu - \nu_i$ is the baseband frequency. The maximum of $|D_R(\tau)|$ gives the maximum-likelihood estimate of the interferometer delay (Rogers 1970). The a priori normalized delay resolution function, obtained from Eq. (9.161) by setting $\mathcal{S}_{12} = 1$

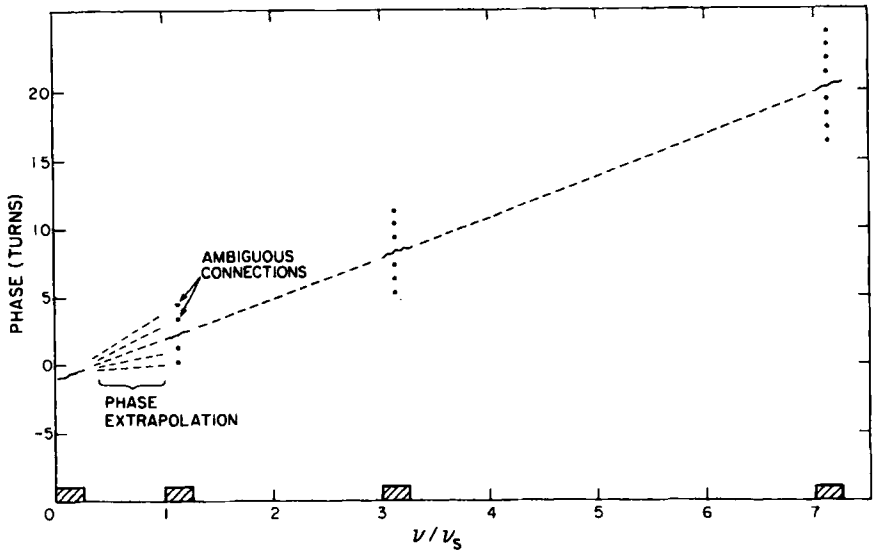


Figure 9.20 Fringe phase versus frequency for a bandwidth synthesis system. The phase is measured over discrete bands (crosshatched) spaced at multiples of the fundamental band separation frequency ν_s. The turn ambiguities give rise to sidelobes in the delay resolution function defined in Eq. (9.161) and shown in Fig. 9.21.

at frequencies where it is measured and $\mathcal{S}_{12} = 0$ otherwise, is

$$|D_R(\tau)| = \Delta\nu \frac{\sin \pi \Delta\nu\tau}{\pi \Delta\nu\tau} \left| \sum_{i=1}^{N} e^{j2\pi\nu_i\tau} \right| . \tag{9.162}$$

The sinc-function envelope is the delay resolution function for a single channel. The frequencies ν_i should be chosen to minimize the width of $D_R(\tau)$ while not allowing any subsidiary maximum to rise above a level such that it could be confused with the principal peak. In situations with low signal-to-noise ratio, the minimum unit spacing should be about four times the bandwidth of a single channel. The delay resolution function for a five-channel system is shown in Fig. 9.21.

Burst Mode Observing

For certain observations there are advantages in limiting the observing time to short bursts during which the bit rate can be much higher than the mean data acquisition rate as limited by the tape recorder [see, e.g., Wietfeldt and Frail (1991)]. In pulsar observations the duration of the pulsed emission is typically ~3% of the total time, so by recording data taken only during pulsar-on time the bandwidth can be increased by a factor of ~33 over the maximum bandwidth for continuous observation. This technique requires the use of a high-speed sampler, high-speed memory, and pulse-timing circuitry at each antenna. During the pulse the data are

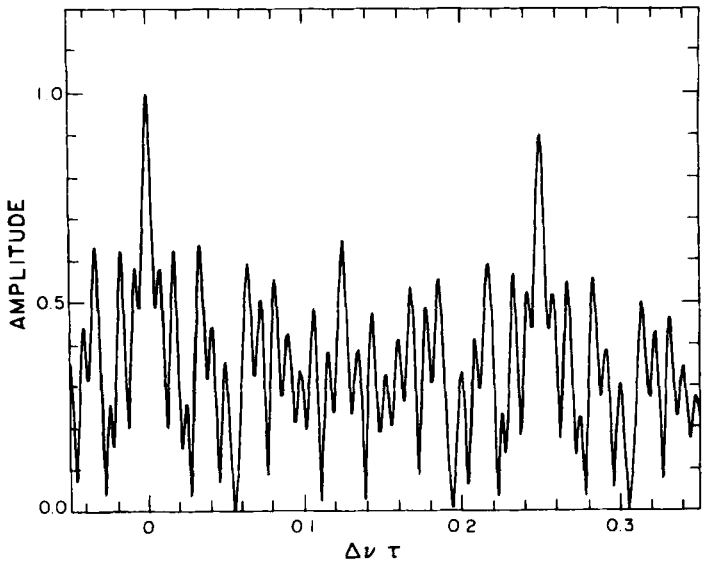


Figure 9.21 Delay resolution function for five-channel system with a unit spacing $\nu_s = 4\Delta\nu$ and spacing of 0, 1, 3, 7, and $15\nu_s$, as shown in part in Fig. 9.20. The "grating" lobe at $\tau\Delta\nu = 0.25$ need only be reduced sufficiently below unity to avoid delay ambiguity.

then read out continuously at a lower rate. If the ratio of these two rates is a factor w, then the bandwidth can be increased by the same factor over constant-rate observing. For pulsars this results in an increase in sensitivity by a factor w, of which $\sqrt{w}$ can be attributed to the increased bandwidth, and $\sqrt{w}$ to the fact that noise is not being recorded during the pulse-off time. The second of these $\sqrt{w}$ factors can be obtained without an increase in the data rate by simply deleting data during the pulse-off periods. Burst mode observing is also useful for astrometry and geodesy because it increases the accuracy of measurement of the geometric delay, and it has been used for this purpose in observations of continuum sources at millimeter wavelengths.

9.9 PHASED ARRAYS AS VLBI ELEMENTS

A phased array is a series of antennas for which the received signals are combined, as indicated in Fig. 5.4. The phase and delay of the signal from each antenna can be adjusted so that the signals from a particular direction in the sky combine in phase, thereby maximizing the sensitivity. It is important to consider the use of phased arrays as VLBI elements for two reasons. First, the elements of a connected-element synthesis array can be combined to form a phased array, thus improving the signal-to-noise ratio of a very-long-baseline interferometer in which they participate as a single station. Second, if elements with very large col-

lecting area are desired to achieve a high signal-to-noise ratio on each baseline, it may be advantageous to build phased arrays rather than monolithic antennas because the cost of a parabolic reflector antenna increases approximately as the diameter to the power 2.7 (Meinel 1979).

Synthesis arrays such as the Westerbork Array, the VLA, and several others are also used as phased arrays to provide a large collecting area for one element in a VLBI system. Phasing the array consists of adjusting the phase and delay of the signal from each antenna so as to compensate for the different geometric paths for a wavefront from the desired direction. These corrections are easily made through the delay and fringe rotation systems that are used for synthesis imaging. The signals are then summed and go to a VLBI recorder.

We can analyze the performance of a phased array that is used to simulate a single large antenna. Consider an array of n_a identical antennas for which the system temperature is T_S and the antenna temperature for a given source that is unresolved by the longest spacings in the array is T_A. The output of the summing port is

$$V_{\text{sum}} = \sum_i (s_i + \epsilon_i), \tag{9.163}$$

where s_i and ϵ_i represent the random signal and random noise voltages, respectively, from antenna i. Now $\langle s_i \rangle = \langle \epsilon_i \rangle = 0$ and, omitting constant gain factors, we can write $\langle s_i^2 \rangle = T_A$ and $\langle \epsilon_i^2 \rangle = T_S$. The power level of the combined signals is represented as the average squared value of Eq. (9.163),

$$\langle V_{\text{sum}}^2 \rangle = \sum_{i,j} [\langle s_i s_j \rangle + \langle s_i \epsilon_j \rangle + \langle s_j \epsilon_i \rangle + \langle \epsilon_i \epsilon_j \rangle]. \tag{9.164}$$

If the array is accurately phased, $s_i = s_j$. Also, since we are considering an unresolved source, $\langle s_i s_j \rangle = T_A$. If the array is unphased, that is, if the signal phases at the combination point are random, then $\langle s_i s_j \rangle = T_A$ only for $i = j$ and is otherwise zero. In either case $\langle s_i \epsilon_i \rangle = 0$ and $\langle \epsilon_i \epsilon_j \rangle = 0$. Thus Eq. (9.164) can be reduced to

$$\langle V_{\text{sum}}^2 \rangle = n_a^2 T_A + n_a T_S \qquad \text{(array phased)} \tag{9.165}$$

$$\langle V_{\text{sum}}^2 \rangle = n_a T_A + n_a T_S \qquad \text{(array unphased)}, \tag{9.166}$$

where the first term on the right-hand side represents the signal and the second term represents the noise. When the array is phased the signal-to-noise (power) ratio is $n_a T_A / T_S$, and when it is unphased it is T_A / T_S. Thus the collecting area of the phased array is equal to the sum of the collecting areas of the individual antennas, but when it is unphased it is, on average, equal to that of a single antenna.

A question of interest concerns the case where the antennas have different sensitivities resulting from different effective collecting areas and/or system temperatures. This is a matter of practical importance even for nominally uniform arrays, since maintenance or upgrading programs can result in differences in sensitivity. Consider a phased array in which the individual system temperatures and antenna

temperatures are represented by T_{Si} and T_{Ai}, respectively. Here T_{Ai} is defined as the signal from a point source of *unit* flux density*, so T_{Ai} is a characteristic of the antenna alone, and is proportional to the collecting area. We consider only the small-signal case for which $T_A \ll T_S$. For antenna i the output voltage from a source of flux density S is $V_i = s_i + \epsilon_i$ and we can write $\langle s_i^2 \rangle = ST_{Ai}$ and $\langle \epsilon_i^2 \rangle = T_{Si}$.

It is convenient to think of the output of each antenna as providing a measure of the flux density of the source which is equal to V_i^2/T_{Ai}. The expectation of the measured value of S should be the same for each antenna. The corresponding voltages are $\sqrt{S} = V_i/\sqrt{T_{Ai}}$ for the signal and $\epsilon_i/\sqrt{T_{Ai}}$ for the noise. In the cross-correlation of the array output with another VLBI antenna, the signal-to-noise ratio at the correlator output is proportional to the signal-to-noise *voltage* ratio of the signal from the array. Thus, in combining the signal voltages in the array, we are, in effect, interested in maximizing the signal-to-noise ratio in an estimate of $\sqrt{S}$. Because the array antennas are not identical, we should use weighting factors w_i in combining their signals. The weights should be chosen to maximize the signal-to-noise ratio of the combined array signals which, in voltage, is

$$\mathcal{R}_{\rm sn} = \sum_i \frac{w_i V_i}{\sqrt{T_{Ai}}} \bigg/ \sqrt{\sum_i \frac{w_i^2 T_{Si}}{T_{Ai}}}. \tag{9.167}$$

Note that we add the signal voltages and the squares of the rms noise voltages. Selecting the weights to provide the best signal-to-noise ratio for $V_i/\sqrt{T_{Ai}}$ is mathematically equivalent to the general problem of obtaining the best estimate of a measured quantity from a series of measurements for which the rms error levels are different, but are known. The optimum procedure is to take a mean in which the weight of each measurement is inversely proportional to the variance of the error of that measurement [see Eq. (A12.6)]. The variance of V_i is proportional to T_{Si}, and thus the variance of $V_i/\sqrt{T_{Ai}}$ is T_{Si}/T_{Ai}. Thus we insert $w_i = T_{Ai}/T_{Si}$ in Eq. (9.167) and obtain

$$\begin{aligned} \mathcal{R}_{\rm snl} &= \sum_i \frac{V_i}{\sqrt{T_{Ai}}} \frac{T_{Ai}}{T_{Si}} \bigg/ \sqrt{\sum_i \frac{T_{Si}}{T_{Ai}} \left(\frac{T_{Ai}}{T_{Si}}\right)^2} \\ &= \sum_i \frac{V_i \sqrt{T_{Ai}}}{T_{Si}} \bigg/ \sqrt{\sum_i \frac{T_{Ai}}{T_{Si}}}. \end{aligned} \tag{9.168}$$

Note that in the numerator V_i is multiplied by $\sqrt{T_{Ai}}/T_{Si}$, which is therefore the (voltage) weighting factor for optimum sensitivity in the signal combination. This conclusion is in agreement with an analysis by Dewey (1994). (Note that the

*Since it is only the relative values of the weighting factors that matter, T_{Ai} could be defined with respect to any source that is common to all antennas, but consideration of unit flux density simplifies the explanation.

weighting factors for the signal voltages at the combination point are not w_i but $w_i/\sqrt{T_{Ai}}$.) The corresponding weighting of the signal *power* at the combination point is proportional to T_{Ai}/T_{Si}^2.

In synthesis arrays such as the VLA, the IF signals from the antennas are each delivered to a digital sampler at the same power level (of signal plus noise), and the signals are combined after that point so that the time delays required can be inserted digitally. Thus, to avoid modifying the receiving system (which is designed for synthesis mapping), the signals are combined with equal powers when the array is used in the phased mode. For the case of $T_A \ll T_S$ that we are considering, the corresponding weighting is $w_i = 1/\sqrt{T_{Si}}$, and the signal-to-noise ratio becomes

$$\mathcal{R}_{\mathrm{sn2}} = \sum_i \frac{V_i}{\sqrt{T_{Ai}T_{Si}}} \Bigg/ \sqrt{\sum_i \frac{1}{T_{Ai}}}. \tag{9.169}$$

Equal-power weighting usually provides sensitivity within a few percent of optimum weighting.

With optimum weighting in the signal combination, all antennas make some contribution to increasing the signal-to-noise ratio. With other weighting, the overall sensitivity may be improved by omitting antennas with poor performance. Moran (1989) has investigated this effect for equal-power weighting. To simplify the situation it was assumed that T_A is the same for all antennas and only T_S varies. Consider an array undergoing an upgrade of the receiver input stages, in which a fraction n_1/n_a have been refitted with new input stages that reduce the system temperature from T_S to T_S/ξ. After a certain fraction of the antennas have been refitted, the array sensitivity is improved by omitting the unimproved antennas because their input stages are noisier. When T_A does not vary, we can represent the signal voltage received by each antenna by V, and Eq. (9.169) for equal-power weighting becomes

$$\mathcal{R}_{\mathrm{sn2}} = \frac{V}{\sqrt{N}} \sum_i \frac{1}{\sqrt{T_{Si}}}. \tag{9.170}$$

Thus we can write

$$\frac{\mathcal{R}_{\mathrm{sn2}}(n_1 \text{ refits only})}{\mathcal{R}_{\mathrm{sn2}}(\text{ all } n_a \text{ antennas})} = \frac{1}{\sqrt{n_1}}\left(\frac{n_1\sqrt{\xi}}{\sqrt{T_S}}\right) \Bigg/ \frac{1}{\sqrt{n_a}}\left(\frac{n_1\sqrt{\xi}}{\sqrt{T_S}} + \frac{n_a - n_1}{\sqrt{T_S}}\right). \tag{9.171}$$

The unimproved antennas should be omitted if the expression above is greater than unity, which occurs for

$$\frac{n_1}{n_a} > \left(\frac{\sqrt{\xi}}{2} + \sqrt{1 - \sqrt{\xi} + \frac{\xi}{4}}\right)^{-2}. \tag{9.172}$$

Figure 9.22 shows n_1/n_a as a function of ξ. Thus, for example, if the refitting reduces T_S by a factor of six, then when about half the antennas have been refitted

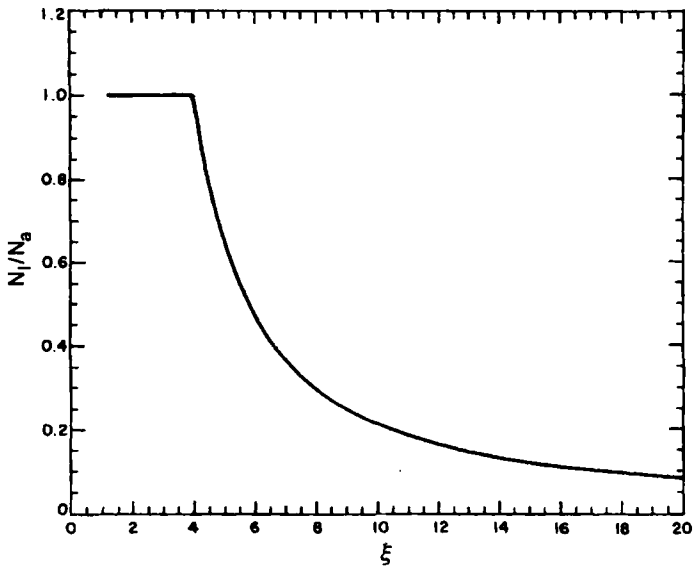


Figure 9.22 The fraction of antennas, n_l/n_a, in a phased array with equal-power weighting, for which the system temperature must be reduced by a factor ξ before the remaining antennas should be omitted. From Moran (1989), ©1989 by Kluwer Academic Publishers, reproduced with permission.

the others should be omitted. However, unless $\xi > 4$ all antennas should be retained. In practice, a factor of four would be an unusually big improvement, so it can be concluded that omitting antennas is rarely useful. A similar analysis based on Eq. (9.168) shows that with optimum weighting the sensitivity is never improved by omitting antennas.

For tape recording in VLBI, the output of a phased array is usually requantized to reduce the number of bits. The first quantization of the signals, before they are combined, introduces quantization noise which, after combination, has a probability distribution that tends to Gaussian as the number of antennas becomes large. Thus for such arrays, the additional loss in sensitivity in requantizing is close to the values of η_Q derived in Chapter 8, for which Gaussian noise is assumed. For other cases, see Kokkeler, Fridman, and van Ardenne (2001).

9.10 ORBITING VLBI (OVLBI)

The basic requirements for a VLBI station, whether orbiting or terrestrial, include a timing system so that the time associated with each digital sample of the received signal is recoverable, and a position for the antenna known with sufficient accuracy that the fringe frequency (but not necessarily the fringe phase) can be determined. The timing system must be stable to a fraction of the period of the

received signal frequency over a coherence time of tens or hundreds of seconds. If it is not possible to put a precise frequency standard on a satellite, then a timing link of equivalent stability must be implemented. Establishing this timing system, which provides the local oscillators and the sampling clock at the satellite, is a major technical challenge in OVLBI. The radial motion of the satellite introduces Doppler shifts and the tangential motion causes the link path to move relative to the atmospheric irregularities. One or more reference frequencies are transmitted to the satellite over a radio link. The position of the satellite at any time is known from standard orbit-tracking procedures to an accuracy of some tens of meters. This is sufficient to determine the (u, v) coordinates of the baseline, but not sufficient for the timing accuracy required. To solve the timing problem a round-trip phase system implemented by radio link is required. This is identical in principle to the round-trip systems for cables discussed in Section 7.2. A discussion of the basic requirements of the timing system is given by D'Addario (1991).

Figure 9.23 shows a simplified example of a system at the satellite and earth station, which illustrates the essential functions. In this case a frequency standard is not included in the satellite. A frequency standard in the earth station provides a reference frequency to synthesizer S_X from which a signal is transmitted to the satellite. This signal provides a reference for synthesizers S_y, S_L, and S_s that produce signals for the round-trip phase measurement, the local oscillator (LO)

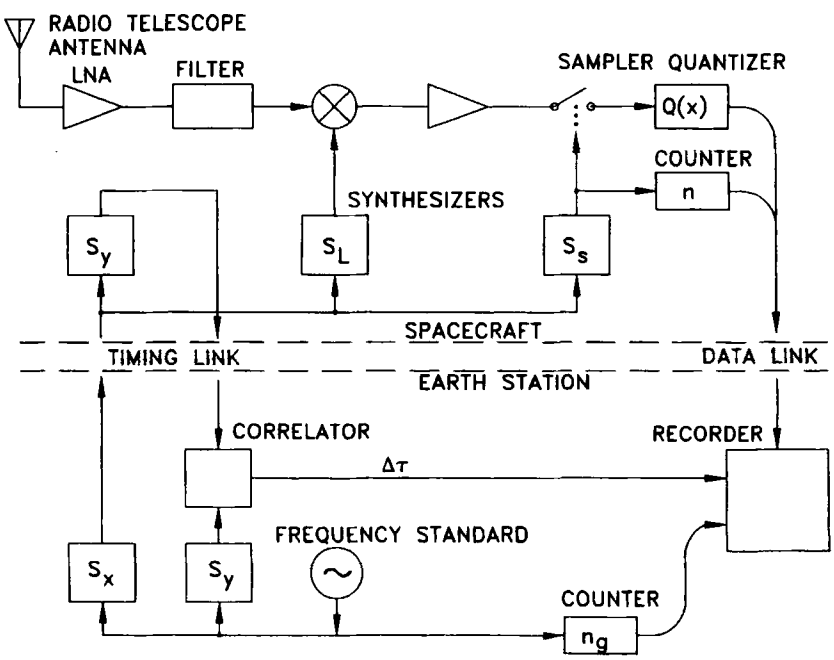


Figure 9.23 Simplified block diagram of the basic signal transmission and processing required on an OVLBI spacecraft and at the earth station. See text for further explanation. From D'Addario (1991), ©1991 IEEE.

of the radio astronomy receiver, and the sampling clock, respectively. The signal from S_y is radiated to the earth station, where its phase is compared in a correlator with a locally generated signal at the same frequency. The correlator output is a measure of $\Delta\tau$, the change in the time delay of the round-trip path. The signal from the radio telescope on the spacecraft goes to a low-noise amplifier (LNA), a filter, and a mixer in which it is converted to intermediate frequency (IF) by the LO signal from S_L. The IF signal then goes to an IF amplifier, a sampler (represented by a switch), and a quantizer, Q(x). The counter n is driven by the sampler clock signal from synthesizer S_s and provides timing signals. These provide a record of when each data point was taken, information for formatting the data, and other timing functions required on the satellite. The counter n_g provides timing at the ground location. Some complications with the operation of the scheme just outlined are

1. The round-trip phase measures the length of the round-trip path with an ambiguity of an integral number of wavelengths. It provides a measure of changes in path length that are continuous.
2. Unless the frequencies generated by the three synthesizers at the satellite are harmonics of one or more reference frequencies supplied (so that no frequency division is necessary in the synthesizers), then the phases of the frequencies will be ambiguous.
3. The transmission times for the reference frequencies and the data may differ because of dispersion in the path or differences in the electronics.

These limitations cause problems when there are discontinuities in the link contact between the satellite and the earth station. If there is continuous contact during an observing period, then once fringes are found the combined effect of the ambiguities is determined. The continuous monitoring of the variation of the path enables the solution to be extended throughout the observing period. However, if signal contact is lost due to interference, atmospheric effects, or equipment problems, phase-locked loops in the synthesizers lose lock and a phase discontinuity will result when the signals are regained. If the round-trip tracking is interrupted for a long period, another fringe search of the data may be required.

The first satellite designed specifically for use as an orbiting element in a VLBI array, the HALCA satellite of Japan came into operation during 1997 (Hirabayashi et al. 1998). As an example, some of the frequencies that were chosen for HALCA are listed as follows:

1. Observing frequencies: 1.7, 5, 22 GHz
2. Total observing bandwidth: 64 MHz (maximum)
3. Reference frequency, earth-to-space: 15.3 GHz
4. Reference frequency, space-to-earth: 14.2 GHz
5. Frequency of data downlink carrier: 14.2 GHz

6. Data rates: 128 Mbits s^{-1} (maximum)
7. Data modulation scheme, QPSK (quadri-phase-shift keying)

In QPSK data modulation the carrier phase takes one of four different values at multiples of 90°, thus representing two bits. The phase is switched between consecutive values at a rate of 64 MHz, or either 32 or 16 MHz if narrower bandwidths are used. Since the data values are essentially random, the carrier phasor averages to zero over many cycles. At the earth station a carrier-frequency oscillator is phase-locked to the received sidebands, and both the digital data streams and the carrier are recovered. In the HALCA satellite the recovered carrier acts as the reference frequency downlink. In phase locking the carrier to the data sidebands a four-phase Costas loop or similar circuit is used [see, e.g., Gardner (1979)]. There is a 90° ambiguity in the phase of the carrier recovered from a QPSK-modulated signal. This ambiguity did not prove to be a serious problem in the HALCA project, but use of a separate downlink reference would eliminate one source of phase discontinuities induced by link dropouts. Since the uplink and downlink frequencies differ, the one-way path cannot be assumed to be exactly equal to half the measured round-trip path. A model of the ionization along the path is required to correct for differences at the two frequencies.

D'Addario (1991) has pointed out that there are ways of designing a system in which the ambiguities could be eliminated. The orbital parameters of the satellite provide an estimate of the round-trip delay within an uncertainty $\delta\tau$, of order 10^{-7} s, and if one were to include a round-trip measurement at a frequency no greater than $\delta\tau^{-1}$ Hz, the ambiguity could be resolved. Round-trip measurements at higher frequencies would still be required to provide sufficient accuracy in the phase of the local oscillator on the satellite, which would be at a frequency similar to that of the radio astronomy signal, that is, possibly tens of gighertz or more. Thus a system in which the ambiguities are eliminated would require round-trip phase measurements at two or more frequencies. For any round-trip measurement, use of the same frequency in both directions would simplify the determination of the one-way propagation time, since the effects of dispersion would be largely eliminated. This would be technically feasible with time sharing or a very small frequency offset to allow signals in the two directions to be separated. However, the international radio regulations usually allocate different frequency bands for the two directions of transmission. Measurement of the round-trip path at two frequencies is therefore important in determining the relative contributions of the neutral and ionized media to the propagation time. If a high-stability frequency standard is included on a satellite it could serve as the primary clock, or as a backup to a radio-link timing system to help keep time at the satellite during link dropouts. Relativistic effects are a complication in the use of an on-board clock, causing its time to vary with respect to earth-station clocks as the satellite moves through regions of differing strength of the earth's gravitational field (Ashby and Allan 1979, Vessot 1991).

A correlator used in OVLBI measurements is in principle the same as one used in terrestrial VLBI, but it must be capable of handling the large Doppler shifts,

time delays, and rates of change of these quantities associated with the space station. A description of a lag-type correlator designed specifically to include OVLBI stations is given by Carlson et al. (1999).

BIBLIOGRAPHY

Biraud, F., Ed., *Very Long Baseline Interferometry Techniques*, Cepadues, Toulouse, France, 1983.

Chi, A. R., Ed., *Proc. IEEE*, Special Issue on Frequency Stability, **54**, No. 2, 1966.

Enge, P. and P. Misra, Eds., *Proc. IEEE*, Special Issue on Global Positioning System, **87**, No. 1, 16–172, 1999.

Felli, M. and R. E. Spencer, Eds., *Very Long Baseline Interferometry, Techniques and Applications*, NATO ASI Series, Kluwer, Dordrecht, 1989.

Hirabayashi, H., M. Inoue, and H. Kobayashi, Eds., *Frontiers of VLBI*, Universal Academy Press, Tokyo, 1991.

Jespersen, J. and D. W. Hanson, Eds., *Proc. IEEE*, Special Issue on Time and Frequency, **79**, No. 7, 1991.

Kroupa, V. F., Ed., *Frequency Stability: Fundamentals and Measurement*, IEEE Press, New York, 1983.

Morris, D., Ed., *Radio Sci.*, Special Issue Devoted to the Open Symposium on Time and Frequency, **14**, No. 4, 1979.

Zensus, J. A., P. J. Diamond, and P. J. Napier, Eds., *Very Long Baseline Interferometry and the VLBA*, Astron. Soc. Pacific Conf. Ser., **82**, 1995.

REFERENCES

Allan, D. W., Statistics of Atomic Frequency Standards, *Proc. IEEE*, **54**, 221–230, 1966.

Ashby, N. and D. W. Allan, Practical Implications of Relativity for a Global Coordinate Time Scale, *Radio Sci.*, **14**, 649–669, 1979.

Bare, C., B. G. Clark, K. I. Kellermann, M. H. Cohen, and D. L. Jauncey, Interferometer Experiments with Independent Local Oscillators, *Science*, **157**, 189–191, 1967.

Barnes, J. A., A. R. Chi, L. S. Cutler, D. J. Healey, D. B. Leeson, T. E. McGunigal, J. A. Mullen, W. L. Smith, R. L. Sydnor, R. F. C. Vessot, and G. M. R. Winkler, Characterization of Frequency Stability, *IEEE Trans. Instrum. Meas.*, **IM-20**, 105–120, 1971.

Berkeland, D. J., J. D. Miller, J. C. Bergquist, W. M. Itano, and D. J. Wineland, Laser-Cooled Mercury Ion Frequency Standard, *Phys. Rev. Lett.*, **80**, 2089–2092, 1998.

Blair, B. E., *Time and Frequency: Theory and Fundamentals*, NBS Monograph 140, U.S. Government Printing Office, Washington, DC, 1974, pp. 223–313.

Broten, N. W., T. H. Legg, J. L. Locke, C. W. McLeish, R. S. Richards, R. M. Chisholm, H. P. Gush, J. L. Yen, and J. A. Galt, Long Baseline Interferometry: A New Technique, *Science*, **156**, 1592–1593, 1967.

Cannon, W. H., The Classical Analysis of the Response of a Long Baseline Radio Interferometer, *Geophys. J. R. Astron. Soc.*, **53**, 503–530, 1978.

Cannon, W. H., D. Baer, G. Feil, B. Feir, P. Newby, A. Novikov, P. Dewdney, B. Carlson, W. P. Petrachencko, J. Popelar, P. Mathieu, R. D. Wietfeldt, The S2 VLBI System, *Vistas in Astronomy*, **41**, 297–302, 1997.

Carlson, B. R., P. E. Dewdney, T. A. Burgess, R. V. Casoro, W. T. Petrachenko, and W. H. Cannon, The S2 VLBI Correlator: A Correlator for Space VLBI and Geodetic Signal Processing, *Pub. Astron. Soc. Pacific*, **111**, 1025–1047, 1999.

Clark, B. G., Radio Interferometers of Intermediate Type, *IEEE Trans. Antennas Propag.*, **AP-16**, 143–144, 1968.

Clark, B. G., The NRAO Tape-Recorder Interferometer System, *Proc. IEEE*, **61**, 1242–1248, 1973.

Clark, B. G., R. Weimer, and S. Weinreb, *The Mark II VLB System*, NRAO Electronics Division Internal Report 118, National Radio Astronomy Observatory, Green Bank, WVA, 1972.

Clark, T. A., B. E. Corey, J. L. Davis, G. Elgered, T. A. Herring, H. F. Hinteregger, C. A. Knight, J. I. Levine, G. Lundqvist, C. Ma, E. F. Nesman, R. B. Phillips, A. E. E. Rogers, B. O. Ronnang, J. W. Ryan, B. R. Schupler, D. B. Shaffer, I. I. Shapiro, N. R. Vandenberg, J. C. Webber, and A. R. Whitney, Precision Geodesy Using the Mark III Very Long Baseline Interferometry System, *IEEE Trans. Geodesy Remote Sens.*, **GE-23**, 438–449, 1985.

Clark, T. A., C. C. Counselman, P. G. Ford, L. B. Hanson, H. F. Hinteregger, W. J. Klepczynski, C. A. Knight, D. S. Robertson, A. E. E. Rogers, J. W. Ryan, I. I. Shapiro, and A. R. Whitney, Synchronization of Clocks by Very Long Baseline Interferometry, *IEEE Trans. Instrum. Meas.*, **IM-28**, 184–187, 1979.

Cohen, M. H., Introduction to Very-Long-Baseline Interferometry, *Proc. IEEE*, **61**, 1192–1197, 1973.

Cohen, M. H. and D. B. Shaffer, Positions of Radio Sources from Long-Baseline Interferometry, *Astron. J.*, **76**, 91–100, 1971.

Cutler, L. S. and C. L. Searle, Some Aspects of the Theory and Measurement of Frequency Fluctuations in Frequency Standards, *Proc. IEEE*, **54**, 136–154, 1966.

D'Addario, L. R., *Minimizing Storage Requirements for Quantized Noise*, VLBA Memo. No. 332, National Radio Astronomy Observatory, Charlottesville, VA, 1984.

D'Addario, L. R., Time Synchronization in Orbiting VLBI, *IEEE Trans. Instrum. Meas.*, **IM-40**, 584–590, 1991.

Davis, M. M., J. H. Taylor, J. M. Weisberg, and D. C. Backer, High Precision Timing of the Millisecond Pulsar PSR 1937 + 21, *Nature*, **315**, 547–550, 1985.

Dewey, R. J., The Effects of Correlated Noise in Phased-Array Observations of Radio Sources, *Astron. J.*, **108**, 337-345, 1994.

Drullinger, R. E., S. L. Rolston, and W. M. Itano, Primary Atomic Frequency Standards: New Developments, in *Review of Radio Science 1993–1996*, W. R. Stone, Ed., Oxford Univ. Press, Oxford, UK, 1996, pp. 11–41.

Dutta, P. and P. M. Horn, Low-Frequency Fluctuations in Solids: $1/f$ Noise, *Rev. Mod. Phys.*, **53**, 497–516, 1981.

Edson, W. A., Noise in Oscillators, *Proc. IRE*, **48**, 1454–1466, 1960.

Forman, P., Atomichron: The Atomic Clock from Concept to Commercial Product, *Proc. IEEE*, **73**, 1181–1204, 1985.

Frank, R. L., Current Developments in Loran-C, *Proc. IEEE*, **71**, 1127–1139, 1983.

Gardner, F. M., *Phaselock Techniques*, 2nd ed., Wiley, New York, 1979.

Hellwig, H., Microwave Time and Frequency Standards, *Radio Sci.*, **14**, 561–572, 1979.

Hellwig, H., R. F. C. Vessot, M. W. Levine, P. W. Zitzewitz, D. W. Allen, and D. J. Glaze, Measurement of the Unperturbed Hydrogen Hyperfine Transition Frequency, *IEEE Trans. Instrum. Meas.*, **IM-19**, 200–209, 1970.

Herring, T. A. *Precision and Accuracy of Intercontinental Distance Determinations Using Radio Interferometry*, Air Force Geophysics Laboratory, Hanscom Field, MA, AFGL-TR-84-0182, 1983.

Hinteregger, H. F., A. E. E. Rogers, R. J. Capallo, J. C. Webber, W. T. Petrachenko, and H. Allen, A High Data Rate Recorder for Astronomy, *IEEE Trans. Magn.*, **MAG-27**, 3455–3465, 1991.

Hirabayashi, H. and 52 coauthors, Overview and Initial Results of the Very Long Baseline Interferometry Space Observatory Program, *Science*, **281**, 1825–1829, 1998.

Kartashoff, P. and J. A. Barnes, Standard Time and Frequency Generation, *Proc. IEEE*, **60**, 493–501, 1972.

Kawaguchi, N., VLBI Recording System in Japan, in *Frontiers of VLBI*, H. Hirabayashi, M. Inoue, and H. Kobayashi, Eds.,Universal Academy Press, Tokyo, 1991, pp. 75–77.

Keshner, M. S., $1/f$ Noise, *Proc. IEEE*, **70**, 212–218, 1982.

Klemperer, W. K., Long Baseline Radio Interferometry with Independent Frequency Standards, *Proc. IEEE*, **60**, 602–609, 1972.

Kleppner, D., H. C. Berg, S. B. Crampton, N. F. Ramsey, R. F. C. Vessot, H. E. Peters, and J. Vanier, Hydrogen-Maser Principles and Techniques, *Phys. Rev. A*, **138**, 972–983, 1965.

Kleppner, D., H. M. Goldenberg, and N. F. Ramsey, Theory of the Hydrogen Maser, *Phys. Rev.*, **126**, 603–615, 1962.

Kogan, L. R. and L. S. Chesalin, Software for VLBI Experiments for CS-Type Computers, *Sov. Astron.*, **25**, 510–513, 1982, transl. from *Astron. Zh.*, **58**, 898–903, 1981.

Kokkeler, A. B. J., P. Fridman, and A. van Ardenne, Degradation due to Quantization Noise in Radio Astronomy Phased Arrays, *Experimental Astronomy*, **11**, 33–56, 2001.

Kulkarni, S. R., Self Noise in Inteferometers: Radio and Infrared, *Astron. J.*, **98**, 1112–1130, 1989.

Leick, A., *GPS Satellite Surveying*, 2nd ed., Wiley, New York, 1995.

Lesage, P. and C. Audoin, Characterization and Measurement of Time and Frequency Stability, *Radio Sci.*, **14**, 521–539, 1979.

Lewandowski, W., J. Azoubib, and W. J. Klepczynski, GPS: Primary Tool for Time Transfer, *Proc. IEEE*, Special Issue on Global Positioning System, **87**, No. 1, 163–172, 1999.

Lewandowski, W. and C. Thomas, GPS Time Transfer, *Proc. IEEE*, **79**, 991–1000, 1991.

Lewis, L. L., An Introduction to Frequency Standards, *Proc. IEEE*, **79**, 927–935, 1991.

Lindsey, W. C. and C. M. Chie, Frequency Multiplication Effects on Oscillator Instability, *IEEE Trans. Instrum. Meas.*, **IM-27**, 26–28. 1978.

Meinel, A. B., Multiple Mirror Telescopes of the Future, in *MMT and the Future of Ground-Based Astronomy*, T. C. Weeks, Ed., *SAO Special Report*, Vol. 385, Harvard-Smithsonian Astrophysical Obs., Cambridge, MA, 1979, pp. 9–22.

Moran, J. M., Spectral-Line Analysis of Very-Long-Baseline Interferometric Data, *Proc. IEEE*, **61**, 1236–1242, 1973.

Moran, J. M., Very Long Baseline Interferometric Observations and Data Reduction, in *Methods of Experimental Physics*, Vol. 12C, M. L.Meeks, Ed., Academic Press, New York, 1976, pp. 228–260.

Moran, J. M., Introduction to VLBI, in *Very Long Baseline Interferometry, Techniques and Applications*, M. Felli and R. E. Spencer, Eds., Kluwer, Dordrecht, 1989, pp. 27–45.

Moran, J. M. and V. Dhawan, An Introduction to Calibration Techniques for VLBI, in *Very Long Baseline Interferometry and the VLBA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron. Soc. Pacific Conf. Ser., **82**, 161–188, 1995.

Napier, P. J., D. S. Bagri, B. G. Clark, A. E. E. Rogers, J. D. Romney, A. R. Thompson, and R. C. Walker, The Very Long Baseline Array, *Proc. IEEE*, **82**, 658–672, 1994.

Papoulis, A., *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 1965.

Parkinson, B. W. and S. W. Gilbert, NAVSTAR: Global Positioning System—Ten Years Later, *Proc. IEEE*, **71**, 1177–1186, 1983.

Pierce, J. A., A. A. McKenzie, and R. H. Woodward, *Loran*, Radiation Laboratory Series, Vol. 4, McGraw-Hill, New York, 1948.

Press, W. H., Flicker Noises in Astronomy and Elsewhere, *Comments Astrophys.*, **7**, 103–119, 1978.

Reid, M. J., Spectral-Line VLBI, in *Very Long baseline Interferometry and the VLBA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron Soc. Pacific Conf. Ser., **82**, 209–225, 1995.

Reid, M. J., Spectral-Line VLBI, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron Soc. Pacific Conf. Ser., **180**, 481–497, 1999.

Reid, M. J., A. D. Haschick, B. F. Burke, J. M. Moran, K. J. Johnston, and G. W. Swenson, Jr., The Structure of Interstellar Hydroxyl Masers: VLBI Synthesis Observations of W3(OH), *Astrophys. J.*, **239**, 89–111, 1980.

Rogers, A. E. E., Very Long Baseline Interferometry with Large Effective Bandwidth for Phase Delay Measurements, *Radio Sci.*, **5**, 1239–1247, 1970.

Rogers, A. E. E., Theory of Two-Element Interferometers, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 139–157.

Rogers, A. E. E., The Sensitivity of a Very Long Baseline Interferometer, *Radio Interferometry Techniques for Geodesy*, NASA Conf. Pub. 2115, National Aeronautics and Space Administration, Washington, DC, 1980, pp. 275–281.

Rogers, A. E. E., Very Long Baseline Fringe Detection Thresholds for Single Baselines and Arrays, in *Frontiers of VLBI*, H. Hirabayashi, M. Inoue, and H. Kobayashi, Eds., Universal Academy Press, Tokyo, 1991, pp. 341–349.

Rogers, A. E. E., VLBA Data Flow: Formatter to Tape, in *Very Long Baseline Interferometry and the VLBA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron. Soc. Pacific Conf. Ser., **82**, 93–115, 1995.

Rogers, A. E. E., R. J. Cappallo, H. F. Hinteregger, J. I. Levine, E. F. Nesman, J. C. Webber, A. R. Whitney, T. A. Clark, C. Ma, J. Ryan, B. E. Corey, C. C. Counselman, T. A. Herring, I. I. Shapiro, C. A. Knight, D. B. Shaffer, N. R. Vandenberg, R. Lacasse, R. Mauzy, B. Rayhrer, B. R. Schupler, and J. C. Pigg, Very-Long-Baseline Interferometry: The Mark III System for Geodesy, Astrometry, and Aperture Synthesis, *Science*, **219**, 51–54, 1983.

Rogers, A. E. E., S. S. Doeleman, and J. M. Moran, Fringe Detection Methods for Very Long Baseline Arrays, *Astron. J.*, **109**, 1391–1401, 1995.

Rogers, A. E. E. and J. M. Moran, Coherence Limits for Very-Long-Baseline Interferometry, *IEEE Trans. Instrum. Meas.*, **IM-30**, 283–286, 1981.

Rutman, J., Characterization of Phase and Frequency Instability in Precision Frequency Sources: Fifteen Years of Progress, *Proc. IEEE*, **66**, 1048–1075, 1978.

Schwab, F. R. and W. D. Cotton, Global Fringe Search Techniques for VLBI, *Astron. J.*, **88**, 688–694, 1983.

Shapiro, I. I., Estimation of Astrometric and Geodetic Parameters, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 261–276.

Shimoda, K., T. C. Wang, and C. H. Townes, Further Aspects of the Theory of the Maser, *Phys. Rev.*, **102**, 1308–1321, 1956.

Siegman, A. E., *An Introduction to Lasers and Masers*, McGraw-Hill, New York, 1971, p. 404

Sovers, O. J., J. L. Fanselow, and C. S. Jacobs, Astrometry and Geodesy with Radio Interferometry: Experiments, Models, Results, *Rev. Mod. Phys.*, **70**, 1393–1454, 1998.

Thomas, J. B., *An Analysis of Radio Interferometry with the Block O System*, JPL Pub. 81–49, Jet Propulsion Laboratory, Pasadena, CA, 1981.

Thompson, A. R., The VLBA Receiving System: Antenna to Data Formatter, in *Very Long Baseline Interferometry and the VLBA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron. Soc. Pacific Conf. Ser., **82**, 73–92, 1995.

Thompson, A. R. and D. S. Bagri, A Pulse Calibration System for the VLBA, in *Radio Interferometry: Theory, Techniques and Applications*, T. J. Cornwell and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **19**, 55–59, 1991.

Vanier, J., M. Têtu, and L. G. Bernier, Transfer of Frequency Stability from an Atomic Frequency Reference to a Quartz-Crystal Oscillator, *IEEE Trans. Instrum. Meas.*, **IM-28**, 188–193, 1979.

Vessot, R. F. C., Frequency and Time Standards, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 198–227.

Vessot, R. F. C., Relativity Experiments with Clocks, *Radio Sci.*, **14**, 629–647, 1979.

Vessot, R. F. C., Applications of Highly Stable Oscillators to Scientific Measurements, *Proc. IEEE*, **79**, 1040–1053, 1991.

Vessot, R. F. C. and M. W. Levine, A Method for Eliminating the Wall Shift in the Atomic Hydrogen Maser, *Metrologia*, **6**, 116–117, 1970.

Vessot, R. F. C., M. W. Levine, E. M. Mattison, T. E. Hoffman, E. A. Imbier, M. Têtu, G. Nystrom, J. J. Kelt, H. F. Trucks, and J. L. Vaniman, Space-Borne Hydrogen Maser Design, *Proc. 8th Annual Precise Time and Interval Meeting*, U.S. Naval Research Laboratory, X-814-77-149, 1976, pp. 227–333.

Walker, R. C., Very Long Baseline Interferometry I: Principles and Practice, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron Soc. Pacific Conf. Ser., **6**, 355–378, 1989a.

Walker, R. C., Calibration Methods, in *Very Long Baseline Interferometry, Techniques and Applications*, M. Felli and R. E. Spencer, Eds., Kluwer, Dordrecht, 1989b, pp. 141–162.

Whitney, A. R., The Mark IV Data-Acquisition and Correlation System, in *Developments in Astrometry and Their Impact on Astrophysics and Geodynamics*, IAU Symp. 156, I. I. Mueller and B. Kolaczek, Eds., Kluwer, Dordrecht, 1993, 151–157.

Whitney, A. R., A. E. E. Rogers, H. F. Hinteregger, C. A. Knight, J. I. Levine, S. Lippincott, T. A. Clark, I. I. Shapiro, and D. S. Robertson, A Very Long Baseline Interferometer System for Geodetic Applications, *Radio Sci.*, **11**, 421–432, 1976.

Wietfeldt, R. D. and L. R. D'Addario, Compatibility Issues in VLBI, in *Radio Interferometry: Theory, Techniques, and Applications*, T. J. Cornwell and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **19**, 98–101, 1991.

Wietfeldt, R. D., D. Baer, W. H. Cannon, G. Feil, R. Jakovina, P. Leone, P. S. Newby, and H. Tan, The S2 Very Long Baseline Interferometry Tape Recorder, *IEEE Trans. Instrum. Meas.*, **IM-45**, 923–929, 1996.

Wietfeldt, R. D. and D. A. Frail, Burst Mode VLBI and Pulsar Applications, in *Radio Interferometry: Theory, Techniques, and Applications*, T. J. Cornwell and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **19**, 76–80, 1991.

Yen, J. L., K. I. Kellermann, B. Rayhrer, N. W. Broten, D. N. Fort, S.H. Knowles, W. B. Waltman, and G. W. Swenson, Jr., Real-Time, Very-Long-Baseline Interferometry Based on the Use of a Communcations Satellite, *Science*, **198**, 289–291, 1977.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

10 Calibration and Fourier Transformation of Visibility Data

This chapter is concerned with details of the calibration and Fourier transformation of visibility data, mainly as applied to earth-rotation synthesis. The use of the fast algorithm for the discrete Fourier transform (FFT) and methods for evaluation of the visibility at rectangular grid points are included. Special considerations for certain observing modes, including spectral line, are described. Some practical hints on the recognition and avoidance of errors in maps and the planning of observations are also discussed. The chapter is concerned principally with linear methods, and nonlinear image processing is discussed in Chapter 11.

10.1 CALIBRATION OF THE VISIBILITY

The purpose of calibration is to remove, insofar as possible, the effects of instrumental and atmospheric factors in the measurements. Such factors depend largely on the individual antennas or antenna pairs and their associated electronics, so correction must be applied to the visibility data before they are combined into an image. Editing the visibility data to delete any that show evidence of radio interference or equipment malfunction is usually performed before the calibration proper. This mainly entails examining samples of data for unexpected levels or phase variations. Data taken on calibration sources are particularly useful here, since the response to such a source is predictable and should vary only slowly and smoothly with time.

In the calibration procedure we first consider instrumental factors that are stable with time over periods of weeks or more. These include the following:

1. Antenna position coordinates that specify the baselines.
2. Antenna pointing corrections resulting from axis misalignments or other mechanical tolerances.
3. Zero-point settings of the instrumental delays, that is, the settings for which the delays from the antennas to the correlator inputs are equal.

These parameters vary only as a result of major changes such as the relocation of an antenna. They can be calibrated by observing unresolved sources with known positions. We assume here that they have been determined in advance of the mapping observations. We also assume that correction for the nonlinearity of signal quantization, which is discussed in Section 8.3 under *Quantization Correction*, is applied automatically if required.

Corrections for Calculable or Directly Monitored Effects

Calibration of the visibility measurements for effects that vary during an observation principally involves correction of the complex gains of the antenna pairs. Such factors can be divided into those for which the behavior can be predicted or directly measured and those for which it must be determined by observing a calibration source during the observation period. Examples of effects that can be corrected for by calculation of their effects include the following:

1. The constant component of atmospheric attenuation as a function of zenith angle (see Section 13.1 under *Absorption*).
2. Variation of antenna gain as a function of elevation caused by elastic deformation of the structure under gravity.
3. Shadowing of one antenna by another at close spacings and low elevation angles.

In the case of shadowing, where one antenna partially blocks the aperture of another, correction is generally difficult. The effect of the geometrical blockage is complicated by diffraction, the shape of the primary beam is modified, and the position of the phase center of the aperture is shifted, thus affecting the baseline. Data from shadowed antennas are frequently discarded.

Effects within the receiving system or external to it that can be continuously monitored during an observation include:

1. Variation of system noise temperature, which can result from changes in the ground radiation picked up in the sidelobes as the antenna tracks. This effect may also cause variation in the gain as a result of ALC (automatic level control) action that is used in some instruments to adjust the signal levels at the sampler or correlator (see Section 7.6). Monitoring can be performed by injection of a low-level, switched noise signal at the receiver input and detection of it later in the system.
2. Phase variations in the local oscillator system monitored by round-trip phase measurement (see Section 7.2).
3. The variable component of atmospheric delay monitored by using water vapor radiometers mounted at the antennas (see Section 13.1 under *Water Vapor Radiometry*).

Corrections for these effects are usually performed at an early stage of the calibration procedure.

Use of Calibration Sources

Further steps in the calibration involve parameters that may vary on timescales of minutes or hours and require the observation of one or more calibration sources. Note that the source that is the subject of the astronomical investigation will be referred to as the *target source* to distinguish it from the calibration source, or *calibrator*. From Eq. (3.9) we can write the expression for the interferometer response as follows:

$$[\mathcal{V}(u,v)]_{\text{uncal}} = G_{mn}(t)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{A_N(l,m)I(l,m)}{\sqrt{1-l^2-m^2}}e^{-j2\pi(ul+vm)}\,dl\,dm, \quad (10.1)$$

where $[\mathcal{V}(u,v)]_{\text{uncal}}$ is the uncalibrated visibility and $I(l,m)$ is the source intensity. The complex gain factor $G_{mn}(t)$ is a function of the antenna pair (m,n) and, as a result of unwanted effects, may vary with time. A_N is the antenna aperture normalized to unity for the direction of the main beam. It can be removed from the source image as a final step in the image processing. The factor $A_N(l,m)/\sqrt{1-l^2-m^2}$ in the intensity–visibility relationship is close to unity, and from here on we generally omit it, except in the case of wide-field mapping discussed in Section 11.8. To calibrate $G_{mn}(t)$, an unresolved calibrator can be observed, for which the measured response is

$$\mathcal{V}_c(u,v) = G_{mn}(t)S_c, \quad (10.2)$$

where the subscript c indicates the calibrator, and S_c is the flux density of the calibrator. In calibrating the gain it is best to consider the amplitude and phase separately, since the errors in these two quantities generally arise through different mechanisms. For example, at short centimeter wavelengths atmospheric fluctuations cause phase fluctuations but have little effect on the amplitudes. To calibrate the visibility of the target source, we can write

$$\mathcal{V}(u,v) = \frac{[\mathcal{V}(u,v)]_{\text{uncal}}}{G_{mn}(t)} = [\mathcal{V}(u,v)]_{\text{uncal}}\left[\frac{S_c}{\mathcal{V}_c}\right]. \quad (10.3)$$

To observe the calibration source it is placed at the phase center of its field. Then assuming that the calibrator is unresolved, the phase is a direct measure of the instrumental phase. Thus phase calibration for the target source simply requires subtracting the calibrator phase from the observed phase. The visibility amplitude can be calibrated by using the moduli of the visibility terms in Eq. (10.3). The response to the calibrator should be corrected for the calculable and/or directly monitored effects before the gain calibration is performed. Where there are separate receiving channels for two opposite polarizations at each antenna, the calibration must be performed separately for each one. For measurements of

source polarization further calibration procedures are necessary, as described in Section 4.8 under *Calibration of Instrumental Polarization.*

Calibration observations require periodic interruption of observations of the target source. At centimeter wavelengths the interval between calibration observations depends on the stability of the instrument, and typically falls within the range of 15 min to 1 h. At meter and millimeter wavelengths the ionosphere and the neutral atmosphere introduce gain and phase changes, and elimination of these may require observation of a calibrator at time intervals as short as one or two minutes.

As indicated by Eq. (7.38), $G_{mn} = g_m g_n^*$, so the measured gains for antenna pairs can be used to determine gain factors for the individual antennas. Using the antenna gain factors rather than the baseline gain factors reduces the calibration data to be stored, and helps in monitoring the performance of individual antennas. Also, with this technique, some of the spacings can be omitted from the calibration observation so long as each of the antennas is included. In practice, gain tables including both amplitude and phase are generated for the antennas as a function of time, and the values are interpolated to the times at which data from the target source were taken. The interpolation should be done separately for the amplitude and phase, not for the real and imaginary parts of the gain; otherwise the phase errors can degrade the amplitude, and vice versa. The desirable characteristics of a calibration source are the following.

Flux density. The calibrator should be strong, so that a good signal-to-noise ratio is obtained in a short time, to reduce the (u, v) coverage lost from the target source. The gaps in the (u, v) coverage are more serious for a linear array, in which complete sectors are lost, than for a two-dimensional array, in which the instantaneous coverage is more widely distributed in u and v.

Angular width. The calibrator should, if possible, be unresolved so that precise details of its visibility are not required.

Position. The position of the calibrator should be close to that of the target source. Effects in the atmosphere or antennas that cause the gain to vary with pointing angle are then more effectively removed, and time lost in driving the antennas between the target source and calibrator positions is kept small. At millimeter wavelengths, where the atmospheric phase path is the main factor being calibrated, the calibrator distance must be within the angular scale of the irregularities. This usually means a distance of no more than a few degrees on the sky.

It is not always possible to find a calibrator that satisfies all of the above requirements. In such cases it may be necessary to find a source that is largely unresolved and close to the target source, and then calibrate it against one of the more commonly used flux density references such as 3C48, 3C147, 3C286, and 3C295. The last of these is the most reliable with regard to non-variability. Thermal sources such as the compact planetary nebula NGC 7027 may be useful as amplitude cali-

brators for short baselines. At millimeter wavelengths it is particularly difficult to find a source that provides a strong signal for test purposes or calibration. Disks of planets become resolved at rather short baselines, but the limb of the moon or a planet can be useful; see Appendix 10.1.

For VLBI observations with milliarcsecond resolution, there are few suitable calibrators. Angular structure on this scale is sometimes variable over periods of months, and caution is necessary if a previously measured and partially resolved source is to be used as a calibrator. An alternative approach to amplitude calibration of VLBI data involves use of the system temperatures and collecting areas of the individual antennas, as follows. The cross-correlation data should first be normalized to unity for the case where the two input data streams are fully correlated. To obtain this normalization, the data are divided by the product of the rms values of the data streams at the two correlator inputs. (For two-level sampling this rms value is unity, and for other types of sampling the rms depends on the setting of the sampler thresholds with respect to the level of the analog signal.) Then, to convert the normalized correlation to visibility $\mathcal{V}$ with units of flux density (janskys), the amplitude is multiplied by the geometric mean of the system equivalent flux density values for the two antennas involved. The system equivalent flux density, $S_E = 2kT_S/A$, is defined in Eq. (1.6). Determination of the system temperature T_S and the collecting area A usually requires measurements in total-power mode with each antenna. If the value of T_S corresponds to a signal plane above the atmosphere, then the resulting visibility values will be corrected for atmospheric losses. For VLBI data in which the phase is not calibrated, the closure relationships in Section 10.3 allow maps to be formed if absolute position is not required.

10.2 DERIVATION OF INTENSITY FROM VISIBILITY

Mapping by Direct Fourier Transformation

The most straightforward method of obtaining an intensity distribution from measured visibility data is by *direct* Fourier transformation, that is, by performing the transformation without putting the visibility into any special form such as that for the fast algorithm described in Section 5.2. The measured visibility $\mathcal{V}_{\text{meas}}(u, v)$ can be written

$$\mathcal{V}_{\text{meas}}(u, v) = W(u, v)w(u, v)\mathcal{V}(u, v), \tag{10.4}$$

where $W(u, v)$ is the transfer function or spatial sensitivity function introduced in Section 5.3, and $w(u, v)$ represents any applied weighting. The Fourier transform of Eq. (10.4) is the measured intensity distribution, which is

$$I_{\text{meas}}(l, m) = I(l, m) ** b_0(l, m). \tag{10.5}$$

Here the double asterisk indicates two-dimensional convolution and b_0 is the synthesized beam, which is the Fourier transform of the weighted transfer function:

$$b_0(l, m) \rightleftharpoons W(u, v)w(u, v), \tag{10.6}$$

where $\rightleftharpoons$ indicates the Fourier transform relationship. Effects such as those of non-coplanar baselines, signal bandwidth, and visibility averaging are unimportant in many observations and are not included here.

The visibility is measured at an ensemble of n_d pairs of points symmetric about the (u, v) origin, and the direct Fourier transform of these data is represented by

$$\sum_{i=1}^{n_d} w_i \left[\mathcal{V}_{\text{meas}}(u_i, v_i)e^{j2\pi(u_i l+v_i m)} + \mathcal{V}_{\text{meas}}(-u_i, -v_i)e^{-j2\pi(u_i l+v_i m)}\right]. \tag{10.7}$$

The weighting factor w_i is introduced to control the form of the synthesized beam. Since the visibility at $(-u_i, -v_i)$ is the complex conjugate of the visibility at (u_i, v_i), the derived intensity is real. (Here we are considering the case where the antennas are identically polarized; for other cases, see the general analysis in Section 4.8.) In the Fourier transformation of the visibility, the intensity is usually computed at points in a rectangular grid with uniform increments in l and m, since this is a very convenient form for subsequent processing.

Weighting of the Visibility Data

To obtain the best signal-to-noise ratio in the summation of measurements that contain Gaussian noise, the individual data values should be weighted inversely as their variances. The same is true for the combination of sinusoidal components of a source map, the amplitudes of which are proportional to the corresponding visibility points. Thus, for best signal-to-noise ratio, the weights w_i in (10.7) should be inversely proportional to the variances. If the data are obtained with a uniform array of antennas and receivers, and the averaging time is the same for all data, then the variances should all be the same and maximum signal-to-noise ratio is obtained by including all measurements with the same weight. This is known as *natural weighting*. For most arrays natural weighting results in a poor beam shape with wide skirts because the shorter spacings are overemphasized. Thus the usual approach is to include in the weighting a factor that is inversely related to the area density of the data in the (u, v) plane. The area density $\rho_\sigma(u, v)$ can be defined such that the number of points in the range $u \pm \frac{1}{2}du$, $v \pm \frac{1}{2}dv$ is $\rho_\sigma(u, v)\,du\,dv$ (Thompson and Bracewell 1974). Although ρ_σ at any given point depends on the size of the increments du and dv, it is usually possible to specify the variation of relative density and correct for it satisfactorily. As a simple example, in the observation of a high-declination source with an east–west array in which the antenna spacings are nonredundant integral multiples of a unit value, the visibility points lie on concentric circles as in Fig. 10.1. Then, if the visibility is measured at uniform increments in hour angle, the area density at any ring is inversely proportional to the radius of the ring. With $w(u, v)$ proportional to $1/\rho_\sigma(u, v)$, the effective density of the data is uniform within a circle of radius $u_{\max}$ determined by the maximum spacing. The beam then closely approximates

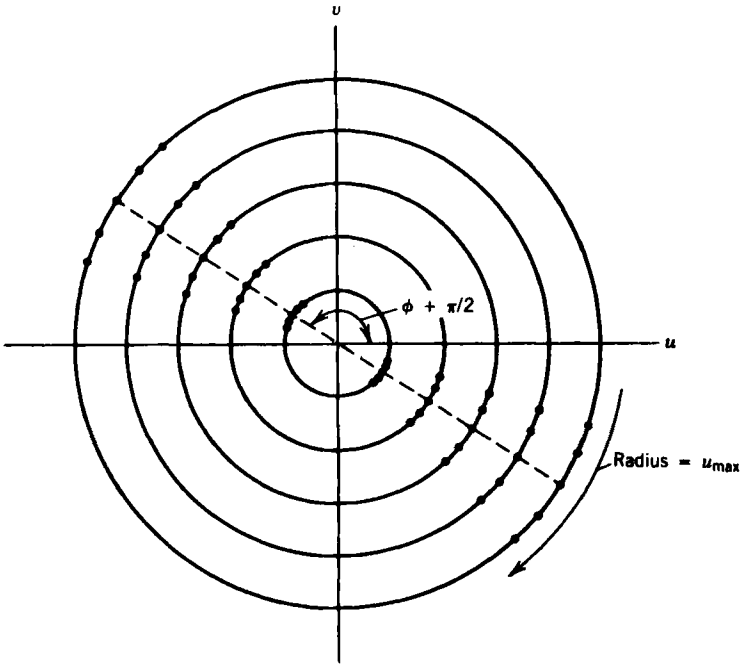


Figure 10.1 Transfer function (spacing loci) in the (u, v) plane for observations of a high-declination source using an east–west array with uniform increments in antenna spacing. The points indicate visibility measurements, and their (u, v) positions reflected through the origin, for uniform intervals of time. The angle ϕ indicates data for a specific hour angle. If the visibility values are weighted in proportion to the radii of the loci, the density of the visibility data is effectively uniform out to a radius u_{max}.

the Fourier transform of a circular disk function which, normalized to unity at the maximum, is given by

$$\frac{J_1(2\pi l u_{max})}{\pi l u_{max}}, \tag{10.8}$$

where J_1 is the Bessel function of the first kind and first order. The full width of the beam at half maximum is $0.705u_{max}^{-1}$, and the first sidelobe response is 13.2% of the main beam.* Similarly, if the effective density of measurements is uniform within a rectangular area of dimensions $2u_{max} \times 2v_{max}$, the synthesized beam is

*This synthesized response should not be confused with the power pattern of a uniformly illuminated antenna with circular aperture of radius r, which is proportional to $[J_1(2\pi rl/\lambda)/(\pi rl/\lambda)]^2$ and has a full width at half maximum of $0.514\lambda/r$, first null at $0.610\lambda/r$, and first sidelobe of 1.7%. The antenna pattern is proportional to the Fourier transform of the autocorrelation function of a uniform circular aperture.

closely approximated by

$$\frac{\sin(2\pi u_{\max} l)}{2\pi u_{\max} l} \times \frac{\sin(2\pi v_{\max} m)}{2\pi v_{\max} m}. \tag{10.9}$$

This beam is not circularly symmetrical, and the first sidelobe has a maximum value of 22% in the east–west and north–south directions through the beam center.

With uniform weighting the strong, near-in sidelobes (close to the main beam) in Fig. 10.2 obscure low-level detail and thereby reduce the range of intensity levels that can be reliably measured. The near-in sidelobes of the functions in expressions (10.8) and (10.9) can be reduced at the expense of some increase in the width of the synthesized beam by introducing a Gaussian or similar taper into the weighting function. The effect of such tapering of the visibility is shown in Fig. 10.2. The taper can be specified in terms of the amplitude of the tapering

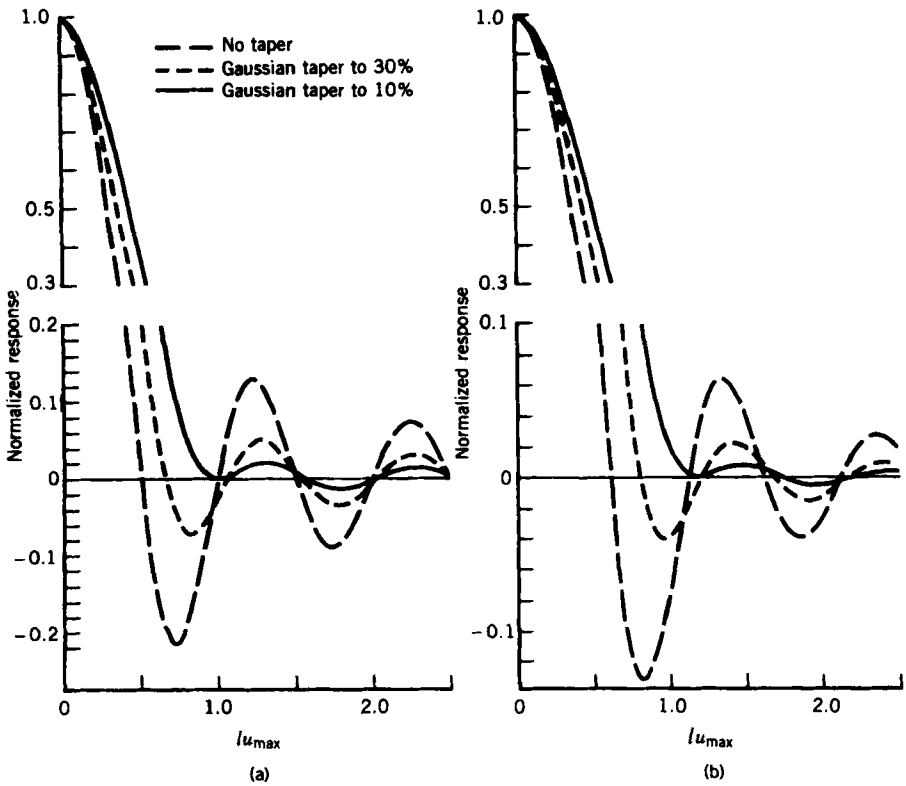


Figure 10.2 Examples of synthesized beam profiles. Curves for no taper correspond to a visibility distribution that is uniform within (a) a rectangular area of width $2u_{\max}$, and (b) a circular area of diameter $2u_{\max}$. For no taper the responses correspond to expression (10.9) for (a) and (10.8) for (b). The effects of Gaussian tapers that reduce the visibility at the edge of the distribution to 30% and to 10% are also shown. Note the difference in the ordinate scales.

function at a distance $u_{\max}$ from the (u, v) origin; a taper to ~−13 dB of the central value is commonly used. With such a taper the weighting $w(u, v)$ is the product of two functions: $w_u(u, v)$, the weighting required to obtain uniform effective density, and $w_t(u, v)$, the tapering function. Thus, the synthesized beam is the Fourier transform of $W(u, v)w_u(u, v)w_t(u, v)$:

$$b_0(l, m) = \overline{W}(l, m) ** \overline{w}_u(l, m) ** \overline{w}_t(l, m), \tag{10.10}$$

where the bar denotes a Fourier transform. The Fourier transform of $W(u, v)w_u(u, v)$ is simply the beam obtained with uniform effective density, for example, as in (10.8) or (10.9). If $w_t(u, v)$ is a two-dimensional Gaussian function, its Fourier transform is also a Gaussian. Thus the sidelobe reduction results from convolution with a Gaussian in the (l, m) domain. The variances of functions are additive under convolution [see, e.g., Bracewell (2000)], so the beam obtained by convolution with $\overline{w}_t$ is broader than that with no tapering, as is evident in Fig. 10.2.

An interesting property of the uniform weighting is that it minimizes the mean-squared deviation of the resulting intensity from the true intensity, within the constraint that unmeasured visibility values remain zero. This can be understood as follows. Since the true intensity distribution $I(l, m)$ and the true visibility function $\mathcal{V}(u, v)$ are a Fourier pair, and the weighted measured visibility and the derived intensity $I_0(l, m)$ are a Fourier pair, it follows that the differences between these quantities in the two domains are also a Fourier pair, to which we can apply Parseval's theorem. Recall that $W(u, v)$ is the transfer function, $w_u(u, v)$ is the weighting required to obtain effective uniform density of data in the (u, v) plane, and $w_t(u, v)$ is an applied taper. Thus, we can write

$$\begin{aligned} &\int\int_{\text{meas}} |\mathcal{V}(u, v) - \mathcal{V}(u, v)W(u, v)w_u(u, v)w_t(u, v)|^2 \, du\, dv \\ &\quad + \int\int_{\text{unmeas}} |\mathcal{V}(u, v)|^2 \, du\, dv \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |I(l, m) - I_0(l, m)|^2 \, dl\, dm. \end{aligned} \tag{10.11}$$

The first and second lines of Eq. (10.11) represent the measured and unmeasured areas of the (u, v) plane, respectively. In the measured area, $W(u, v)w_u(u, v) = 1$. For the case of uniform weighting, $w_t = 1$, so the integral on the first line is zero. This condition minimizes the squared difference between the true and observed intensity distributions on the third line. If $I(l, m)$ is an unresolved point source, then $I_0(l, m)$ is equal to the synthesized beam. The uniform weighting minimizes the squared difference, over 4π sr, between the synthesized beam and the response to a point source as it would be observed with unlimited (u, v) coverage. In this sense it is sometimes said that uniform weighting minimizes the sidelobes of the synthesized beam. However, as shown in Fig. 10.2, a Gaussian taper reduces the sidelobes outside of the main beam at the expense of widening

the beam. Seemingly contradictory statements about "minimizing the sidelobes" can occur if it is not clear exactly what is meant. Maps derived from visibility data that are uniformly weighted within the measured area of the (u, v) plane have been referred to as the *principal solution* or *principal response* (Bracewell and Roberts 1954).

Briggs (1995) has developed a procedure known as *robust weighting* in response to a problem that is encountered in the use of uniform weighting,. Visibility measurements that are isolated in (u, v) space can occur, for example, as a result of loss of neighboring measurements by malfunction or sporadic interference. With uniform weighting such points are assigned high weights. If they occur at (u, v) points where the visibility is low, such data may consist mainly of noise. The Fourier transform of an erroneous (u, v) point and its conjugate introduce a cosine ripple component into the intensity background that limits the dynamic range of the image. Robust weighting introduces an algorithm that takes account of the signal-to-noise ratio of individual points in the assignment of weights, and reduces the weighting of noisy points. More generally, robust weighting can be viewed as optimizing the combined effect of noise and extended sidelobes, by varying the weighting of individual points between the extremes of natural weighting, which optimizes sensitivity, and uniform weighting, which improves the poor beam shape of natural weighting (Briggs, Sramek, and Schwab 1999). A similar approach is useful in the case of an array with different sizes of antennas, or quality of receivers, since, to obtain the maximum signal-to-noise ratio, it is then necessary to weight the data in inverse proportion to their variance. The related process of reducing the sidelobe reponse in optical imaging is called apodization, for which there is an extensive literature; see, for example, Jacquinot and Roizen-Dossier (1964), Slepian (1965).

Mapping by Discrete Fourier Transformation

The speed of the fast algorithm for the discrete Fourier transform (FFT), briefly discussed in Section 5.2, is a major advantage in computing large maps. However, the use of the FFT introduces two complications in addition to those discussed for the direct transform: (1) the necessity to evaluate the visibility at points on a rectangular grid and (2) the resulting possibility of aliasing of parts of the image from outside the synthesized field. The evaluation at the grid points is often referred to as *gridding*. The output of such a process can be represented by the following expression:

$$\frac{w(u, v)}{\Delta u \Delta v} {}^2\mathrm{III}\left(\frac{u}{\Delta u}, \frac{v}{\Delta v}\right) \{C(u, v) ** [W(u, v)\mathcal{V}(u, v)]\}. \tag{10.12}$$

Here the visibility $\mathcal{V}(u, v)$, measured at the points denoted by the transfer function $W(u, v)$, is convolved with a function $C(u, v)$ to produce a continuous visibility distribution. This is then resampled at points in a rectangular grid with incremental spacings Δu and Δv. This process is often referred to as *convolutional gridding*. The resampling is here represented by the two-dimensional shah function ${}^2\mathrm{III}$ (Bracewell 1956a), defined by

$$^2\text{III}\left(\frac{u}{\Delta u}, \frac{v}{\Delta v}\right) = \Delta u \Delta v \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} {}^2\delta(u - i\Delta u,\ v - k\Delta v), \tag{10.13}$$

where $^2\delta$ is the two-dimensional delta function. The weighting to optimize the beam is applied to the resampled data. Although this process is described mathematically in terms of convolution and resampling, in practice the convolution is evaluated only at the grid points. The Fourier transform of (10.12) represents the measured intensity:

$$I_{\text{meas}}(l, m) = {}^2\text{III}(l\Delta u, m\Delta v) ** \overline{w}(l, m) ** \left\{\overline{C}(l, m)\left[\overline{W}(l, m) ** I(l, m)\right]\right\}. \tag{10.14}$$

As a result of the Fourier transformation, the intensity function $I(l, m)$ is convolved with the Fourier transform of the transfer function, multiplied by $\overline{C}(l, m)$ which is the Fourier transform of the convolving function, and then convolved with the Fourier transforms of the weighting and resampling functions. This last convolution causes the whole map to be replicated at intervals Δu^{-1} in l and Δv^{-1} in m. These intervals are equal to the dimensions of the map in the (l, m) plane; that is, $\Delta u^{-1} = M\Delta l$ and $\Delta v^{-1} = N\Delta m$, for an $M \times N$ point array. The function $\overline{C}(l, m)$ takes the form of a taper applied to the map, and if this function does not vary greatly on the scale of the width of $\overline{w}(l, m)$, which is usually the case for large maps, then $\overline{w}(l, m)$ in Eq. (10.14) can be convolved directly with $\overline{W}(l, m) ** I(l, m)$, and Eq. (10.14) becomes

$$I_{\text{meas}}(l, m) \simeq {}^2\text{III}(l\,\Delta u, m\,\Delta v) ** \left\{\overline{C}(l, m)\left[I(l, m) ** b_0(l, m)\right]\right\}, \tag{10.15}$$

where the synthesized beam $b_0(l, m)$ enters through the relationship in Eq. (10.6). Comparison with Eq. (10.5) shows that the effect of the gridding and resampling is to multiply the map by $\overline{C}(l, m)$ and replicate it. This replication introduces the aliasing.

Returning to the estimation of the visibility at the grid points, we might perhaps expect the best technique to be some form of exact interpolation so that the resulting values are equal to those that would be obtained by measurement at the grid points. A method of this type has been described by Thompson and Bracewell (1974). However, the problem of aliasing remains, and the most effective way to deal with this is to convolve the data in the (u, v) plane with the Fourier transform of a function that, in the (l, m) plane, varies very little over the map and then falls off rapidly at the map edges. We therefore look for a convolving function $C(u, v)$ for which the Fourier transform $\overline{C}(l, m)$ has these properties. An ideal function with infinitely sharp cutoff at the field edges would completely eliminate the aliasing since there would be no overlap of the replicated maps. Unfortunately, this ideal is not practical because the required convolving function is not bounded in the (u, v) plane. Nevertheless, a very worthwhile degree of suppression of the aliasing is possible with a careful choice of functions. A common and convenient practice is to combine both the gridding and the convolution to minimize aliasing into a single operation. Note, however, that at the (u, v) points

at which the measurements are made the function $C(u, v) * * [W(u, v)\mathcal{V}(u, v)]$, in general, is not equal to the measured visibility $\mathcal{V}(u, v)$. Thus the gridding process cannot be described as interpolation. Also, because of the convolution, the sampled points represent averages of the visibility local to the grid points, rather than samples of the visibility function. Finally, note also that although convolution is effective in suppressing artifacts that result from gridding of the data, it does not reduce sidelobe or ringlobe responses to sources located outside the area of the map.

Convolving Functions and Aliasing

From the foregoing discussion we can conclude that the point of principal concern in the use of the FFT is the choice of convolving function. A detailed discussion of convolving functions is given by Schwab (1984). It is convenient to consider those that are separable into one-dimensional functions of the same form for u and v, that is,

$$C(u, v) = C_1(u)C_1(v). \tag{10.16}$$

We therefore discuss some examples of the function C_1.

Rectangular Function. This function is the one used in cell averaging discussed in Section 5.2. It can be written

$$C_1(u) = (\Delta u)^{-1}\Pi\left(\frac{u}{\Delta u}\right), \tag{10.17}$$

where Π is the unit rectangle function defined by

$$\Pi(x) = \begin{cases} 1, & |x| \leq \frac{1}{2} \\ 0, & |x| > \frac{1}{2}. \end{cases} \tag{10.18}$$

The Fourier transform of $C_1(u)$ is

$$\overline{C}_1(l) = \frac{\sin(\pi \Delta u l)}{\pi \Delta u l}. \tag{10.19}$$

At the edge of the synthesized field, $l = (2\Delta u)^{-1}$ and $\overline{C}_1(1/2\Delta u) = 2/\pi$. The map is tapered by a sinc-function profile in the l and m directions and a sinc-squared profile along the diagonals. Equation (10.19) is plotted in Fig. 10.3, and the value at the first maximum outside the edge of the map is 0.22 of the value at the map center. The effect of aliasing is shown more directly in Fig. 10.4a, which is a plot of $\overline{C}_1(l)/\overline{C}_1[f(l)]$, where $f(l)$ is the value of l within the map [i.e., $|f(l)| < (2\Delta u)^{-1}$] at which the alias of a feature of l would appear. This quantity

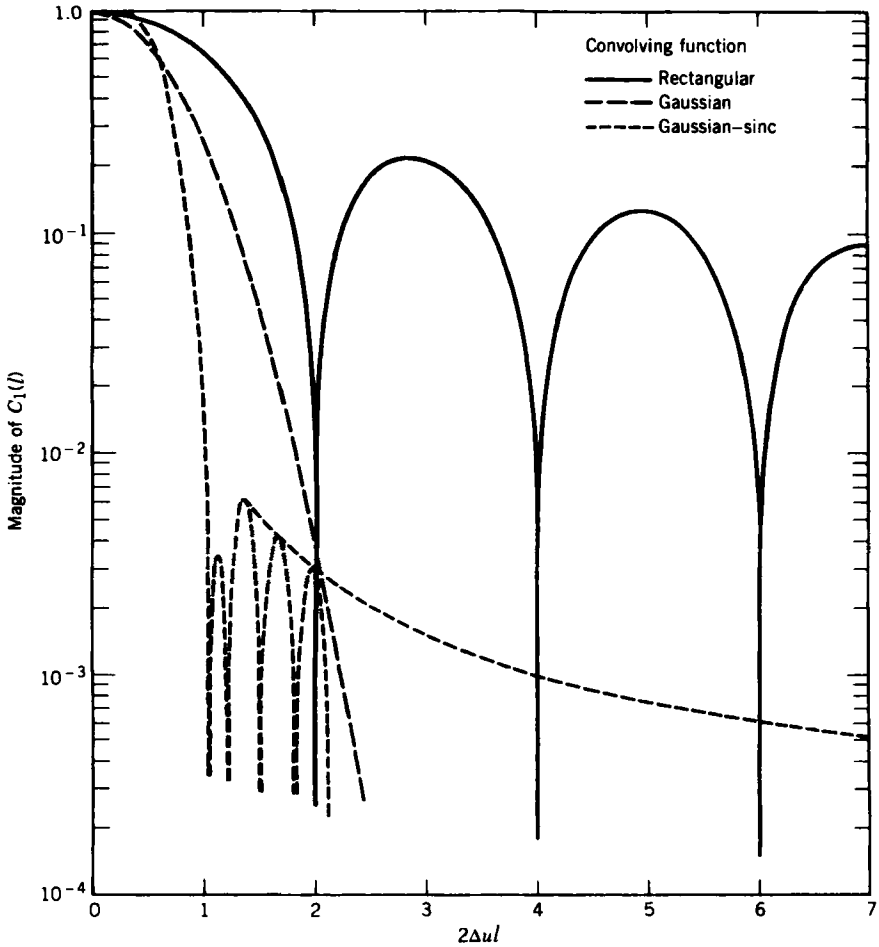


Figure 10.3 Three examples of the tapering function $\overline{C}_1(l)$, which is the Fourier transform of the convolving function $C_1(u)$. For the Gaussian convolving function, $\alpha = 0.75$. For the Gaussian-sinc convolving function, $\alpha_1 = 1.55$, $\alpha_2 = 2.52$, and beyond the fourth subsidiary maximum only the envelope of the maxima is shown. On the abscissa scale the center of the map is at zero and the edge at 1.0. The data for the Gaussian-sinc function were computed by F. R. Schwab.

gives the relative response to an aliased feature in a map that has been corrected for the taper imposed by $\overline{C}_1(l)$. It is clear that simple averaging of points within a rectangular cell performs poorly in suppressing aliasing.

Gaussian Function. Here we have

$$C_1(u) = \frac{1}{\alpha \Delta u \sqrt{\pi}} e^{-(u/\alpha \Delta u)^2} \tag{10.20}$$

and

$$\overline{C}_1(l) = e^{-(\pi\alpha\Delta u l)^2}. \tag{10.21}$$

The value of the constant α can be chosen to vary the widths of the functions as desired. If α is too small $C_1(u)$ will be too narrow, and only visibility measurements that are close to grid points will be used effectively in the mapping. If α is too large the function $\overline{C}_1(u)$ will taper the resulting map too severely. The Gaussian convolving function was used in the early years of the Westerbork array with $\alpha = 2\sqrt{\ln 4}/\pi = 0.750$ (Brouw 1971). The value of the factor $e^{-(u/\alpha\Delta u)^2}$ in $C_1(u)$ is then equal to 0.41 for a point on a diagonal in the (u, v) plane midway between two grid points. Thus, all measured points enter into the map with significant weights, and at the edge of the map the tapering factor $\overline{C}_1 = \frac{1}{4}$. A curve for the Gaussian function is shown in Fig. 10.3.

Gaussian-Sinc Function. The ideal form for the map tapering function $\overline{C}_1(l)$ would be a rectangle, which corresponds to convolution with a sinc function as in Eq. (10.19). However, the envelope of a sinc function falls to zero slowly as its argument increases, and the computation required for the convolution becomes large. Truncation of the sinc function is undesirable because in the l domain the desired rectangular function is convolved with the Fourier transform of the truncation function, and this destroys the sharp cutoff at the map edges. A better procedure is to multiply the sinc function with a Gaussian, which gives

$$C_1(u) = \frac{\sin(\pi u/\alpha_1\Delta u)}{\pi u} e^{-(u/\alpha_2\Delta u)^2} \tag{10.22}$$

and

$$\overline{C}_1(l) = \Pi(\alpha_1\Delta u l) * \left[\sqrt{\pi}\alpha_2\Delta u e^{-(\pi\alpha_2\Delta u l)^2}\right]. \tag{10.23}$$

Good performance is obtained with $\alpha_1 = 1.55$ and $\alpha_2 = 2.52$, with the convolution extending over an area about $6\Delta u$ in width. Corresponding curves for $\overline{C}_1(l)$ and the resulting aliasing are given in Figs. 10.3 and 10.4b. This convolving function is much better than either of the two previous examples.

Spheroidal Functions. Various other functions can be found that have the features desirable for convolution. As a measure of the effectiveness of the suppression of aliasing, Brouw (1975) has suggested the following quantity:

$$\frac{\iint_{\text{map}} \left[\overline{C}(l, m)\right]^2 dl\, dm}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\overline{C}(l, m)\right]^2 dl\, dm}, \tag{10.24}$$

which shows the fraction of the integrated squared amplitude of the tapering function that falls within the map. Maximization of (10.24) provides a criterion for choosing a convolving function. This approach led to consideration of the

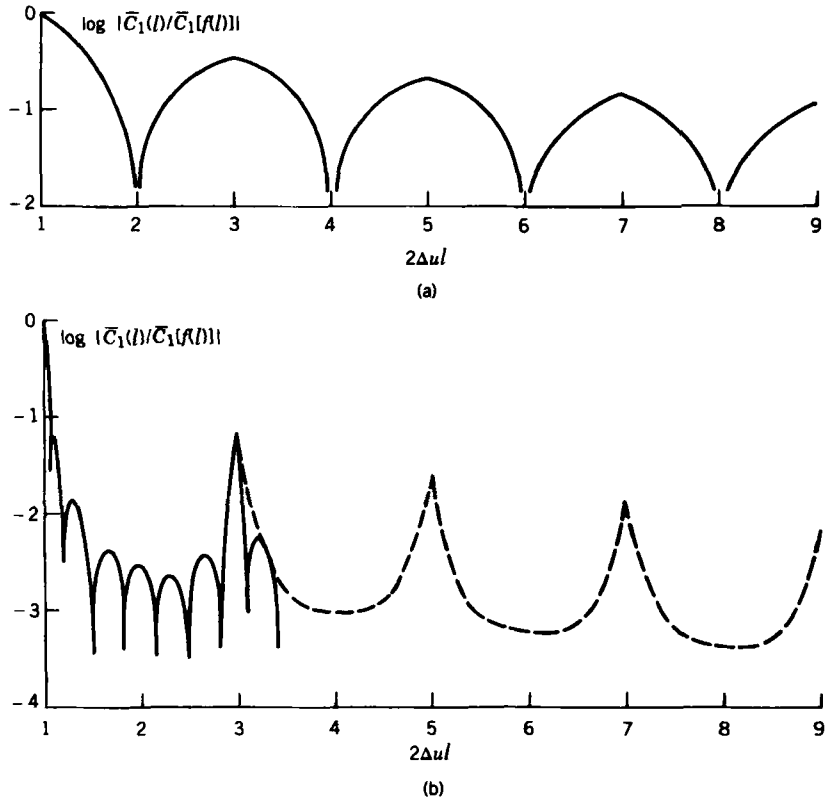


Figure 10.4 Logarithmic plot of the factor by which the amplitudes of structures outside the map are multiplied when aliased into the map. On the abscissa scale 1.0 is the edge of the map and 2, 4, 6, ... are the centers of the adjacent replications. (a) Aliasing factor for a rectangular convolving function of width equal to Δu (cell averaging). (b) Aliasing factor for a Gaussian-sinc convolving function with the optimized parameters given in the text. The broken line indicates the envelope of the maxima. Data computed by F. R. Schwab.

prolate spheroidal wave functions [see, e.g., Slepian and Pollak (1961)] and the spheroidal functions (Rhodes 1970). Schwab (1984) found that among functions investigated, the latter provide the best approach to an optimum convolving function. The spheroidal functions are solutions to certain differential equations and are not expressible in simple analytic form. In applying such functions for convolution of visibility data, they are computed in advance to provide a lookup table. Comparison of some functions of this type with the Gaussian-sinc function shows that the aliasing factor $\overline{C}_1(l)/\overline{C}_1[f(l)]$ falls off about as rapidly from the center to the edge of the map, but as l increases beyond the edge of the map, it reaches values an order of magnitude or more lower than those for the Gaussian-sinc function (Briggs, Sramek, and Schwab 1999). Computational capacity complicates the choice of the optimal function, since it limits the area of the (u, v) plane over which the convolution can be performed. Commonly this area is six to eight

grid cells wide and centered on the point to be interpolated. Roundoff errors in the Fourier transform are amplified in the removal of the tapering function and may limit the allowable taper at the map edges.

Aliasing and the Signal-to-Noise Ratio

Features aliased into a map from outside the boundary include not only the images of features on the sky but also the random variations resulting from the system noise. If we consider a direct Fourier transform of the noise component of the measured visibility, it is clear from expression (10.7) that for any point (l, m) the visibility data are weighted by complex exponential factors, all of which have the same modulus. Since the noise is independent at each data point in the (u, v) plane, the variance of the noise in the (l, m) plane is statistically constant in all parts of the map. If the FFT is used, however, the rms noise level across the map is multiplied by the function $\overline{C}(l, m)$, and details beyond the map edge are aliased into the map. Note that the noise contributions combine additively in the variance. Thus, in one dimension the noise variance as a function of l is proportional to

$$\mathrm{III}(l\Delta u) * |C_1(l)|^2. \tag{10.25}$$

The replication resulting from the FFT can also be written in terms of a summation, and the variance of the noise at a point l within the map is then proportional to

$$\sum_{i=-\infty}^{\infty} |\overline{C}_1(l + i\Delta u^{-1})|^2. \tag{10.26}$$

Usually $\overline{C}_1(l)$ decreases sufficiently with l that only the noise from the adjacent replication of the map makes a serious contribution through aliasing. This contribution is greatest near the edge of the map, as shown in Fig. 10.5 (Crane and Napier 1989).

If the convolving function is the Gaussian-sinc type, we see from Fig. 10.4b that, except for values of $2\Delta ul$ between 1.0 and 1.1, aliased features are reduced in amplitude by a factor $< 10^{-1}$, and in the square of the amplitude by $< 10^{-2}$. Thus, there is no significant increase in the noise level as a result of aliasing, except in a narrow zone at the edge of the map.

At the other extreme, the aliasing is most serious in the case of cell averaging, for which $C_1(u)$ is the sinc function given by Eq. (10.19). Expression (10.26) then becomes

$$\sum_{i=-\infty}^{\infty} \frac{\sin^2[\pi(\Delta ul + i)]}{[\pi(\Delta ul + i)]^2} = 1, \tag{10.27}$$

which indicates that the aliasing exactly cancels the taper, and the variance of the noise is constant with l, that is, before any correction for tapering of the astronomical features in the image is applied. (This result could also be deduced from the fact that in cell averaging each visibility measurement contributes to one grid

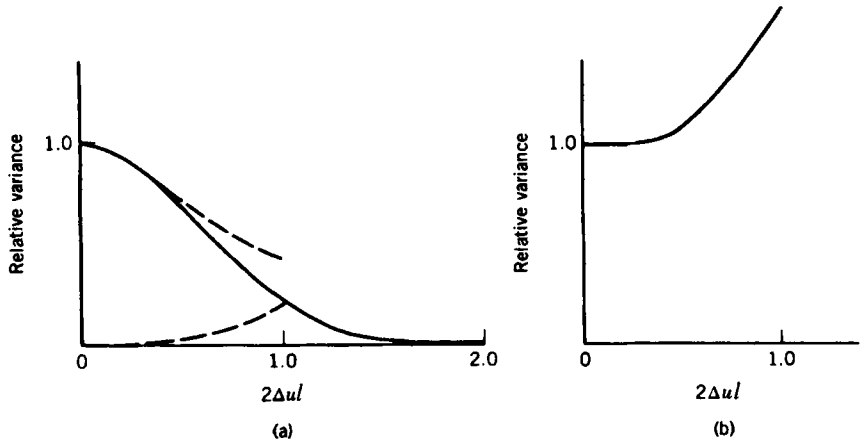


Figure 10.5 Effect of aliasing on the variance of the noise across a map. The abscissa in each case is l in units of half the map width; the map center is at 0, the edge at 1.0, and the center of the adjacent replication at 2.0. (a) Solid curve shows the taper for a Gaussian convolving function C_1, and broken curves show the effect of aliasing. (b) Variance of the noise including aliased component after correction for taper C_1. After Napier and Crane (1982).

point only, and the noise components of the visibility at the grid points are therefore independent.) However, the intensity distribution of the sky within the field being mapped is tapered by the function $\overline{C}_1(l)$, and correction for this taper then causes the noise to increase toward the map edges. For the sinc-function taper the noise is increased by a factor of $\pi/2$ at the edge of the map on the l and m axes and by $(\pi/2)^2$ at the corners. At the center of the map the aliased contribution originates at points for which $2\Delta ul$ is an even integer in the plots in Fig. 10.4, and in both cases shown the aliasing factor $\overline{C}_1(l)/\overline{C}_1[f(l)]$ drops to a very low value. With any of the convolving functions that we have considered, there is no significant increase in the noise at the center of the map, and the signal-to-noise ratio for a source at that point is determined by the factors discussed in Section 6.2.

10.3 CLOSURE RELATIONSHIPS

Closure effects are relationships between visibility values for baselines that form a closed figure, for example, a triangle or quadrilateral with the antennas at the vertices. As shown by Eqs. (7.37) and (7.38), the correlator output for antenna pair (m, n) can be written as

$$r_{mn} = G_{mn}\mathcal{V}_{mn} = g_m g_n^* \mathcal{V}_{mn}, \tag{10.28}$$

where G_{mn} is the complex gain for the antenna pair, and g_m and g_n are gain factors for the individual antennas. We ignore any gain terms that do not factor into the

terms for individual antennas (see Section 7.3 under *Tolerances on Variation of the Frequency Response: Gain Errors*). Considering first the phase relationships, we represent the arguments of the exponential terms of r_{mn}, g_m, g_n, and $\mathcal{V}_{mn}$ by ϕ_{rmn}, ϕ_{gm}, ϕ_{gn}, and ϕ_{vmn} respectively. Thus we can write

$$\phi_{rmn} = \phi_{gm} - \phi_{gn} + \phi_{vmn}. \tag{10.29}$$

Now for three antennas m, n, and p the phase closure relationship is

$$\phi_{rmn} + \phi_{rnp} + \phi_{rpm} = \phi_{vmn} + \phi_{vnp} + \phi_{vpm}. \tag{10.30}$$

The antenna gain terms, g_m and so on, contain the effects of the atmospheric paths to the antennas as well as instrumental effects, and since these terms do not appear in Eq. (10.30), it is evident that the combination of the three correlator output phases constitutes an observable quantity that depends only on the phase of the visibility. This property of the phase closure relationships was first recognized and used by Jennison (1958) in the experiments mentioned in Section 1.3 under *Early Measurements of Angular Width*.

If we have n_a antennas and we measure the correlation of all pairs, the number of independent phase closure relationships is equal to the number of correlator output phases less the number of unknown instrumental phases, one of which can be arbitrarily chosen. If there are no redundant spacings, then each closure relationship provides different information on the source structure. The number of phase closure relationships is

$$\tfrac{1}{2}n_a(n_a - 1) - (n_a - 1) = \tfrac{1}{2}(n_a - 1)(n_a - 2). \tag{10.31}$$

This number can also be obtained by taking one antenna and considering the number of different groups of three that can be formed that include it. Each of these groups must contain one baseline that is not in any other group so that the relationships are independent.

An amplitude closure relationship involves four antenna pairs, for which four antennas m, n, p, and q are required:

$$\frac{|r_{mn}||r_{pq}|}{|r_{mp}||r_{nq}|} = \frac{|\mathcal{V}_{mn}||\mathcal{V}_{pq}|}{|\mathcal{V}_{mp}||\mathcal{V}_{nq}|}. \tag{10.32}$$

The proof of Eq. (10.32) is obtained by substituting terms of the form $g_m g_n^* \mathcal{V}_{mn}$ into the left-hand side of Eq. (10.32), using Eq. (10.28). The moduli of the g terms then cancel out. There are two independent closure relationships for the four antennas; a second one may be obtained by replacing the subscripts in the numerator (or denominator) of Eq. (10.32) by mq and np. The number of independent amplitude closure relationships for n_a antennas with no redundant baselines is equal to the number of measured amplitudes, $\frac{1}{2}n_a(n_a - 1)$, less the number of unknown

antenna gain factors n_a, that is,

$$\tfrac{1}{2}n_a(n_a - 1) - n_a = \tfrac{1}{2}n_a(n_a - 3). \tag{10.33}$$

For early usage of the principle of taking ratios of observed visibility amplitudes to eliminate instrumental gains, see Smith (1952) and Twiss, Carter, and Little (1960).

Note that a fundamental requirement for the validity of the closure relationships is that at any instant it should be possible to represent the effect of any signal path from the source to the correlator by a single complex gain factor. Thus the effects of the atmosphere must be constant over the source under observation, that is, the angular width of the source should be no greater than the isoplanatic patch size for the atmosphere. The isoplanatic patch is the area of sky within which the path length for an incident wave remains constant to within a small fraction of a wavelength; see also Section 11.9 under *Low-Frequency Mapping*. The size of the isoplanatic patch varies with frequency. At a few hundred megahertz or less it is common to have more than one source within an antenna beam, and these may be separated sufficiently in angle that ionospheric conditions may be different for each one. The closure conditions will then be different for each source, and use of the closure principle then becomes much more complicated than in the single-source case discussed above.

The closure relationships have proved to be very important in synthesis mapping. When applied to unresolved point sources, the phase closure should be zero and the amplitude closure unity. Thus they are useful in checking the accuracy of calibration and examining instrumental effects. For resolved sources they can be used as observables in situations where direct calibration by observation of a calibration source is not practicable, as is sometimes the case in VLBI. Most importantly, they can be used to improve calibration accuracy for observations where high dynamic range is required, as discussed in Section 11.5. The amplitude closure relationships are less frequently used because it is generally easier to calibrate the visibility amplitudes than the phases. However, they provide a useful check in cases where the amplitude is required with especially high accuracy.

10.4 MODEL FITTING

The fitting of intensity models to visibility data was practiced extensively in early radio interferometry, especially when the visibility phase was poorly calibrated or the data were not sufficiently complete to allow Fourier transformation. Examples of simple models are shown in Figs. 1.5, 1.10, and 1.14. In the absence of phase information there is an ambiguity of 180° in position angle of the model. However, there are a number of circumstances in which model fitting offers advantages in the interpretation of interferometer data, as follows:

- Interpretation of VLBI observations with extreme angular resolution, for example, those made between the earth and deep space in which the (u, v) plane may not be well sampled.

- For certain types of sources the radio emission can be specified with reasonable accuracy in terms of a physical model that involves only a small number of parameters. In such cases, when the source is only partly resolved, the parameters of the physical model can best be determined by fitting the model visibility function directly to the observed values. An example is radiation from the extended atmosphere of a star, related to the stellar wind. Measurements by White and Becker (1982) of P Cygni, for which the relative visibility amplitude was not less than ~0.35 at their longest baseline, provide a good example.
- Determination of change in a parameter of a source in which time-separated observations may not have identical (u, v) coverage. Fitting the same model (allowing the parameters of interest to vary) to both data sets is likely to give the best evidence of change. An interesting example is provided by Masson (1986) in a measurement of angular expansion of a compact planetary nebula. From several data sets obtained at different epochs, the image from the one with the best (u, v) coverage was used as a model to fit to the others, thereby avoiding direct comparison of images made with different synthesized beams.
- Determination of probable error in a measured parameter. Processing by nonlinear algorithms, in particular CLEAN (see Section 11.2), which are usually necessary to maximize dynamic range, may result in image-plane noise for which the characteristics are not well understood. However, in model fitting in the (u, v) plane the noise is generally Gaussian.
- Provision of a starting point for maximum entropy deconvolution and self-calibration described in Chapter 11. A simplified model is often all that is necessary.

Basic Considerations for Models

Gaussian functions are convenient model components for source intensity. They are always positive and vary smoothly with angle, as do many of the structures in nebulae and radio galaxies. A circularly symmetrical Gaussian function of half-amplitude width $\sqrt{8 \ln 2}\sigma$, centered at (l_1, m_1), is represented by

$$I_G(l, m) = I_0 \exp\left[\frac{-(l - l_1)^2 - (m - m_1)^2}{2\sigma^2}\right]. \tag{10.34}$$

The corresponding visibility function is

$$\mathcal{V}_G(u, v) = \sqrt{2\pi}\sigma I_0 \exp\left\{-\left[2\pi^2\sigma^2(u^2 + v^2) + j2\pi(ul_1 + vm_1)\right]\right\}. \tag{10.35}$$

The visibility has real and imaginary components that are sinusoidal corrugations, the ridges of which are normal to the radius vector to the point (l_1, m_1) in the image domain. These visibility components are modulated in amplitude by a Gaussian function centered on the (u, v) origin and of width inversely proportional to σ. Examination of the visibility distribution can thus indicate the form

and position of the main intensity components. For discussions and examples of this type of model fitting, see, for example, Maltby and Moffet (1962), Fomalont (1968), and Fomalont and Wright (1974).

A relationship between the visibility function and the moments of the intensity distribution provides some further insight into model fitting. In one dimension, for simplicity, the visibility function can be expressed as a Taylor series:

$$\mathcal{V}(u) = \mathcal{V}(0) + u\mathcal{V}'(0) + \frac{u^2}{2!}\mathcal{V}''(0) + \cdots + \frac{u^n}{n!}\mathcal{V}^{(n)}(0) + \cdots. \tag{10.36}$$

The derivatives of the visibility are related to the moments of the intensity distribution as follows:

$$\mathcal{V}^{(n)}(0) = (-j2\pi)^n \int_{-\infty}^{\infty} l^n I_1(l)\, dl. \tag{10.37}$$

Equation (10.37) follows from a general relationship between the derivatives of a function at the origin and the moments of its Fourier transform [see, e.g., Bracewell (2000)].

The zero-order moment is equal to the flux density S, the odd-order moments contribute to the imaginary components of the visibility, and the even-order moments contribute to the real part. If the source is symmetrical in l, the odd-order terms are zero. If, in addition, the source is only slightly resolved, the decrease in $\mathcal{V}$ results mainly from the second-moment term. Then the source can be represented by any symmetrical model with an appropriate second moment (Moffet 1962). Examination of the visibility functions of simple symmetrical models, as shown in Fig. 1.5, indicates that it is not practical to differentiate between them unless the function is measured down to a visibility amplitude of ~0.25. Differences are best revealed by the rate at which the function dies away with increasing baseline length and by the amplitude of the oscillations as it does so.

A review of methods of model fitting is given by Pearson (1999). After a model is chosen, the next step is to choose a function that will provide a measure of the quality of the fit. Assuming Gaussian errors (noise) in the (u, v) plane, we express the likelihood of the model as a product of Gaussian terms:

$$\prod_{i=1}^{n_d}\left\{\exp\left[-\frac{1}{2}\left(\frac{\mathcal{V}(u_i, v_i) - \overline{M}(u_i, v_i; a_1, \ldots, a_p)}{\sigma_i}\right)^2\right]\right\}, \tag{10.38}$$

where n_d is the number of independent visibility data, $\overline{M}$ is the Fourier transform of the model, which has p parameters a_1 to a_p, and σ_i is the standard deviation of $\mathcal{V}(u_i, v_i)$. Maximizing the likelihood is equivalent to minimizing the negative log of (10.38), that is, minimizing

$$\chi^2 = \sum_{i=1}^{n_d}\left(\frac{\mathcal{V}(u_i, v_i) - \overline{M}(u_i, v_i; a_1, \ldots, a_p)}{\sigma_i}\right)^2. \tag{10.39}$$

Here χ^2 has the usual statistical definition [see, e.g., Taylor (1982)]. Thus, with Gaussian errors, the method becomes one of least squares. Note that this may not apply if interference is present. Methods for computing least-squares solutions are discussed in Appendix 12.1, and can also be found, for example, in Bevington and Robinson (1992). The expected minimum for χ^2 is $n_d - p$, and the standard deviation of χ^2 is $\sqrt{2(n_d - p)}$. If the measured visibility is not satisfactorily calibrated, closure values for the visibility phase and/or amplitude, rather than individual visibility values, can be used in the model fitting.

Cosmic Background Anisotropy

In studies of the anisotropy of the cosmic microwave background radiation (CMB), for example, by using an array of the type shown in Fig. 5.24, the objective is to determine the statistical properties of the angular variation of the brightness temperature. These can be obtained directly from the visibility values without generation of sky brightness images. The required statistics are completely specified by the amplitudes of spherical harmonics, in terms of which the CMB variations can be represented. For a spherical harmonic of order ℓ, the amplitude is proportional to the visibility $\mathcal{V}(u, v)$, where $\ell = 2\pi\sqrt{u^2 + v^2}$. The (u, v) coverage therefore determines the range of ℓ that can be examined. Extending the measurements over several contiguous areas of the primary-element beams, as in mosaicking (see Section 11.6), increases the spatial resolution in u and v, and hence also the resolution of the harmonic amplitudes in ℓ. A major concern is the removal of brightness components resulting from "foreground" objects such as the Galaxy and discrete radio sources; see, for example, White et al. (1999).

10.5 SPECTRAL LINE OBSERVATIONS

General Considerations

A spectral line correlator produces separate visibility measurements at many points across the receiver passband, and for each of these a different intensity distribution can be obtained. The data reduction involved is in principle the same as used in continuum mapping, but differs in some practical details. The number of channels into which the received signal is divided is typically 100–1000. The discussion in this section is largely based upon Ekers and van Gorkom (1984) and van Gorkom and Ekers (1989).

Calibration of the instrumental bandpass response is perhaps the most important step in obtaining accurate spectral line data. Generally the channel-to-channel differences are relatively stable with time and need not be calibrated as frequently as the time-variable effects of the overall receiver gain. The overall gain variations require periodic observation of a calibration source as described for continuum observations. For this purpose the summed response of the individual channels is often used, since a much longer observing time would be required to obtain a sufficient signal-to-noise ratio in each narrow channel. For the bandpass calibration

a longer observation of a calibrator can be made to determine the relative gains of the spectral channels. Since the relative gains of the different channels into which the bandpass is divided change very little with time, the bandpass calibration need only be performed once or twice during, say, an eight-hour observation. The bandpass calibration source should be unresolved, strong enough to provide good signal-to-noise ratio in the spectral channels, and have a sufficiently flat spectrum. However, it need not be close in position to the source being mapped.

Bandpass ripples resulting from standing waves between the antenna feed and the reflector, which pose a serious problem for single-antenna total-power systems, are much less important for interferometers. This is because the instrumental noise, including thermal noise picked up in the antenna sidelobes, is not correlated between antennas. On the other hand, for digital correlators, the Gibbs-phenomenon ripples in the bandpass, which arise in Fourier transformation from the delay to the frequency domains, introduce a problem not found in autocorrelators. Because the cross-correlation of the signals from two antennas is real but not symmetrical as a function of delay, the cross power spectrum as a function of frequency is complex. (The autocorrelation function of the signal from a single antenna is real and symmetrical, and the power spectrum is real.) As explained in Section 8.7 (see Fig. 8.12), the imaginary part of the cross power spectrum changes sign at the origin, but the real part does not. Because of this large discontinuity at the frequency origin, ripples in the imaginary part of the frequency spectrum are of larger relative amplitude than those in the real part. The peak overshoot in the imaginary part is 18% (9% of the full step size); see also Bos (1984, 1985). Figure 10.6 shows a calculated example. The ratio of the real and

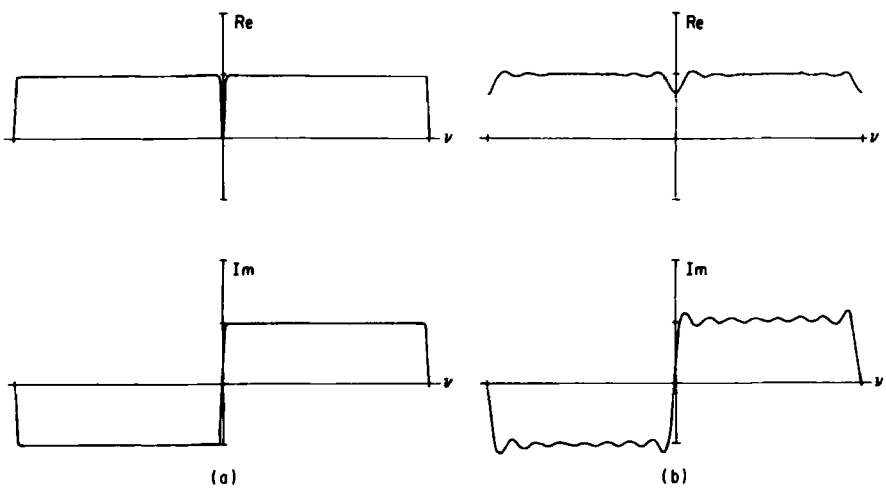


Figure 10.6 (a) The cross power spectrum resulting from a continuum source in which the phase is arbitrarily chosen such that the amplitudes of the real and imaginary parts are equal. (b) Computed response of a cross-correlator with 16 channels to the spectrum in (a). Note the difference in amplitude of the ripples in the real and imaginary parts. From D'Addario (1989), courtesy of the Astron. Soc. Pacific Conf. Ser.

imaginary parts depends on the instrumental phase (which is not calibrated out at this stage of the analysis), and on the position of the source of the radiation relative to the phase center of the field.

Increasing the number of lags of a lag correlator, or the size of the FFT in an FX correlator, improves the spectral resolution and confines the Gibbs-phenomenon ripples more closely to the passband edges. The data from the channels at the band edges are often discarded because of the ripples and the roll-off of the frequency response. An effective way to reduce the amplitude of the ripples is to taper the cross-correlation function and thus introduce smoothing into the cross power spectrum. For this smoothing the Hanning function (see Table 8.4) is often used. Van Gorkom and Ekers (1989) draw attention to the following examples:

1. If the field contains a line source but no continuum, and the line is confined to the central part of the passband, then the spectrum has no discontinuity at the passband edges. This is the only case where it is advisable to use different tapering of the cross-correlation function for the source and the continuum calibrator.
2. If in addition to the line source the field contains one continuum point source, and if both this source and the bandpass calibrator are at the centers of their respective fields, then an accurate calibration of the bandpass ripples is possible. The same weighting must be used for the source and calibrator.
3. In more complicated cases, for example, when there is both a line source and an extended continuum source within the field, the ripples will be different in the two cases and exact calibration is not possible. Hanning smoothing of the spectra of both the source and the calibrator is recommended.

VLBI Observations of Spectral Lines

Since VLBI observations are limited to sources of very high brightness temperature, spectral line measurements in VLBI are used mainly for the study of masers and absorption of emission from bright extragalactic sources by molecular clouds. Frequently observed maser lines include those arising from OH, H_2O, CH_3OH, and SiO. For absorption studies, many atomic and molecular species can be observed since the brightness temperature requirement is fulfilled by the background source. The formalism of spectral line signal processing is described in Section 9.3. Special considerations for astrometric measurements are given in Section 12.5. Here we discuss several practical issues related to the handling of spectroscopic data. The use of independent frequency standards at the antennas results in time-dependent timing errors, which introduce linear phase slopes across the basebands. The difference in Doppler shifts among the antennas can be large, and hence the residual fringe rates can also be large, which may necessitate short integration times for calibration. For masers, the phase calibration can usually be obtained from the use of the phase of a particular spectral feature as a reference. The amplitude calibration can be obtained from the measurement of

the spectra derived from the data recorded at individual antennas. More details of procedures for handling spectral line data can be found in Reid (1995, 1999).

In spectral line VLBI it is usual to observe a compact continuum calibrator several times an hour, preferably one strong enough to give an accurate fringe measurement in one or two minutes of integration. If a lag-type correlator is used to cross-correlate the signals, the output is a function of time and delay. Equation (9.16), in which $\Delta\tau_g$ and θ_{21} are functions of time, shows cross-correlation as a function of time and delay. By Fourier transformation, the arguments t and τ can be changed to the corresponding conjugate variables, which are fringe frequency (or fringe rate) ν_f and the frequency of the spectral feature ν, respectively. Thus the correlator output can be expressed as a function of (t, τ), (ν_f, τ), (t, ν), or (ν_f, ν), and can be interchanged between these domains by Fourier transformation. This is important because some steps in the calibration are best performed in particular domains. Note that the fringe frequency in VLBI observations results mainly from the difference between the true fringe frequency and the model fringe frequency used to stop the fringes. Consider first the data from the continuum calibrator. In fringe fitting for a continuum source it is advantageous to use visibility data as a function of fringe frequency and delay, (ν_f, τ), as shown in Fig. 9.4. In that domain the visibility data are most compactly concentrated and therefore most easily identified in the presence of the noise. In the absence of errors the visibility will be concentrated at the origin in the (ν_f, τ) domain. A shift from the origin in the τ coordinate indicates timing errors resulting from clock offsets or baseline errors. The shift $\Delta\tau$ represents the difference in the errors for the two antennas. Values of $\Delta\tau$ determined from the continuum calibrator are used to apply corrections to the spectral line data. Variation of the $\Delta\tau$ values over time requires interpolation to the times of the spectral line data. The continuum data can also be used for bandpass calibration, to determine the relative amplitude and phase characteristics of the spectral channels.

For fringe fitting the spectral line data, it is advantageous to transform to the (t, ν) domain since, in contrast to the continuum case, the spectral line data represent features that are narrow in frequency. The cross-correlation function is therefore correspondingly broad in the delay dimension, and generally more compact in frequency. Note that in the τ-to-ν transformation, ν is not the frequency of the radiation as received at the antenna, since the frequency of a local oscillator (or a combination of more than one local oscillators) ν_{LO} has been subtracted. Thus ν here represents the frequency within the IF band that is sampled and recorded for transmission to the correlator. The (t, ν) domain is also appropriate for inserting corrections for the timing errors $\Delta\tau$ determined from the continuum data. These corrections are made by inserting phase offsets that are proportional to frequency. Thus the data as a function of (t, ν) are multiplied by† $\exp(j2\pi\Delta\tau\nu)$. If the variation in the $\Delta\tau$ values over time results from a *clock rate* error at one or both of the antennas, correction should be made for the associated error in the frequency ν_{LO} at the antennas. The resulting phase error is corrected by multiplying the correlator output data by $\exp(j2\pi\Delta\tau\nu_{LO})$.

† Note that the required sign of the exponent in this and similar expressions used in this subsection may be positive or negative depending on other sign conventions used.

Since Doppler shift corrections (see Appendix 10.2) are rarely made as local oscillator offsets at the antennas, these corrections must be made at the correlator or subsequently in the post-processing analysis. The diurnal Doppler shift is normally removed at the station level in the precorrelation fringe rotation, where the signals are delayed and frequency-shifted to a reference point at the center of the earth. Correction for the Doppler shift due to the earth's orbital motion and the local standard of rest, as well as any other frequency offset, can conveniently be made on the post-correlation data by use of the shift theorem, that is, multiplication of the correlation functions by $\exp(j2\pi\Delta\nu\tau)$, where $\Delta\nu$ is the total frequency shift desired.

The visibility spectra can be calibrated in units of flux density by multiplication of the normalized visibility spectra by the geometric mean of the system equivalent flux densities (SEFDs) of the two antennas concerned, as discussed in Section 10.1 under *Use of Calibration Sources*. The SEFD is defined in Eq. (1.6). It can be determined from occasional supplemental measurements at the antennas, and the results interpolated in time. A better method for strong sources is to calculate the total-power spectrum of the source from the autocorrelation functions of the data from each antenna. These must be corrected for the bandpass response, which can be obtained from the autocorrelation functions on a continuum fringe calibrator. The amplitude of a specific spectral feature is proportional to the reciprocal of the SEFD. If greater sensitivity is required, then each measured spectrum can be matched to a spectral template obtained from a global average of all the single-antenna data or from a spectrum taken with the most sensitive antenna in the array. The disadvantage of this method is that it is seldom convenient to acquire bandpass spectra often enough to ensure sufficiently accurate baseline subtraction on weak sources.

If the total frequency bandwidth in the measurements is covered by using two or more IF bands of the receiving system, it is necessary to correct for differences in their instrumental phase responses. This can be done using the continuum calibrator measurements, by averaging the phase values for the different channels in each IF band, and subtracting these averages from the corresponding spectral line visibility data.

Finally, it is necessary to correct for remaining instrumental phases and for the different atmospheric and ionospheric phase shifts, which may be large for widely separated sites. In mapping strong continuum sources, this can be achieved by using phase closure as described in Section 10.3. A similar approach can be used in mapping a distribution of maser point sources, by selecting a strong spectral component that is seen at all baselines and assuming that it represents a single point source. Then if the phase for this component at one arbitrarily chosen antenna is assumed to be zero, the relative phases for the other antennas can be deduced from the fringe phases. Since these phases are attributed to the atmosphere over each antenna, the correction can be applied to all frequency components within the measured spectrum. This method of using one maser component to provide a phase reference is discussed in more detail in Section 12.5, together with fringe frequency mapping, a technique that is useful in determining the positions of major components in a large field of masers.

Variation of Spatial Frequency over the Bandwidth

The effect of using the center frequency of the receiver passband in calculating the values of u and v for all frequencies within the passband is discussed in Section 6.3. Consider, for example, a single discrete source for which the visibility function has a maximum centered on the (u, v) origin and decreases monotonically for a range of increasing u and v. If we use the frequency at the band center ν_0 to calculate u and v for a frequency at the high end of the band, that is, $\nu > \nu_0$, then the values of u and v will be underestimated. The measured visibility will fall off too quickly with u and v and the central peak of the visibility function will be too narrow. Hence the width of the image in l and m will be too wide. Thus if the source radiates a spectral line at the blue-shifted side of the bandwidth the angular dimensions may be overestimated, and similarly underestimated at the red-shifted side. This effect can be described as *chromatic aberration*.

As discussed in Section 6.3, for observations with a spectral line (multichannel) correlator the visibility measured for each channel can be expressed as a function of the (u, v) values appropriate for the frequency of the channel. This corrects the chromatic aberration, but causes the (u, v) range over which the visibility is measured to increase over the bandwidth in proportion to the frequency. Thus the width of the synthesized beam (i.e., the angular resolution) and the angular scale of the sidelobes vary over the bandwidth. The variation of the resolution can, if necessary, be corrected by truncation or tapering of the visibility data to reduce the resolution to that of the lowest frequency within the passband.

Accuracy of Spectral Line Measurements

The *spectral dynamic range* of an image after final calibration is an estimate of the accuracy of the measurement of spectral features expressed as a fraction of the maximum signal amplitude. It can be defined as the variation in the response of different channels to a continuum signal divided by the maximum response, the variation being a result of noise and instrumental errors. When the amplitude of a spectral line is only a few percent of the continuum that is present, as in the case of a recombination line or a weak absorption line, the accuracy of spectral line features depends on the accuracy with which the response to the continuum can be separated from that to the line. In such a case a dynamic range of order 10^3 is required to measure a line profile to an accuracy of 10%. Hence the importance of accurate bandpass calibration and of correction for chromatic aberration.

Various techniques have been used to help subtract the continuum response from a map. It is necessary to choose the receiver bandwidth so as to include some channels that contain continuum only, at frequencies on either side of the line features. A straightforward method is to use an average of the line-free channel data to make a continuum map, and subtract this map from each of the maps derived for a channel with line emission. Unless the receiver bandwidth is sufficiently small compared to the center frequency, it is likely that a correction for chromatic aberration should be used in making the continuum map. If the contin-

uum emanates from point sources, the positions and flux densities of the sources provide a convenient model. For the most precise subtraction, the continuum response should be calculated separately for each line channel, using the individual channel frequencies in determining the (u, v) values. The subtraction should be performed in the visibility data. Use of deconvolution algorithms in the continuum subtraction is briefly discussed in Section 11.9 under *Use of CLEAN and Self-Calibration with Spectral Line Data.*

Presentation and Analysis of Spectral Line Observations

Spectral line data can be presented as three-dimensional distributions of pixels in (l, m, v). For physical interpretation, the Doppler shift in the frequency dimension is often converted to radial velocity v_r with respect to the rest frequency of the line. The relationship between frequency and velocity is given in Appendix 10.2. A model of such a three-dimensional distribution is shown in Fig. 10.7. Continuum sources are represented by cylindrical functions of constant cross section in l and m.

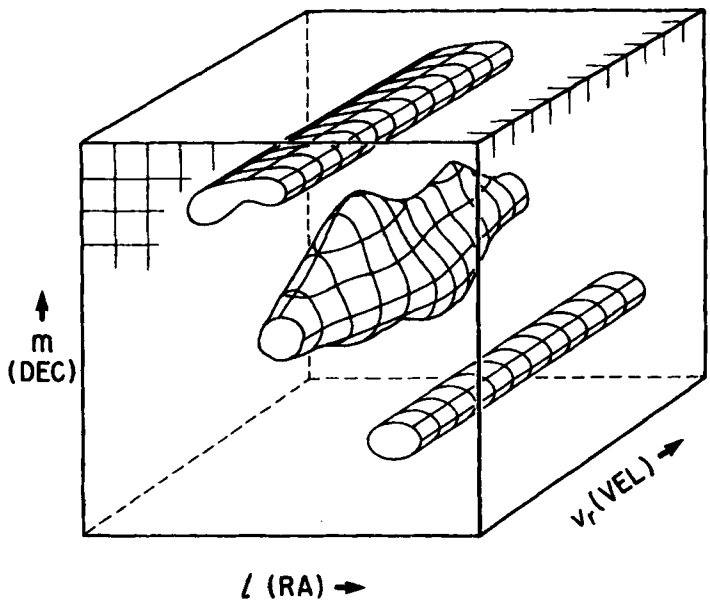


Figure 10.7 Three-dimensional representation of spectral line data in right ascension, declination, and frequency. The frequency axis is calibrated in velocity corresponding to the Doppler shift of the rest frequency of the line. The flux density or intensity of the radiation is not shown but could be represented by color or shading. The indicated velocity has no physical meaning for continuum sources, which are represented by cylindrical forms of constant cross section normal to the velocity dimension. Spectral line emission is indicated by the variation of position or intensity with velocity. From Roelfsema (1989), courtesy of the Astron. Soc. Pacific Conf. Ser.

The three-dimensional data cube that contains the maps for the individual channels can be thought of as representing a line profile for each pixel in two-dimensional (l, m) space. To simplify the ensemble of maps it is often useful to plot a single (l, m) map of some feature of the line profile. This feature might be the integrated intensity

$$\Delta\nu \sum_i I_i(l, m), \tag{10.40}$$

where i indicates the spectral channels, which are spaced at intervals $\Delta\nu$ in frequency. For an optically thin radiating medium such as neutral hydrogen, this is proportional to the column density of radiating atoms or molecules. The intensity-weighted mean velocity is an indicator of large scale motion,

$$\langle v_r(l, m)\rangle = \frac{\sum_i I_i(l, m) v_{r_i}}{\sum_i I_i(l, m)}. \tag{10.41}$$

The intensity-weighted velocity dispersion

$$\sqrt{\frac{\sum_i I_i(l, m)(v_{r_i} - \langle v_r\rangle)^2}{\sum_i I_i(l, m)}} \tag{10.42}$$

is an indicator of random motions within the source. The summation in the velocity dimension is performed separately for each (l, m) pixel of the maps. In each of the three quantities in (10.40) to (10.42) the intensity values correspond to the specific line of interest, continuum features having been separated out. In obtaining the best estimates for these three quantities, it should be noted that including ranges of (l, m, v_r) that contain no discernable emission only add noise to the results.

Exploring the relationships between three-dimensional images in (l, m, v_r) and the three-dimensional distribution of the radiating material is an astronomical concern. As a simple example, consider a spherical shell of radiating material. If the material is at rest, it will appear in (l, m, v_r) space as a circular disk in the plane of zero velocity, with brightening at the outer edge. If the shell is expanding with the same velocity in all directions, it will appear in (l, m, v_r) space as a hollow ellipsoidal shell or, with appropriate adjustment of the velocity scale, a spherical shell. Interpretation of observations of rotating spiral galaxies are more complex. An example of a model galaxy is given by Roelfsema (1989) and a more extensive discussion can be found in Burton (1988).

10.6 MISCELLANEOUS CONSIDERATIONS

Interpretation of Measured Intensity

The quantity measured in a synthesized map is the radio intensity, but $\mathcal{V}$ is usually calibrated in terms of the equivalent flux density of a point source, and the

intensity unit in the resulting map is in units of flux density per beam area Ω_0, which is given by

$$\Omega_0 = \int\int_{\substack{\text{main}\\ \text{lobe}}} \frac{b_0(l, m)\, dl\, dm}{\sqrt{1 - l^2 - m^2}}. \tag{10.43}$$

The response to an extended source is the convolution of the sky intensity $I(l, m)$ with the synthesized beam $b_0(l, m)$. Note that since there is often no measured visibility value at the (u, v) origin, the integral of $b_0(l, m)$ over all angles is zero; that is to say, there is no response to a uniform level of intensity. At any point on the extended source where the intensity varies slowly compared with the width of the synthesized beam, the convolution with $b_0(l, m)$ results in a flux density that is approximately $I\Omega_0$. Thus the scale of the map can also be interpreted as intensity measured in units of flux density per beam area Ω_0. For a discussion of mapping wide sources and measuring the intensity of extended components of low spatial frequency, see Section 11.6.

Errors in Maps

A very useful technique for investigating suspicious or unusual features in any synthesis image, continuum or spectral line, is to compute an inverse Fourier transform (i.e., from intensity to visibility) including only the feature in question. A distribution in the (u, v) plane concentrated in a single baseline, or in a series of baselines with a common antenna, could indicate an instrumental problem. A distribution corresponding to a particular range of hour angle of the source could indicate the occurrence of sporadic interference.

An aid in identifying erroneous features is a familiarity with the behavior of functions under Fourier transformation; see, for example, Bracewell (2000) and the discussion by Ekers (1999). A persistent error in one antenna pair will, for an east–west spacing, be distributed along an elliptical ring centered on the (u, v) origin, and in the (l, m) plane will give rise to an elliptical feature with a radial profile in the form of the zero-order Bessel function. An error of short duration on one baseline introduces two delta functions representing the measurement and its conjugate. In the image these produce a sinusoidal corrugation over the (l, m) plane. The amplitude in the image plane may be only small, since in an $M \times N$ visibility matrix the effect of the two erroneous points is diluted by a factor of $2(MN)^{-1}$, which is usually of order 10^{-3}–10^{-6}. Thus a single short-duration error could be acceptable if, in the image plane, it is small compared with the noise.

Errors of an additive nature combine by addition with the true visibility values. In the map the Fourier transform of the error distribution $\varepsilon_{\text{add}}(u, v)$ is added to the intensity distribution, and we have

$$\mathcal{V}(u, v) + \varepsilon_{\text{add}}(u, v) \rightleftharpoons I(l, m) + \bar{\varepsilon}_{\text{add}}(l, m). \tag{10.44}$$

Other types of additive errors result from interference, cross-coupling of system noise between antennas, and correlator offset errors. The sun is many orders of magnitude stronger than most radio sources and can produce interference of a different character from that of terrestrial sources because of its diurnal motion. The response to the sun is governed mainly by the sidelobes of the primary beam, the difference in fringe frequencies for the sun and the target source, and the bandwidth and visibility averaging effects. Solar interference is most severe for low-resolution arrays with narrow bandwidths. Cross-coupling of noise (crosstalk) occurs only between closely spaced antennas and is most severe for low elevation angles when shadowing of antennas may occur.

A second class of errors comprises those that combine with the visibility in a multiplicative manner, and for these we can write

$$\mathcal{V}(u, v)\varepsilon_{\text{mul}}(u, v) \rightleftharpoons I(l, m) * * \overline{\varepsilon}_{\text{mul}}(l, m). \tag{10.45}$$

The Fourier transform of the error distribution is convolved with the intensity distribution, and the resulting distortion produces erroneous structure connected with the main features in the map. In contrast, the distribution of errors of the additive type is unrelated to the true intensity pattern. Multiplicative errors mainly involve the gain constants of the antennas, and result from calibration errors including antenna pointing and, in the case of VLBI systems, radio interference (see Section 15.3).

Distortions that increase with distance from the center of the map constitute a third category of errors. These include the effects of non-coplanar baselines (see Sections 3.1 and 11.8), bandwidth (see Section 6.3), and visibility averaging (see Section 6.4), which are predictable and therefore somewhat different in nature from the other distortions mentioned above.

Hints on Planning and Reduction of Observations

Making the best use of synthesis arrays and similar instruments requires an empirical approach in some areas, and the best procedures for analyzing data are often gained by experience. Much helpful information exists in the handbooks on specific instruments, symposium proceedings, and so on; see, for example, Bridle (1989). A few examples are discussed below.

In choosing the observing bandwidth for continuum observations, the radial smearing effect should be considered, since the signal-to-noise ratio for a point source near the edge of the field is not necessarily maximized by maximizing the bandwidth. Then in choosing the data-averaging time the resulting circumferential smearing can be about equal to the radial effect. The required condition is obtained from Eqs. (6.75) and (6.80) and for high declinations is

$$\frac{\Delta\nu}{\nu_0} \simeq \omega_e \tau_a. \tag{10.46}$$

Here ν_0 is the center frequency of the observing band, $\Delta\nu$ is the bandwidth, ω_e is the earth's rotation velocity, and τ_a is the averaging time. When attempting to detect a weak source of measurable angular diameter, or an extended emission, it is important not to choose an angular resolution that is too high. The signal-to-noise ratio for an extended source is approximately proportional to $I\Omega_0$, as discussed in the previous section. The observing time required to obtain a given signal-to-noise ratio is proportional to Ω_0^{-2}, or to θ_b^{-4}, where θ_b is the synthesized beamwidth.

If the antenna beam contains a source that is much stronger than the features to be studied, the response to the strong source can be subtracted, provided it is a point source or one that can be accurately modeled. This is best done by subtracting the computed visibility before gridding the measurements for the FFT. The subtracted response will then accurately include the effect of the sidelobes of the synthesized beam. Nevertheless, the precision of the operation will be reduced if the source response is significantly affected by bandwidth, visibility averaging, and similar effects, so it may be best to place the source to be subtracted at the center of the field.

When observing a very weak source, it may be advisable to place the source a few beamwidths away from the (l, m) origin to avoid confusion with residual errors from correlator offsets. Experience with any particular instrument will show whether this is necessary.

As part of the procedure in making any map it is useful also to make a low-resolution map covering the entire area of the primary antenna beam. For this map, the data can be heavily tapered in the (u, v) plane to reduce the resolution and thus also the computation. Such a map will reveal any sources outside the field of the final map that may introduce aliased responses in the FFT. Aliasing of these sources can be suppressed by subtraction of their visibility or use of a suitable convolving function. The sidelobe or ringlobe responses to such a source are also eliminated by subtraction of the source, but not by convolution in the (u, v) plane. The low-resolution map will also emphasize any extended low-intensity features that might otherwise be overlooked.

APPENDIX 10.1 THE EDGE OF THE MOON AS A CALIBRATION SOURCE

During the test phase of bringing an interferometer into operation, it is useful to observe sources that produce fringes with high signal-to-noise ratio. At frequencies above ~100 GHz there are not many such sources. The sun, moon, and planets, the disks of which are resolved by the interferometer fringes, can nevertheless provide significant correlated flux density because of their sharp edges. Consider the limb of the moon, and the case where the primary beam of the interferometer elements is much smaller than 30 arcmin, the lunar diameter. When the antenna beam tracks the moon's limb, the apparent source distribution is the antenna pattern multiplied by a step function; it is assumed that the brightness temperature of the moon is constant within the beam. Approximating the antenna

pattern as a Gaussian function, assuming that the antenna is pointed at the west limb of the moon, and ignoring the curvature of the lunar limb, we can express the effective source distribution as

$$\begin{aligned} I(x, y) &= I_0 e^{-4(\ln 2)(x^2+y^2)/\theta_b^2} \qquad & x \geq 0, \\ &= 0 & x < 0, \end{aligned} \tag{A10.1}$$

where x and y are angular coordinates centered on the beam axis, θ_b is the full width of the beam at the half-power level, and in the Rayleigh–Jeans regime $I_0 = 2kT_m/\lambda^2$, where T_m is the temperature of the moon. The visibility function is then

$$\begin{aligned} \mathcal{V}(u, v) = 2I_0 &\left[\int_0^\infty e^{-4(\ln 2)x^2/\theta_b^2}(\cos 2\pi ux - j \sin 2\pi ux)dx\right] \\ &\times \left[\int_0^\infty e^{-4(\ln 2)y^2/\theta_b^2} \cos 2\pi vy\, dy\right], \end{aligned} \tag{A10.2}$$

The cosine integral is straightforward, and the sine integral can be written in terms of a degenerate hypergeometric function ${}_1F_1$ (see Gradshteyn and Ryzhik 1994, Eq. 3.896.3). The result is

$$\mathcal{V}(u, v) = I_0 S_0 e^{-\pi^2\theta_b^2(u^2+v^2)/4\ln 2}\left[1 - j\sqrt{\frac{\pi}{\ln 2}}\,(\theta_b u)\,{}_1F_1\left(\frac{1}{2}, \frac{3}{2}, \frac{\pi^2\theta_b^2 u^2}{4\ln 2}\right)\right], \tag{A10.3}$$

where

$$S_0 = \frac{\pi k \tau_m \theta_b^2}{4\lambda^2 \ln 2} \tag{A10.4}$$

is the flux density of the moon in the half-Gaussian beam. In the limit $(u, v) \gg (0, 0)$, the imaginary part of the visibility is zero and $\mathcal{V}(u, v) = S_0$, as expected. For $T_m = 200$ K and $\theta_b = 1.2\lambda/d$, where d is the diameter of the interferometer antennas in meters, $S_0 \simeq 460{,}000/d^2$ Jy. The integral over x in Eq. (A10.2) can also be written in terms of the error function. For the limit where $u \gg d/\lambda$, the asymptotic expansion of the error function leads to the convenient approximation

$$\mathcal{V}(u, v = 0) = j\sqrt{\frac{4\ln 2}{\pi^3}}\,\frac{S_0}{\theta_b u} \simeq j0.41\frac{kT_m}{dD}, \tag{A10.5}$$

where D is the baseline length. Hence we have the interesting situation that the visibility for a given baseline length increases as the antenna diameter decreases, as long as $\theta_b \ll 30$ arcmin. The approximation in Eq. (A10.5) is accurate to 2% for $D > 2d$. The full visibility function as a function of projected baseline length

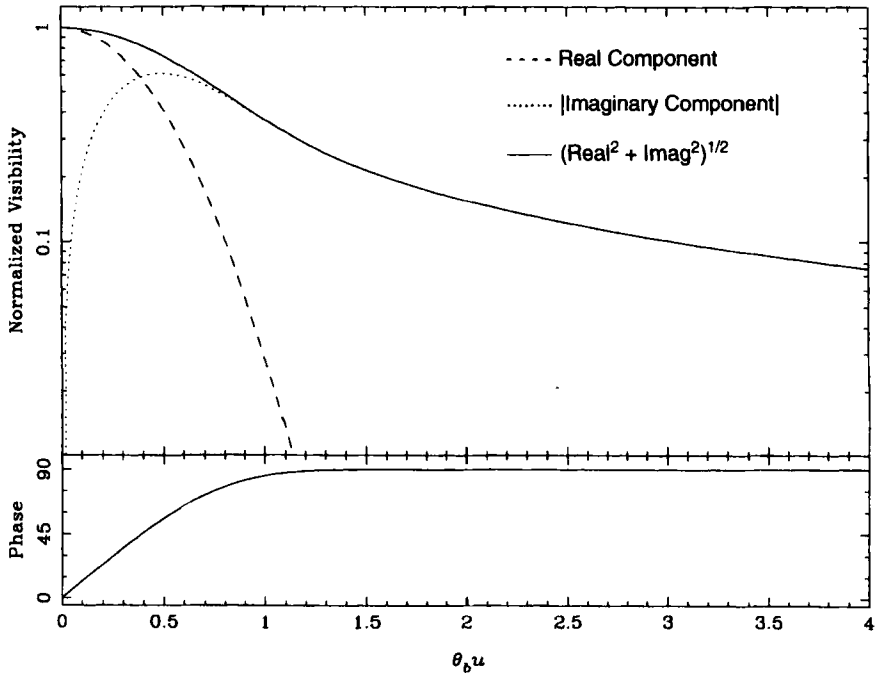


Figure A10.1 Normalized fringe visibility for an interferometer with an east–west baseline observing the west limb of the moon at transit ($v = 0$), versus $\theta_b\, u$, where $\theta_b \simeq 1.2\lambda/d$ is the half-power beamwidth of the antenna, d is the antenna diameter, and $u = D/\lambda$ is the baseline in wavelengths. On the horizontal axis $\theta_b\, u$ is approximately equal to $1.2D/d$. The dotted line is the imaginary component of visibility, the dashed line is the real part, and the solid line is the magnitude. Since the portion of the curve for $D/d < 1$ is not accessible, the measured visibility is almost purely imaginary. For $d = 6$ m and $D/d = 3$, the zero-spacing flux density [see Eq. (A10.4)] is 12,700 Jy, and the visibility is about 1000 Jy [see Eq. (A10.5)]. Adapted from Gurwell (1998).

is shown in Fig. A10.1. Note that the visibility measured with an interferometer having an east–west baseline orientation and tracking the north or south limb of the moon will be essentially zero. In the general case the maximum fringe visibility is obtained by tracking the limb of the moon that is perpendicular to the baseline.

Although the moon may produce strong fringes, it is not an ideal calibration source. First, libration may make it difficult to track the exact edge of the moon. Second, because the apparent source distribution is determined by the antennas, tracking errors introduce amplitude and phase fluctuations. Third, because the temperature of the moon depends upon solar illumination, variations around the mean temperature of 200 K are significant, especially at short wavelengths. For accurate results the lunar temperature variation should be incorporated into the brightness temperature model.

APPENDIX 10.2 DOPPLER SHIFT OF SPECTRAL LINES

Doppler shifts of spectral lines result from the relative velocity between the source and the observer. Four important practical issues are discussed here. First, the use of a first-order expansion of the special relativistic Doppler formula leads to significant errors for large velocities. Second, there are several different approximations in use for correcting measured velocities for the observer's motion. Third, special care must be taken to avoid the introduction of a velocity offset when converting from frequency to velocity. Finally, there are velocity shifts of non-Doppler origin that sometimes need to be considered.

The Doppler shift [e.g., Rybicki and Lightman (1979)] is given by the relation

$$\frac{\lambda}{\lambda_0} = \frac{\nu_0}{\nu} = \frac{1 + \frac{v}{c}\cos\theta}{\sqrt{1 - \left(\frac{v}{c}\right)^2}}, \tag{A10.6}$$

where λ_0 and ν_0 are the rest wavelength and frequency as measured in the reference frame of the source, the corresponding unscripted variables are the wavelength and frequency in the observer's frame, v is the magnitude of the relative velocity between the source and the observer, and θ is the angle between the velocity vector and the line-of-sight direction between source and observer in the observer's frame. The numerator in Eq. (A10.6) is the classical Doppler shift caused by the change in distance between the source and the observer. The denominator is the relativistic time dilation factor, which takes account of the difference between the period of the radiated wave as measured in the rest frame of the source and the rest frame of the observer ($\theta < 90°$ for a receding source).

Because of the time dilation effect, there will be a second-order Doppler shift even if the motion is transverse to the line of sight. For the rest of this discussion we consider only radial velocities; that is, $\theta = 0$ or $180°$. In this case the Doppler shift equation is

$$\frac{\lambda}{\lambda_0} = \frac{\nu_0}{\nu} = \sqrt{\frac{1 + \frac{v_r}{c}}{1 - \frac{v_r}{c}}}, \tag{A10.7}$$

where v_r is the radial velocity (positive for recession). Solving for velocity,

$$\frac{v_r}{c} = \frac{\nu_0{}^2 - \nu^2}{\nu_0{}^2 + \nu^2}, \tag{A10.8}$$

or

$$\frac{v_r}{c} = \frac{\lambda^2 - \lambda_0^2}{\lambda^2 + \lambda_0^2}. \tag{A10.9}$$

Taylor expansions of Eqs. (A10.8) and (A10.9) yield

$$\frac{v_r}{c} \simeq -\frac{\Delta\nu}{\nu_0} + \frac{1}{2}\frac{\Delta\nu^2}{\nu_0^2} \cdots \tag{A10.10}$$

and

$$\frac{v_r}{c} \simeq \frac{\Delta\lambda}{\lambda_0} - \frac{1}{2}\frac{\Delta\lambda^2}{\lambda_0^2} \cdots, \tag{A10.11}$$

where $\Delta\nu = \nu - \nu_0$ and $\Delta\lambda = \lambda - \lambda_0$. For negative $\Delta\nu$, the velocity is positive, or "redshifted." Since $\Delta\nu/\nu_0 \simeq -\Delta\lambda/\lambda_0$, the second-order terms have approximately the same magnitude but opposite signs in Eqs. (A10.10) and (A10.11).

Devices for spectroscopy at radio and optical frequencies usually produce data that are uniformly spaced in frequency and wavelength, respectively. Hence, to first order, the velocity axis can be calculated as a linear transformation of the frequency or wavelength axes. Unfortunately, this has led to two different approximations of the velocity:

$$\frac{v_{r\,\mathrm{radio}}}{c} = -\frac{\Delta\nu}{\nu_0} \tag{A10.12}$$

$$\frac{v_{r\,\mathrm{optical}}}{c} = \frac{\Delta\lambda}{\lambda_0}. \tag{A10.13}$$

The difference between these two approximations can be appreciated by noting that $v_{r\,\mathrm{radio}}/c = -\Delta\lambda/\lambda$. Each velocity scale produces a second-order error in its estimation of the true velocity; that is, the radio definition underestimates the velocity, and the optical definition overestimates the velocity by the same amount. The difference in velocity between the scales as a function of velocity is

$$\delta v_r = v_{r\,\mathrm{optical}} - v_{r\,\mathrm{radio}} \simeq \frac{{v_r}^2}{c}. \tag{A10.14}$$

Hence, the identification of the velocity scale used is very important for extragalactic sources. For example, if $v_r = 10,000$ km s^{-1}, $\delta v_r \simeq 330$ km s^{-1}. Failure to recognize the difference between the velocity conventions can cause considerable problems when observations are made with narrow bandwidth.

To interpret the velocities of spectral lines it is necessary to refer them to an appropriate inertial frame. The rotation velocity of an observer at the equator about the earth's center is about 0.5 km s^{-1}; the velocity of the earth around the sun is about 30 km s^{-1}; the velocity of the sun with respect to the nearby stars is about 20 km s^{-1} [this defines the local standard of rest (LSR)]; the velocity of the LSR around the center of the Galaxy is about 220 km s^{-1}; the velocity of our Galaxy with respect to the local group is about 310 km s^{-1}; and the velocity

TABLE A10.1 Reference Frames for Spectroscopic Observations

Name	Type of Motion	Motion (km s^{-1})	Direction[a] ℓ(°)	b(°)
Topocentric	Rotation of earth	0.5	—	—
Geocentric	Rotation of earth around earth/moon barycenter	0.013	—	—
Heliocentric	Rotation of earth around sun	30	—	—
Barycentric	Rotation of sun around solar system barycenter (planetary perturbations)	0.012	—	—
Local standard of rest[b]	Solar motion with respect to local stars	20	57	23
Galactocentric[c]	LSR around center of the Galaxy	220	90	0
Local galactic standard of rest[b]	Galactic center motion with respect to galaxies of local group	310	146	−23
CMB[b]	Local group of galaxies with respect to CMB	630	276	30

[a]Galactic longitude and latitude.
[b]Standard value adopted by the IAU in 1985 (Kerr and Lynden-Bell 1986).
[c]Cox (2000).

of the local group with respect to the cosmic microwave background radiation (CMB) is about 630 km s^{-1}. The most accurate reference frame beyond the solar system is defined with respect to the CMB. The velocity of the sun with respect to the cosmic microwave background has been determined from measurements of the dipole anisotropy of the CMB, which yields the remarkably precise result of 370.6 ± 0.4 km s^{-1} toward $\ell = 264.31° \pm 0.17°$ and $b = 48.05° \pm 0.10°$ (Lineweaver et al. 1996). Information on these various reference frames is listed in Table A10.1. Most observations are reported with respect to either the solar system barycenter or the local standard of rest. Velocities of stars and galaxies are usually given in the former frame, and observations of non-stellar Galactic objects (e.g., molecular clouds) are usually given in the latter frame. Velocity corrections at many radio observatories are based on a program called DOP [Ball (1969); see also Gordon (1976)], which has an accuracy of ~0.01 km s^{-1} because it does not take planetary perturbations into account. Routines such as CVEL in AIPS are based on this code. Much higher accuracy can be obtained by more sophisticated programs such as the Planetary Ephemeris Program (Ash 1972) or the JPL Ephemeris (Standish and Newhall 1995). Interpretation of pulsar timing measurements also requires precise velocity correction.

There is sometimes confusion in the conversion of baseband frequency to true observed frequency. In the calculation of the spectrum in the baseband by

Fourier transformation of either the data stream or the correlation function with the Fast Fourier Transformation algorithm, the first channel corresponds to zero frequency, the channel increment is $\Delta\nu_{\rm IF}/N$, where $\Delta\nu_{\rm IF}$ is the bandwidth (half the Nyquist sampling rate) and N is the total number of frequency channels. The Nth channel corresponds to frequency $\Delta\nu_{\rm IF}(1 - 1/N)$. If N is an even number (N is usually a power of 2), channel $N/2$ corresponds to the center frequency of the baseband. For a system with only upper-sideband conversions, the sky frequency of the first channel (zero frequency in the baseband) is the sum of the local oscillator frequencies. Note that the velocity axes run in opposite directions ($v \propto -\nu$ and $v \propto \nu$) for systems with net upper- and lower-sideband conversion, respectively.

There are several velocity shifts of non-Doppler origin that sometimes need to be taken into account. For spectral lines originating in deep potential wells—for example, close to black holes—there is an additional time dilation term

$$\gamma_G = \frac{1}{\sqrt{1 - \frac{r_s}{r}}}, \tag{A10.15}$$

where r is the distance from the center of the black hole and r_s is its Schwarzschild radius ($r_s = 2GM/c^2$), which is valid for $r \gg r_s$. The total frequency shift [obtained by generalizing Eq. (A10.6)] is therefore

$$\frac{\nu_0}{\nu} = (1 + \frac{v_r}{c}\cos\theta)\gamma_L\gamma_G, \tag{A10.16}$$

where $\gamma_L = 1/\sqrt{1 - v_r{}^2/c^2}$ is the so-called Lorentz factor. For example, the radiation from the water masers in NGC4258 (see Fig. 1.21), which orbit a black hole at a radius of 40,000 r_s, undergoes a velocity shift of about 4 km s^{-1}.

The most important non-Doppler frequency shift for sources at cosmological distances is due to the expansion of the universe. In the relatively nearby universe this velocity shift is

$$z = \frac{\lambda}{\lambda_0} - 1 \simeq \frac{H_0 d}{c}, \tag{A10.17}$$

where H_0 is the Hubble constant and d is the distance. H_0 is thought to be about 70 km s^{-1} Mpc^{-1} (Mould et al. 2000). For greater distances ($z > 1$), the relations between z and the distance and look-back time depend on the cosmological model used [e.g., Peebles (1993)]. However, given the definition of z, the correct frequency will always be related to it by

$$\nu = \frac{\nu_0}{z + 1}. \tag{A10.18}$$

Other issues regarding observations of cosmologically distant spectral line sources are discussed by Gordon, Baars, and Cocke (1992). An example of spec-

troscopic interferometric observations of a molecular cloud at a cosmological distance ($z = 3.9$) can be found in Downes et al. (1999).

APPENDIX 10.3 HISTORICAL NOTES

Maps from One-Dimensional Profiles

Early maps of the sun and a few other strong sources were made with linear arrays such as the grating array and compound interferometer shown in Fig. 1.13. The results were obtained in the form of fan beam scans. With such an instrument the visibility data sampled at any instant are located on a straight line through the origin in the (u, v) plane, as shown in Fig. 10.1. Fourier transformation of the visibility data sampled along such a line provides a corrugated surface with a profile given by the fan beam scan, as shown in Fig. A10.2. This can be regarded as one component of a two-dimensional map of the sun. As the earth rotates, the angle of the beam on the sky varies, so addition of these components builds up a two-dimensional map. However, in the fan beam scans from such arrays, each pair of antennas contributes with equal weight to the profile, so a map built up from profiles in such a manner exhibits the undesirable characteristics of natural weighting. During the 1950s, before digital computers were generally available, the combination of such data to provide two-dimensional maps with a desirable weighting was a laborious process. Christiansen and Warburton's (1955)

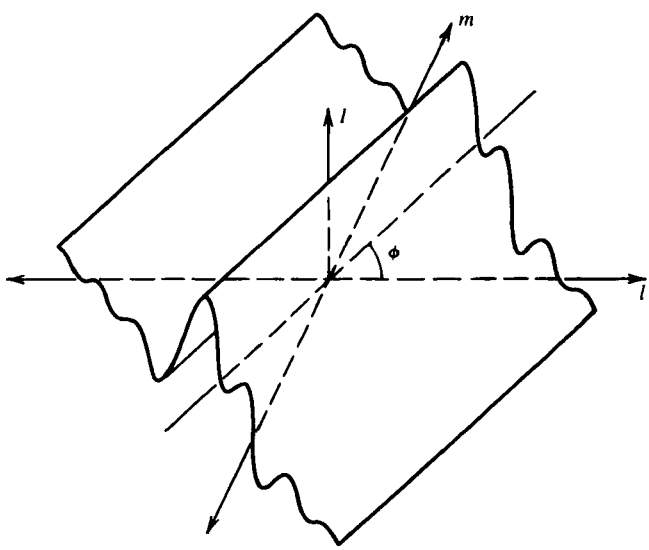


Figure A10.2 A surface in the (l, m) domain that is the Fourier transform of visibility data in the (u, v) plane measured along a line making an angle $\phi + \pi/2$ with the u axis, as shown by the broken line in Fig. 10.1.

solar map involved Fourier transformation, weighting, and retransformation of the data by manual calculation. A method of combining fan-beam scans without Fourier transformation was later devised by Bracewell and Riddle (1967) using convolution to adjust the visibility weighting. Basic relationships between one- and two-dimensional responses (Bracewell 1956b) are discussed in Section 2.4. The same concepts are applicable to image processing in other fields, for example, tomography (Bracewell and Wernecke 1975).

Analog Fourier Transformation

An optical lens can be used as an analog device for Fourier transformation. Analog systems for data processing based on optical, acoustic, or electron-beam processes have been investigated, but generally have not proved successful for synthesis imaging. They lack flexibility, and a further problem is limitation of the *dynamic range*, which is the ratio of the highest intensity levels to the noise in the image. Maintaining image quality in any iterative process that involves successive Fourier transformation and retransformation of the same data, as occurs in some deconvolution processes (see Chapter 11), requires high precision. Analog possibilities for Fourier transformation are discussed by Cole (1979).

BIBLIOGRAPHY

Perley, R. A., F. R. Schwab, and A. H. Bridle. Eds., *Synthesis Imaging in Radio Astronomy*, Astron. Soc. Pacific Conf. Ser., **6**, 1989.

Taylor, G. B., C. L. Carilli, and R. A. Perley, Eds., *Synthesis Imaging in Radio Astronomy II*, Astron. Soc. Pacific Conf. Ser., **180**, 1999.

Thompson, A. R. and L. R. D'Addario, Eds., *Synthesis Mapping, Proc. NRAO Workshop No. 5* (Socorro, NM, June 21–25, 1982), National Radio Astronomy Observatory, Green Bank, WV, 1982.

REFERENCES

Ash, M. E., *Determination of Earth Satellite Orbits*, MIT Lincoln Laboratory Technical Note, 1972–5, 1972.

Ball, J. A., *Some Fortran Subprograms Used in Astronomy*, MIT Lincoln Laboratory Technical Note, 1969–42, 1969.

Bevington, P. R. and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed., McGraw-Hill, New York, 1992.

Bos, A., On Ghost Source Mechanisms in Spectral Line Synthesis Observations with Digital Spectrometers, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Univ. Press, 1984, pp. 239–243.

Bos, A., On Instrumental Effects in Spectral Line Synthesis Observations, Ph.D. Thesis, Univ. of Groningen, 1985, see section 10.

Bracewell, R. N., Two-dimensional Aerial Smoothing in Radio Astronomy, *Aust. J. Phys.*, **9**, 297–314, 1956a.

Bracewell, R. N., Strip Integration in Radio Astronomy, *Aust. J. Phys.*, **9**, 198–217, 1956b.

Bracewell, R. N., *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 2000 (earlier eds. 1965, 1978).

Bracewell, R. N. and A. C. Riddle, Inversion of Fan-Beam Scans in Radio Astronomy, *Astrophys. J.*, **150**, 427–434, 1967.

Bracewell, R. N. and J. A. Roberts, Aerial Smoothing in Radio Astronomy, *Aust. J. Phys.*,**7**, 615–640, 1954.

Bracewell, R. N. and S. J. Wernecke, Image Reconstruction over a Finite Field of View, *J. Opt. Soc. Am.* **65**, 1342–1346, 1975.

Bridle, A. H., Synthesis Observing Strategies, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser., **6**, 443–476, 1989.

Briggs, D. S., *High Fidelity Deconvolution of Moderately Resolved Sources*, Ph.D. thesis, New Mexico Institute of Mining and Technology, Socorro, NM, 1995.

Briggs, D. S., R. A. Sramek, and F. R. Schwab, Imaging, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron Soc. Pacific Conf. Ser., **180**, 127–149, 1999.

Brouw, W. N., *Data Processing for the Westerbork Synthesis Radio Telesscope*, Univ. Leiden, 1971.

Brouw, W. N., Aperture Synthesis, in *Methods in Computational Physics*, Vol. 14, B. Alder, S. Fernbach, and M. Rotenberg, Eds., Academic Press, New York, 1975, pp. 131–175.

Burton, W. B., The Structure of Our Galaxy Derived from Observations of Neutral Hydrogen, in *Galactic and Extragalactic Radio Astronomy*, G. L. Verschuur amd K. I. Kellermann, Eds., Springer-Verlag, Berlin, pp. 295–358, 1988.

Christiansen, W. N., and J. A. Warburton, The Distribution of Radio Brightness over the Solar Disk at a Wavelength of 21 cm. III. The Quiet Sun–Two-Dimensional Observations, *Aust. J. Phys.*, **8**, 474–486, 1955.

Cole, T. W., Analog Processing Methods for Synthesis Observations, in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., Reidel, Dordrecht 1979, pp. 123–141.

Cox, A. N., Ed., *Allen's Astrophysical Quantities*, 4th ed., AIP Press, Springer, New York, 2000.

Crane, P. C. and P. J. Napier, Sensitivity, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser., **6**, 139–165 1989.

D'Addario, L. R., Cross Correlators, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser. **6**, 59–82, 1989.

Downes, D., R. Neri, T. Wiklind, D. J. Wilner, and P. A. Shaver, Detection of CO(4–3), CO(9–8), and Dust Emission in the Broad Absorption Line Quasar APM 08279+5255 at a Redshift of 3.9, *Astrophys. J. Lett.*, **513**, L1–L4, 1999.

Ekers, R. D., Error Recognition, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 321–334, 1999.

Ekers, R. D. and J. H. van Gorkom, Spectral Line Imaging with Aperture Synthesis Radio Telescopes, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Univ. Press, Cambridge, UK, 1984, pp. 21–32.

Fomalont, E. B., The East-West Structure of Radio Sources at 1425 MHz, *Astrophys. J. Suppl.*, **15**, 203–274, 1968.

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

Gordon, M. A., Computer Programs for Radio Astronomy, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks. Ed., Academic Press, New York, 1976.

Gordon, M. A., J. W. M. Baars, and W. J. Cocke, Observations of Radio Lines from Unresolved Sources: Telescope Coupling, Doppler Effects, and Cosmological Corrections, *Astron. Astrophys.*, **264**, 337–344, 1992.

Gradshteyn, I. S. and I. M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, New York, 5th ed., 1994.

Gurwell, M., Lunar and Planetary Fluxes at 230 GHz: Models for the Haystack 15-m Baseline, SMA Technical Memo. 127, Smithsonian Astrophysical Observatory, 1998.

Jacquinot, P. and B. Roizen-Dossier, Apodisation, in *Progress in Optics*, Vol. 3, pp. 29–186, E. Wolf, Ed., North Holland, Amsterdam, 1964.

Jennison, R. C., A Phase Sensitive Interferometer Technique for the Measurement of the Fourier Transforms of Spatial Brightness Distributions of Small Angular Extent, *Mon. Not. R. Astron. Soc.*, **118**, 276–284, 1958.

Kerr, F. J. and D. Lynden-Bell, Review of Galactic Constants, *Mon. Not. Roy. Ast. Soc.*, **221**, 1023–1038, 1986.

Lineweaver, C. H., L. Tenorio, G. F. Smoot, P. Keegstra, A. J. Banday, and P. Lubin, The Dipole Observed in the COBE DMR 4 Year Data, *Atrophys. J.*, **470**, 38–42, 1996.

Maltby, P. and A. T. Moffet, Brightness Distribution in Discrete Radio Sources, III. The Structure of the Sources, *Astrophys. J. Suppl.*, **7**, 141–163, 1962.

Masson, C. R., Angular Expansion and Measurement with the VLA: The Distance to NGC 7027, *Astrophys. J.*, **302**, L27–L30, 1986.

Moffet, A. T., Brightness Distribution in Discrete Radio Sources, I. Observations with an East-West Interferometer, *Astrophys. J. Suppl.*, **7**, 93–123, 1962.

Mould, J. R. and 16 coauthors, The Hubble Space Telescope Key Project on the Extragalactic Distance Scale. XXVIII. Combining the Constraints on the Hubble Constant, *Astrophys. J.*, **529**, 786–794, 2000.

Napier, P. J. and P. C. Crane, Signal-to-Noise Ratios, in *Synthesis Mapping, Proc. NRAO Workshop No. 5* (Socorro, NM, June 21–25, 1982), A. R. Thompson and L. R. D'Addario, Eds., National Radio Astronomy Observatory, Green Bank, WV, 1982.

Pearson, T. J., Non-Imaging Data Analysis, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 335–355, 1999.

Peebles, P. J. E., *Principles of Physical Cosmology*, Princeton Univ. Press, Princeton, NJ, 1993.

Reid, M. J., Spectral-Line VLBI, in *Very Long baseline Interferometry and the VLBA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron. Soc. Pacific Conf. Ser., **82**, 209–225, 1995.

Reid, M. J., Spectral-Line VLBI, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 481–497, 1999.

Rhodes, D. R., On the Spheriodal Functions, *J. Res. Natl. Bur. Stand.* (U.S.) B, **74**, 187–209, 1970.

Roelfsema, P., Spectral Line Imaging I: Introduction, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser., **6**, 315–339, 1989.

Rybicki, G. B. and A. P. Lightman, *Radiative Processes in Astrophysics*, Wiley, New York, 1979 (reprinted 1985).

Schwab, F. R., Optimal Gridding of Visibility Data in Radio Interferometry, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Univ. Press, Cambridge, UK, 1984, pp. 333–346.

Slepian, D., Analytic Solution of Two Apodization Problems, *J. Opt. Soc. Am.*, **55**, 1110–1115, 1965.

Slepian, D. and H. O. Pollak, Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty. I, *Bell Syst. Tech. J.*, **40**, 43–63, 1961.

Smith, F. G., The Measurement of the Angular Diameter of Radio Stars, *Proc. Phys. Soc. B*, **65**, 971–980, 1952.

Standish, E. M. and X X Newhall, New Accuracy Levels for Solar System Ephemerides, in *Proc. IAU Symp. 172, Dynamics, Ephemerides, and Astrometry of Solar System Bodies*, Kluwer, Dordrecht, 1995, pp. 29–36.

Taylor, J. R., *An Introduction to Error Analysis*, University Science Books, Mill Valley, CA, 1982.

Thompson, A. R. and R. N. Bracewell, Interpolation and Fourier Transformation of Fringe Visibilities, *Astron. J.*, **79**, 11–24, 1974.

Twiss, R. Q., A. W. L. Carter, and A. G. Little, Brightness Distribution Over Some Strong Radio Sources at 1427 Mc/s, *Observatory*, **80**, 153–159, 1960.

van Gorkom, J. H. and R. D. Ekers, Spectral Line Imaging II: Calibration and Analysis, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser., **6**, 341–353, 1989.

White, R. L. and R. H. Becker, The Resolution of P Cygni's Stellar Wind, *Astrophys. J.*, **262**, 657–662, 1982.

White, M., J. E. Carlstrom, M. Dragovan, and W. L. Holzapfel, Interferometric Observations of Cosmic Microwave Background Anisotropies, *Astrophys. J.*, **514**, 12–24, 1999.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

11 Deconvolution, Adaptive Calibration, and Applications

This chapter is concerned with techniques of processing that are largely nonlinear. These further improve the image as formed by the procedures described in Chapter 10. There are two principal deficiencies in the visibility data that limit the accuracy of synthesis images. These are (1) the limited distribution of spatial frequencies in u and v and (2) errors in the measurements themselves. The limited spatial frequency coverage can be improved by deconvolution processes that allow the unmeasured visibility to take nonzero values within some general constraints on the image. Calibration can be improved by adaptive techniques in which the antenna gains, as well as the required image, are derived from the visibility data. Wide-field imaging, multifrequency imaging, and some other special applications are also described.

11.1 LIMITATION OF SPATIAL FREQUENCY COVERAGE

The intensity distribution $I_0(l, m)$ obtained in synthesis mapping can be regarded as the true intensity $I(l, m)$ convolved with the synthesized beam $b_0(l, m)$:

$$I_0(l, m) = I(l, m) ** b_0(l, m). \tag{11.1}$$

Knowing $I_0(l, m)$ and $b_0(l, m)$, can we solve for $I(l, m)$? An analytic procedure for deconvolving two functions is to take the Fourier transform of the convolution, which is equal to the product of the Fourier transforms of the components, divide out the Fourier transform of the known function, and transform back. From Eq. (11.1) we have

$$I(l, m) ** b_0(l, m) \rightleftharpoons \mathcal{V}(u, v)\,[W(u, v) w_u(u, v) w_t(u, v)], \tag{11.2}$$

where $\rightleftharpoons$ indicates Fourier transformation, $\mathcal{V}(u, v)$ is the visibility function, $W(u, v)$ is the spatial transfer function, $w_u(u, v)$ is the weighting required to obtain effective uniform density of data in the (u, v) plane, and $w_t(u, v)$ is an applied taper. However, the transfer function contains areas where it is zero, so we cannot divide it out to obtain $\mathcal{V}(u, v)$. The unmeasured visibilities present a fundamental problem, and any procedure that improves the derived intensity

other than weighting of the visibility must involve placing nonzero visibility values in the unmeasured (u, v) areas.

Bracewell and Roberts (1954) pointed out that there are an infinite number of solutions to the convolution in Eq. (11.1), since one can add any arbitrary visibility values in the unsampled areas of the (u, v) plane. The Fourier transform of these added values constitutes an invisible distribution that cannot be detected by any instrument with corresponding zero areas in the transfer function. It may be argued that in interpreting observations from any radio telescope, one should maintain only zeros in the unmeasured regions of spectral sensitivity, to avoid arbitrarily generating information. On the other hand, the zeros are themselves arbitrary values, some of which are certainly wrong. What is wanted is a procedure that allows the visibility at the unmeasured points to take values consistent with the most reasonable or likely intensity distribution, while minimizing the addition of arbitrary detail. Positivity of intensity and confinement of the angular structure of a source are expected characteristics that can be introduced into the imaging process. Negative intensity values and extensive sinusoidal structure are examples of instrumental artifacts to be removed. As suggested in the discussion of Fig. 2.6, procedures for the removal of the effects of sidelobes should be possible. A review of processing algorithms is given by Sault and Oosterloo (1996).

11.2 THE CLEAN DECONVOLUTION ALGORITHM

CLEAN Algorithm

One of the most successful deconvolution procedures is the algorithm CLEAN devised by Högbom (1974). This is basically a numerical deconvolving process applied in the (l, m) domain. The procedure is to break down the intensity distribution into point-source responses, and then replace each one with the corresponding response to a "clean" beam, that is, a beam free of sidelobes. The principal steps are as follows.

1. Compute the map and the response to a point source by Fourier transformation of the visibility and the weighted transfer function. These functions, the synthesized intensity and the synthesized beam, are often referred to as the "dirty map" and the "dirty beam," respectively. The spacing of the sample points in the (l, m) plane should not exceed about one-third of the synthesized beamwidth.
2. Find the highest intensity point on the map and subtract the response to a point source, including the full sidelobe pattern, centered on that position. The peak amplitude of the subtracted point source is equal to γ times the corresponding map amplitude. γ is called the loop gain, by analogy with negative feedback in electrical systems, and commonly has a value of a few tenths. Record the position and amplitude of the component removed by inserting a delta-function component into a model that will become the cleaned map.

3. Return to step 2 and repeat the procedure iteratively until all significant source structure has been removed from the map. There are several possible indicators of this condition. For example, one can compare the highest peak with the rms level of the residual intensity, look for the first time that the rms level fails to decrease when a subtraction is made, or note when significant numbers of negative components start to be removed.
4. Convolve the delta functions in the cleaned model with a clean-beam response, that is, replace each delta function with a clean-beam function of corresponding amplitude. The clean beam is often chosen to be a Gaussian with a half-amplitude width equal to that of the original synthesized (dirty) beam, or some similar function that is free from negative values.
5. Add the residuals (the residual intensity from step 3) into the clean-beam map, which is the output of the process.

It is assumed that each dirty-beam response subtracted represents the response to a point source. As discussed in Section 4.4, the visibility function of a point source is a pair of real and imaginary sinusoidal corrugations that extend to infinity in the (u, v) plane. Any intensity feature for which the visibility function is the same within the (u, v) area sampled by the transfer function would produce a response in the map identical to the point source response. Högbom (1974) has pointed out that much of the sky is a random distribution of point sources on an empty background, and CLEAN was initially developed for this situation. Nevertheless, experience shows that CLEAN also works on extended and complicated sources.

The result of the first three steps in the CLEAN procedure outlined above can be represented by a model intensity distribution that consists of a series of delta functions with magnitudes and positions representing the subtracted components. Since the modulus of the Fourier transform of each delta function extends uniformly to infinity in the (u, v) plane, the visibility is extrapolated as required beyond the cutoff of the transfer function.

The delta-function components do not constitute a satisfactory model for astronomical purposes. Groups of delta functions with separations no greater than the beamwidth may actually represent extended structure. Convolution of the delta-function model by the clean beam, which occurs in step 4, removes the danger of over-interpretation. Thus CLEAN performs, in effect, an interpolation in the (u, v) plane. Desirable characteristics of a clean beam are that it should be free from sidelobes, particularly negative ones, and that its Fourier transform should be constant inside the sampled region of the (u, v) plane and rapidly fall to a low level outside it. These characteristics are essentially incompatible since a sharp cutoff in the (u, v) plane results in oscillations in the (l, m) plane. The usual compromise is a Gaussian beam, which introduces a Gaussian taper in the (u, v) plane. Since this function tapers the measured data and the unmeasured data generated by CLEAN, the resulting intensity distribution no longer agrees with the measured visibility data. However, the absence of large, near-in sidelobes improves the dynamic range of the image, that is, it increases the range of intensity over which the structure of the image can reliably be measured.

As discussed above, we cannot directly divide out the weighted transfer function on the right-hand side of Eq. (11.2) because it is truncated to zero outside the areas of measurement. In CLEAN, this problem is solved by analyzing the measured visibility into sinusoidal visibility components and then removing the truncation so that they extend over the full (u, u) plane. Selecting the highest peak in the (l, m) plane is equivalent to selecting the largest complex sinusoid in the (u, v) plane.

At the point that the component subtraction is stopped, it is generally assumed that the residual intensity distribution consists mainly of the noise. Retaining the residual distribution within the map is, like the convolution with the clean beam, a nonideal procedure that is necessary to prevent misinterpretation of the final result. Without the residuals added in step 5, there would be an amplitude cutoff in the structure corresponding to the lowest subtracted component. Also, the presence of the background fluctuations provides an indication of the level of uncertainty in the intensity values. An example of the effect of processing with the CLEAN algorithm is shown in Fig. 11.1.

Implementation and Performance of the CLEAN Algorithm

As a procedure for removing sidelobe responses, CLEAN is easy to understand. Being highly nonlinear, however, CLEAN does not yield readily to a complete mathematical analysis. Some conclusions have been derived by Schwarz (1978, 1979), who has shown that conditions for convergence of CLEAN are that the synthesized beam must be symmetrical and its Fourier transform, that is, the weighted transfer function, must be non-negative. These conditions are fulfilled in the usual synthesis procedure. Schwartz's analysis also indicates that if the number of delta-function components in the CLEAN model does not exceed the number of independent visibility data, CLEAN converges to a solution that is the least-squares fit of the Fourier transforms of the delta-function components to the measured visibility. In enumerating the visibility data, either the real and imaginary parts or the conjugate values (but not both) are counted independently. In maps made using the FFT algorithm there are equal numbers of grid points in the (u, v) and (l, m) planes, but not all (u, v) grid points contain visibility measurements. To maintain the condition for convergence it is a common procedure to apply CLEAN only within a limited area, or "window," of the original map.

In order to clean a map of a given dimension, it is necessary to have a beam pattern of twice the map dimensions so that a point source can be subtracted from any location in the map. However, it is often convenient for the map and beam to be the same size. In this case only the central quarter of the map can be properly processed. Thus, it is commonly recommended that the map obtained from the initial Fourier transform should have twice the dimensions required for the final map. As mentioned above, the use of such a window also helps to ensure that the number of components removed does not exceed the number of visibility data and, in the absence of noise, allows the residuals within the window area to approach zero.

Several arbitrary choices influence the result of the CLEAN process. These include the parameter γ, the window area, and the criterion for termination. A

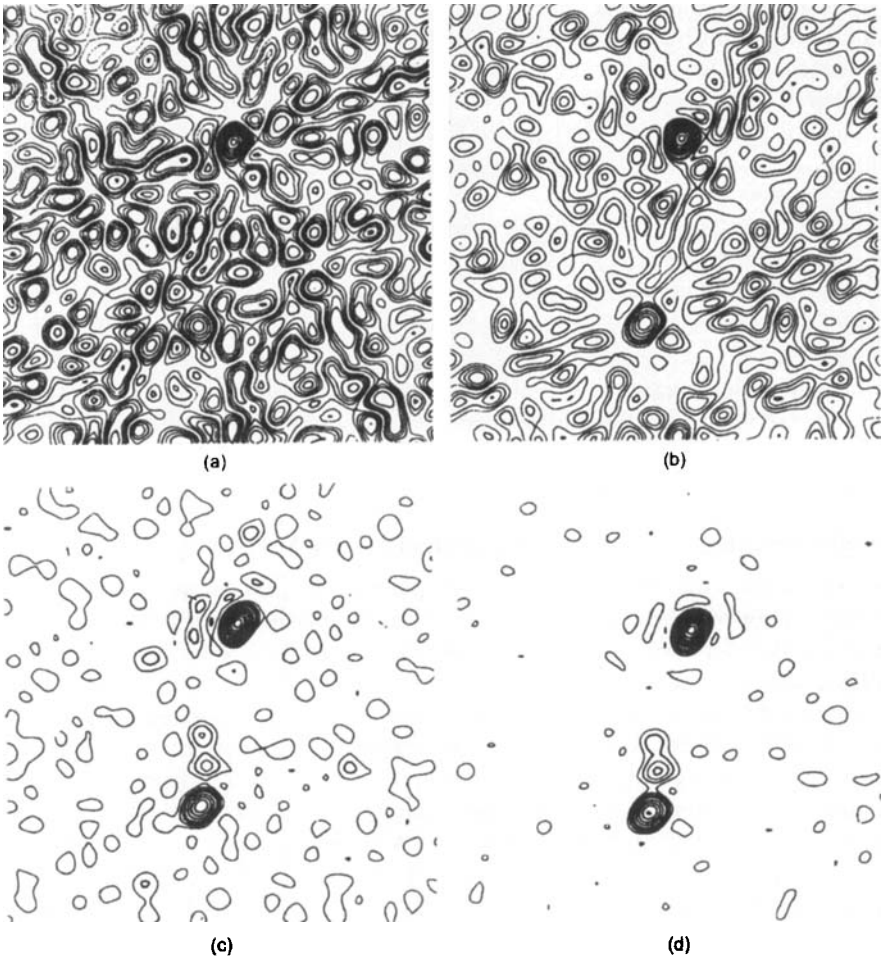


Figure 11.1 Illustration of the CLEAN procedure using observations of 3C224.1 at 2695 MHz made with the interferometer at Green Bank, and rather sparse (u, v) coverage. (a) The synthesized "dirty" map; (b) the map after one iteration with the loop gain $\gamma = 1$; (c) after two iterations; (d) after six iterations. The components removed were restored with a clean beam in all cases. The contour levels are 5, 10, 15, 20, 30, etc. percent of the maximum value. From Högbom (1974), courtesy of *Astron. Astrophys. Suppl.*

value between 0.1 and 0.5 is usually assigned to γ, and it is a matter of general experience that CLEAN responds better to extended structure if the loop gain is in the lower part of this range. The computation time for CLEAN increases rapidly as γ is decreased, because of the increasing number of subtraction cycles required. If the signal-to-noise ratio is $\mathcal{R}_{\rm sn}$, then the number of cycles required for one point source is $-\log \mathcal{R}_{\rm sn}/\log(1-\gamma)$. Thus, for example, with $\mathcal{R}_{\rm sn} = 100$ and $\gamma = 0.2$, a point source requires 21 cycles.

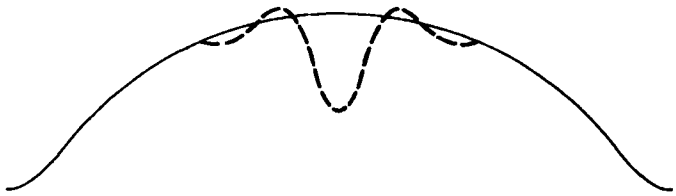


Figure 11.2 Subtraction of the point-source response (broken line) at the maximum of a broad feature, as in the process CLEAN. After Clark (1982).

A well-known problem of CLEAN is the generation of spurious structure in the form of spots or ridges as modulation on broad features. A heuristic explanation of this effect is given by Clark (1982). The algorithm locates the maximum in the broad feature and removes a point-source component, as shown in Fig. 11.2. The negative sidelobes of the beam add new maxima, which are selected in subsequent cycles, and thus there is a tendency for the component subtraction points to be located at intervals equal to the spacing of the first sidelobe of the synthesized (dirty) beam. The resulting map contains a lumpy artifact introduced by CLEAN, but the map is consistent with the measured visibility data. Cornwell (1983) has introduced a modification of the CLEAN algorithm that is intended to reduce this unwanted modulation. The original CLEAN algorithm minimizes

$$\sum_i w_i |\mathcal{V}_i^{\text{meas}} - \mathcal{V}_i^{\text{model}}|^2, \tag{11.3}$$

where $\mathcal{V}_i^{\text{meas}}$ is the measured visibility at (u_i, v_i), w_i is the applied weighting, and $\mathcal{V}_i^{\text{model}}$ is the corresponding visibility of the CLEAN-derived model. The summation is taken over the points with nonzero data in the input transformation for the dirty map. Cornwell's algorithm minimizes

$$\sum_i w_i |\mathcal{V}_i^{\text{meas}} - \mathcal{V}_i^{\text{model}}|^2 - \kappa s, \tag{11.4}$$

where s is a measure of smoothness and κ is an adjustable parameter. Cornwell finds that the mean-squared intensity of the model, taken with a negative sign, is an effective implementation of s.

The effects of visibility tapering appear in both the original map and the beam, and thus the magnitudes and positions of the components subtracted in the CLEAN process should be largely independent of the taper. However, since tapering reduces the resolution, it is a common practice to use uniform visibility weighting for maps that are processed using CLEAN. Alternatively, in difficult cases such as those involving extended, smooth structure, reduction of sidelobes by tapering may improve the performance of CLEAN.

An important reduction in the computation required for CLEAN was introduced by Clark (1980). This is based on subtraction of the point-source responses

in the (u, v) plane and using the FFT for moving data between the (u, v) and (l, m) domains. The procedure consists of minor and major cycles. A series of minor cycles is used to locate the components to be removed by performing approximate subtractions using only a small patch of the synthesized dirty beam that includes the main beam and the major sidelobes. Then in a major cycle the identified point-source responses are subtracted, without approximation, in the (u, v) plane. That is, the convolution of the delta functions with the dirty beam is performed by multiplying their Fourier transforms. The series of minor and major cycles is then repeated until the required stop condition is reached. Clark devised this technique for use with data from the VLA and found that it reduced the computation by a factor of two to ten compared with the original CLEAN algorithm.

Other variations on the CLEAN process have been devised; one of the more widely used is the Cotton–Schwab algorithm [Schwab (1984); see Sect. IV]. The subtractions in the major cycle are performed on the ungridded visibility data, which eliminates aliasing at this point. The algorithm is also designed to permit processing of adjacent fields, which are treated separately in the minor cycles but in the major cycles components are jointly removed from all fields.

To summarize the characteristics of CLEAN, we note that it is simple to understand from a qualitative viewpoint and straightforward to implement, and that its usefulness is well proven. On the other hand, a full analysis of its response is difficult. The response of CLEAN is not unique, and it can produce spurious artifacts. It is sometimes used in conjunction with model-fitting techniques; for example, a disk model can be removed from the image of a planet and the residual intensity processed by CLEAN. It is also used as part of more complex image construction techniques, which are described later in this chapter. For more details, including hints on usage, see Cornwell, Braun, and Briggs (1999).

11.3 MAXIMUM ENTROPY METHOD

MEM Algorithm

An important class of image restoration algorithms operates to produce a map that agrees with the measured visibility to within the noise level, while constraining the result to maximize some measure of image quality. Of these the maximum entropy method (MEM) has received particular attention in radio astronomy. If $I'(l, m)$ is the intensity distribution derived by the maximum entropy method, a function $F(I')$ is defined, which is referred to as the entropy of the distribution. $F(I')$ is determined entirely by the distribution of I' as a function of solid angle and takes no account of structural forms within the map. In constructing the map, $F(I')$ is maximized within the constraint that the Fourier transform of I' should fit the observed visibility values.

In astronomical image formation an early application of the maximum entropy method is that of Frieden (1972) to optical images. In radio astronomy the earliest discussions are by Ables (1974) and Ponsonby (1973). The aim of the technique, as described by Ables, is to obtain an intensity distribution consistent with all

relevant data but minimally committal with regard to missing data. Thus, $F(I')$ must be chosen so that maximization introduces legitimate a priori information but allows the visibility in the unmeasured areas to assume values that minimize the detail introduced.

Several forms of $F(I')$ have been used, which include the following:

$$F_1 = -\sum_i \frac{I'_i}{I'_s} \log\left(\frac{I'_i}{I'_s}\right) \tag{11.5a}$$

$$F_2 = -\sum_i \log I'_i \tag{11.5b}$$

$$F_3 = -\sum_i I'_i \ln\left(\frac{I'_i}{M_i}\right), \tag{11.5c}$$

where $I'_i = I'(l_i, m_i)$, $I'_s = \sum_i I'_i$, M_i represents an a priori model, and the sums are taken over all pixels in the map. F_3 can be described as relative entropy, since the intensity values are specified relative to a model.

A number of papers discuss the derivation of the expressions for entropy from theoretical and philosophical considerations. Bayesian statistics are invoked: see Jaynes (1968, 1982). Gull and Daniell (1979) consider the distributions of intensity quanta scattered randomly on the sky, and they derive the form F_1, which is also used by Frieden (1972). The entropy form F_2 is obtained by Ables (1974) and Wernecke and D'Addario (1977). Other investigators take a pragmatic approach to the maximum entropy method (Högbom 1979, Subrahmanya 1979, Nityananda and Narayan 1982). They view the method as an effective algorithm, even though there may be no underlying physical or information-theoretic basis for the choice of constraints. Högbom (1979) points out that both F_1 and F_2 contain the required mathematical characteristics: the first derivatives tend to infinity as I' approaches zero, so maximizing F_1 or F_2 produces positivity in the image. The second derivatives are everywhere negative, which favors uniformity in the intensity. Narayan and Nityananda (1984) consider a general class of functions F that have the properties $d^2F/dI'^2 < 0$ and $d^3F/dI'^3 > 0$. F_1 and F_2, discussed above, are members of this class.

In the maximization of the entropy expression $F(I')$, the constraint that the resulting intensity model should be consistent with the measured visibility data is implemented through a χ^2 statistic. χ^2 is a measure of the mean squared difference between the measured visibility values, $\mathcal{V}_k^{\text{meas}} = \mathcal{V}(u_k, v_k)$, and the corresponding values for the model $\mathcal{V}_k^{\text{model}}$:

$$\chi^2 = \sum_k \frac{|\mathcal{V}_k^{\text{meas}} - \mathcal{V}_k^{\text{model}}|^2}{\sigma_k^2}, \tag{11.6}$$

where σ_k^2 is the variance of the noise in $\mathcal{V}_k^{\text{meas}}$, and the summation is taken over the visibility data set. Obtaining a solution involves an iterative procedure; for descriptions, see Wernecke and D'Addario (1977), Wernecke (1977), Gull

and Daniell (1978), Skilling and Bryan (1984), and a review by Narayan and Nityananda (1984). As an example, Cornwell and Evans (1985) maximize a parameter J given by

$$J = F_3 - \alpha\chi^2 - \beta S_{\text{model}}, \tag{11.7}$$

where F_3 is defined in Eq. (11.5c). S_{model} is the total flux density of the model and is included because, in order for the process to converge to a satisfactory result, it was found necessary to include a constraint that the total flux density of the model be equal to the measured flux density. Lagrange multipliers α and β are included, the values of which are adjusted as the model-fitting proceeds so that χ^2 and S_{model} are equal to the expected values. Through the use of F_3, a priori information can be introduced into the final image. The various algorithms that have been developed for implementing MEM generally use the gradients of the entropy and of χ^2 to determine the adjustment of the model in each iteration cycle.

A feature of maps derived by the maximum entropy method is that the point-source response varies with position, so the angular resolution is not constant over the map. Comparison of maximum entropy maps with those obtained using direct Fourier transformation often shows higher angular resolution in the former. The extrapolation of the visibility values can provide some increase in resolution over more conventional mapping techniques.

Comparison of CLEAN and MEM

CLEAN is defined in terms of a procedure, so the implementation is straightforward, but because of the nonlinearity in the processing, a noise analysis of the result is very difficult. In contrast, MEM is defined in terms of an image that fits the data to within the noise and is also constrained to maximize some parameter of the image. The noise in MEM is taken into account through the χ^2 statistic, and the resulting effect on the noise is more easily analyzed for MEM; see, for example, Bryan and Skilling (1980). Some further points of comparison are as follows:

- Implementation of MEM requires an initial source model, which is not necessary in CLEAN.
- CLEAN is usually faster than MEM for small images, but MEM is faster for very large images. Cornwell, Braun, and Briggs (1999) give the break-even point as about 10^6 pixels in typical VLA images.
- CLEAN images tend to show a small-scale roughness, attributable to the basic approach of CLEAN, which models all images as ensembles of point sources. In MEM the constraint in the solution emphasizes smoothness in the image.
- Broad, smooth features are better deconvolved using MEM, since CLEAN may introduce stripes and other erroneous detail. MEM does not perform well on point sources, particularly if they are superimposed on a smooth background that prevents negative sidelobes from appearing as negative intensity in the dirty map.

To illustrate the characteristics of the CLEAN and MEM procedures, Fig. 11.3 shows examples of processing of a model jet structure from Cornwell (1995) and Cornwell, Braun, and Briggs (1999), using model calculations by Briggs. The jet model is based on similar structure in M87, and is virtually identical to the contour levels shown in part (e). The left-hand end of the jet is a point source smoothed to the resolution of the simulated observation. Visibility values for the model corresponding to the (u, v) coverage of the VLBA (Napier et al. 1994) were calculated for a frequency of 1.66 GHz and a declination of 50° with essentially full tracking range. Thermal noise was added, but the calibration was assumed to be fully accurate. Fourier transformation of the visibility data and the spatial transfer function provided the dirty image and dirty beam. The image showed the basic structure but fine details were swamped by sidelobes. Parts (a) to (c) of Fig. 11.3 show the effects of processing by CLEAN. In the CLEAN deconvolution 20,000 components were subtracted with a loop gain of 0.1. Part (a) shows the result of application of CLEAN to the whole image, and part (b) shows the result when components are taken only within a tight support region surrounding the source (the technique sometimes referred to as use of a box or window). Note the improvement obtained in (b), which is a result of adding the information that there is no emission outside the box region. The contours approximately indicate the intensity increasing in powers of two from a low value of 0.05%. Part (c) shows the same image as panel (b) but with contours starting a factor of 10 lower in intensity. The roughness visible in the low-level contours is characteristic of CLEAN, in which each component is treated independently and there is no mechanism to relate the result for any one component to those for its neighbors, unlike the case of MEM, where a smoothness constraint is introduced. Parts (d) to (f) result from MEM processing. Part (d) shows the result of MEM deconvolution using the same constraint region as in panel (b) and 80 iterations. The circular pattern of the background artifacts, centered on the point source, clearly shows that MEM does not handle such a feature well. In part (e) the point source was subtracted, using the CLEAN response to the feature, and then the MEM deconvolution performed with the same constraint region as in (d). The source was then replaced. Part (f) shows the same response as (e) with the lowest contours at the same level as panel (c). The low-level contours show the structure contributed by the observation and processing. The contours are smoother in the MEM image than in the CLEAN one. The images in (c) and (f) have comparable *fidelity*, that is, accuracy of reproduction of the initial model. Combinations of procedures, such as the use of CLEAN to remove point source responses from a map and then the use of MEM to process the broader background features, as illustrated above, can sometimes be used to advantage in complex images.

Other Deconvolution Procedures

Briggs (1995) has applied a non-negative, least-squares (NNLS) algorithm for deconvolution. The NNLS algorithm was developed by Lawson and Hanson (1974), and provides a solution to a matrix equation of the form $\mathbf{AX} = \mathbf{B}$, where, in the radio astronomy application, $\mathbf{A}$ represents the dirty beam and $\mathbf{B}$ the dirty map. The

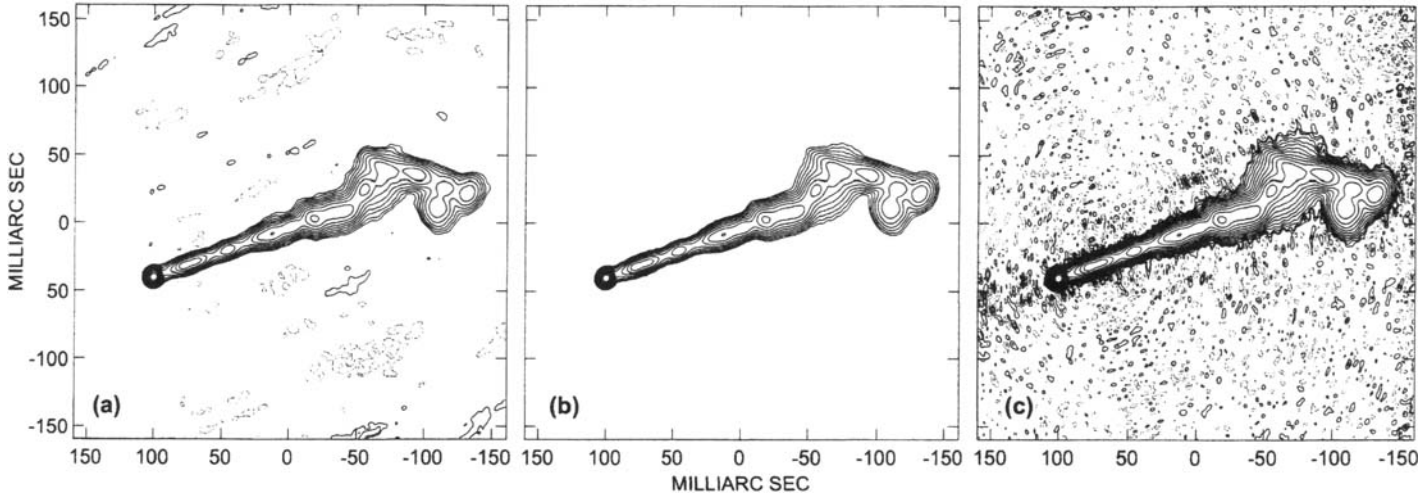


Figure 11.3 ***a, b, c*** Examples of deconvolution procedures applied to a model jet structure that includes a point source at the left-hand end. Part (a) shows the result of application of CLEAN to the whole image, and (b) the result when components are only taken within a tight support region surrounding the source. Note the improvement obtained in (b). The contours approximately indicate the intensity increasing in powers of two from a low value of 0.05%. Part (c) shows the same image as (b) but with contours starting a factor of 10 lower in intensity, and the roughness characteristic of CLEAN is visible in the low-level contours.

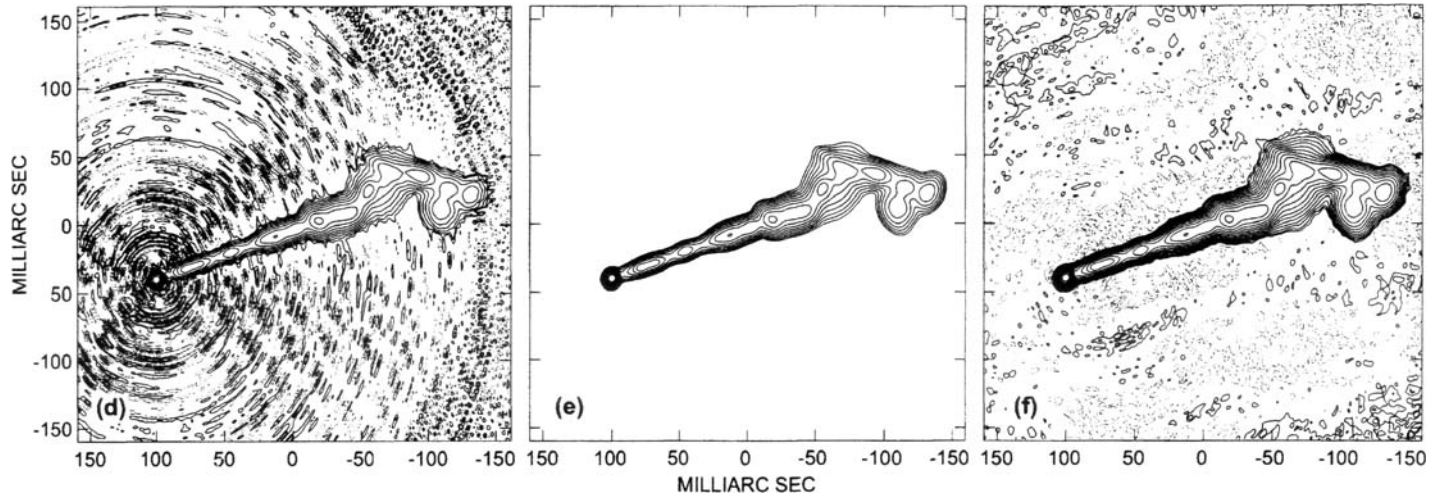


Figure 11.3 *d, e, f* Part (d) shows the result of MEM deconvolution using the same constraint region as in (b) and 80 iterations. The circular artifacts, centered on the point source, illustrate the well-known inability of MEM to handle sharp features well. In part (e) the point source was subtracted, using the CLEAN response to the feature, and then the MEM deconvolution performed with the same constraint region as in (d). The source was then replaced. Part (f) shows the same response as (e) with the lowest contours at the same level as part (c). Note that the low-level contours are smoother in the MEM image than in the CLEAN one. The images in (c) and (f) show comparable fidelity to the model. All six parts are from Cornwell (1995), courtesy of the Astron. Soc. Pacific Conf. Ser.

algorithm provides a least-mean-squares solution for the intensity **X** that is constrained to contain no negative values. However, unlike the case for MEM, there is no smoothness criterion involved. The NNLS solution requires more computer capacity than CLEAN or MEM solutions, but Briggs' investigation indicated that it is capable of superior performance, particularly in cases of compact objects of width only a few synthesized beamwidths. NNLS was found to reduce the residuals to a level close to the system noise in the observations. In certain cases it was found to work more effectively than CLEAN in hybrid mapping and self-calibration procedures (discussed below) and to allow higher dynamic range to be achieved. In MEM the residuals may not be entirely random but may be correlated in the image plane, and this effect can introduce bias in the (u, v) data that limits the achievable dynamic range. CLEAN appears to behave somewhat similarly unless it is allowed to run long enough to work down into the noise. See Briggs (1995) and Cornwell, Braun, and Briggs (1999) for further discussion.

11.4 ADAPTIVE CALIBRATION AND MAPPING WITH AMPLITUDE DATA ONLY

Calibration of the visibility amplitude is often accurate to a few percent, but phase errors expressed as a fraction of a radian may be much larger, as a result of variations in the ionosphere or troposphere. Nevertheless, the relative values of the uncalibrated visibility measured simultaneously on a number of baselines contain information about the intensity distribution that can be extracted through the closure relationships described in Chapter 10, Eqs. (10.30) and (10.32). Following Schwab (1980), we use the term *adaptive calibration* for both the hybrid mapping and self-calibration techniques that make use of this information. Mapping with amplitude data only has also been investigated and is briefly described.

Hybrid Mapping

The rekindling of interest in closure techniques in the 1970s began with their rediscovery by Rogers et al. (1974), who used closure phases to derive model parameters for VLBI data. Fort and Yee (1976) and several later groups incorporated closure data into iterative mapping techniques, of which we describe that by Readhead et al. (1980). The procedure is as follows:

1. Obtain an initial trial map based on inspection of visibility amplitudes and any a priori data such as a map at a different wavelength or epoch. If the trial map is inaccurate, the convergence will be slow, but if necessary, an arbitrary trial map such as a single point source will often suffice.
2. For each visibility integration period, determine a complete set of independent amplitude and/or phase closure equations. For each such set, compute a sufficient number of visibility values from the model such that when added to the closure relationships, the total number of independent equations is equal to the number of antenna spacings.

3. Solve for the complex visibility corresponding to each antenna spacing and make a map from the visibility data by Fourier transformation.
4. Process the map from step 3 using CLEAN, but omitting the residuals.
5. Apply constraints for positivity and confinement (delete components having negative intensity or lying outside the area that is judged to contain the source).
6. Test for convergence and return to step 2 as necessary, using the map from step 5 as the new model.

Note that the solution improves with iteration because of the constraints of confinement and positivity introduced in step 5. These nonlinear processes can be envisioned as spreading the errors in the model-derived visibility values throughout the visibility data, so that they are diluted when combined with the observed values in the next iterative cycle.

In the process described, and most variants of it, the map is formed by using some data from the model and some from direct measurements, and following Baldwin and Warner (1978) the name *hybrid mapping* is widely used as a generic description. With the use of phase closure, there is no absolute position measurement, but there is no ambiguity with respect to the position angle of the image. With the use of amplitude closure, only relative levels of intensity are determined, but it is usually not difficult to calibrate enough of the data to establish an intensity scale. In many cases the amplitude data are sufficiently accurate as observed, and only the phase closure relationships need be used; Readhead and Wilkinson (1978) have described a version of the above program using phase closure only. Other versions of this technique, which differ mainly in detail of implementation from that described, have been developed by Cotton (1979) and Rogers (1980). If there is some redundancy in the baselines, the number of free parameters is reduced, which can be advantageous, as discussed by Rogers.

The number of antennas n_a is obviously an important factor in mapping by the closure relationships since it affects the efficiency with which the data are used. We can quantify this efficiency by considering the number of closure data as a fraction of the number of data that would be available if full calibration were possible, as a function of n_a. The numbers of independent closure data are given by Eqs. (10.31) and (10.33). The number of data with full calibration is equal to the number of baselines, which, if we assume there is no reduncancy, is $\frac{1}{2}n_a(n_a - 1)$. For the phase data the fraction is

$$\frac{\frac{1}{2}(n_a - 1)(n_a - 2)}{\frac{1}{2}n_a(n_a - 1)} = \frac{n_a - 2}{n_a}. \tag{11.8}$$

For the amplitude data the fraction is

$$\frac{\frac{1}{2}n_a(n_a - 3)}{\frac{1}{2}n_a(n_a - 1)} = \frac{n_a - 3}{n_a - 1}. \tag{11.9}$$

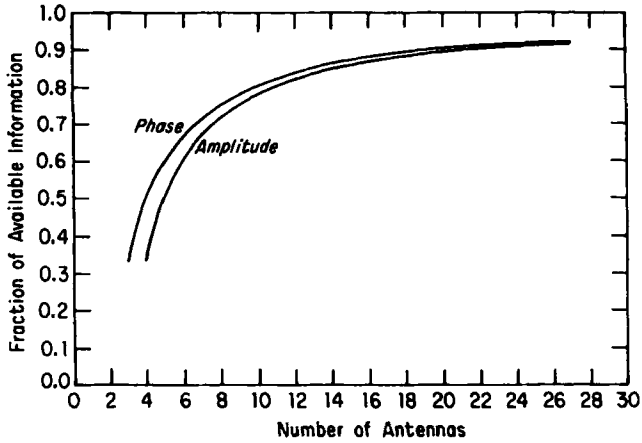


Figure 11.4 Visibility data that can be obtained through adaptive calibration techniques expressed as a fraction of those available from a fully calibrated array. The curves correspond to Eqs. (11.8) and (11.9).

These fractions are also equal to the ratios of observed data to observed plus model-derived data in each iteration of the hybrid mapping procedure. Equations (11.8) and (11.9) are plotted in Fig. 11.4. For $n_a = 4$, the closure relationships yield only 50% of the possible phase data and only 33% of the amplitude data. For $n_a = 10$, however, the corresponding figures are 80% and 78%. Thus, in any array in which the atmosphere or instrumental effects may limit the accuracy of calibration by a reference source, it is desirable that the number of antennas should be at least ten and preferably more. The number of iterations required to obtain a solution with the hybrid technique depends on the complexity of the source, the number of antennas, the accuracy of the initial model, and other factors including details of the algorithm used.

Self-Calibration

Another group of image construction programs that basically performs the same function as hybrid mapping, but with a different approach, is described as *self-calibration*. Here the complex antenna gains are regarded as free parameters to be explicitly derived together with the intensity. In certain cases the process is easily explained. For example, in mapping an extended source containing a compact component (as in many radio galaxies), the broad structure is resolved with the longer antenna spacings, leaving only the compact source. This can be used as a calibrator to provide the relative phases of the long-spacing antenna pairs, but not the absolute phase since the position is not known. Then, if there is a sufficient number of long spacings in the array, the relative gain factors of the antennas can be obtained using long spacings only. Such a special intensity distribution, however, is not essential to the method, and with an iterative technique it is possible

to use almost any source as its own calibrator. Programs of this type were developed by Schwab (1980) and by Cornwell and Wilkinson (1981). Reviews of the techniques are given by Pearson and Readhead (1984) and Cornwell (1989).

The procedure in self-calibration is to use a least-squares method to minimize the square of the modulus of the difference between the observed visibilities $\mathcal{V}_{mn}^{\text{meas}}$ and the corresponding values for the derived model, $\mathcal{V}_{mn}^{\text{model}}$. The expression that is minimized is

$$\sum_{\text{time}} \sum_{m<n} w_{mn} |\mathcal{V}_{mn}^{\text{meas}} - g_m g_n^* \mathcal{V}_{mn}^{\text{model}}|^2, \tag{11.10}$$

where the weighting coefficient w_{mn} is usually chosen to be inversely proportional to the variance of $\mathcal{V}_{mn}^{\text{meas}}$, and the quantities shown are all functions of time within the observing period. Expression (11.10) can be written

$$\sum_{\text{time}} \sum_{m<n} w_{mn} |\mathcal{V}_{mn}^{\text{model}}|^2 |X_{mn} - g_m g_n^*|^2, \tag{11.11}$$

where

$$X_{mn} = \frac{\mathcal{V}_{mn}^{\text{meas}}}{\mathcal{V}_{mn}^{\text{model}}}. \tag{11.12}$$

If the model is accurate, the ratio X_{mn} of the uncalibrated observed visibility to the visibility predicted by the model is independent of u and v but proportional to the antenna gains. Thus the values of X_{mn} simulate the response to a calibrator and enable the gains to be determined. However, since the initial model is only approximate, the desired result must be approached by iteration.

The self-calibration procedure is as follows:

1. Make an initial map as for hybrid mapping.
2. Compute the X_{mn} factors for each visibility integration period within the observation.
3. Determine the antenna gain factors for each integration period.
4. Use the gains to calibrate the observed visibility values and make a map.
5. Use CLEAN and select components to provide positivity and confinement of the image; Cornwell (1982) recommends omitting all features for which $|I(l, m)|$ is less than that for the most negative feature.
6. Test for convergence and return to step 2 as necessary.

The numbers of independent data used in the procedure above are, as in the case of hybrid mapping, equal to the numbers of independent closure relationships given in Eqs. (10.31) and (10.33), that is, $\frac{1}{2}n_a(n_a - 3)$ for amplitude and $\frac{1}{2}(n_a - 1)(n_a - 2)$ for phase. The two procedures, hybrid mapping and self-calibration, are basically equivalent but differ in details of approach and imple-

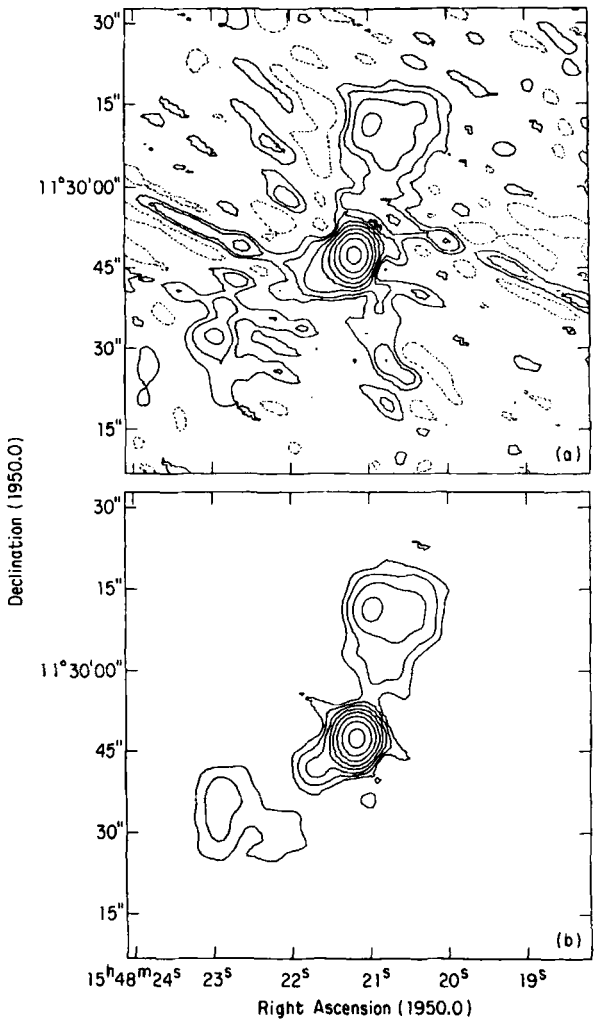


Figure 11.5 Effect of self-calibration on a VLA radio image of the quasar 1548+115. (a) Image obtained by normal calibration techniques, which has spurious detail at the level of 1% of the peak intensity. (b) Image obtained by the self-calibration technique, in which the level of spurious detail is reduced below the 0.2% level. In both (a) and (b) the lowest contour level is 0.6%. From Napier, Thompson, and Ekers (1983); ©1983 IEEE.

mentation. The efficiency as a function of the number of antennas (Fig. 11.4) applies to both. Examples of the performance of the self-calibration technique are shown in Figs. 11.5 and 11.6.

Treating the gain factors, which are the fundamental unknown quantities, as free parameters as in self-calibration is a rather more direct approach than that of hybrid mapping. A global estimate of the instrumental factors is obtained using the entire data set. Cornwell (1982) points out that it is easier to deal correctly

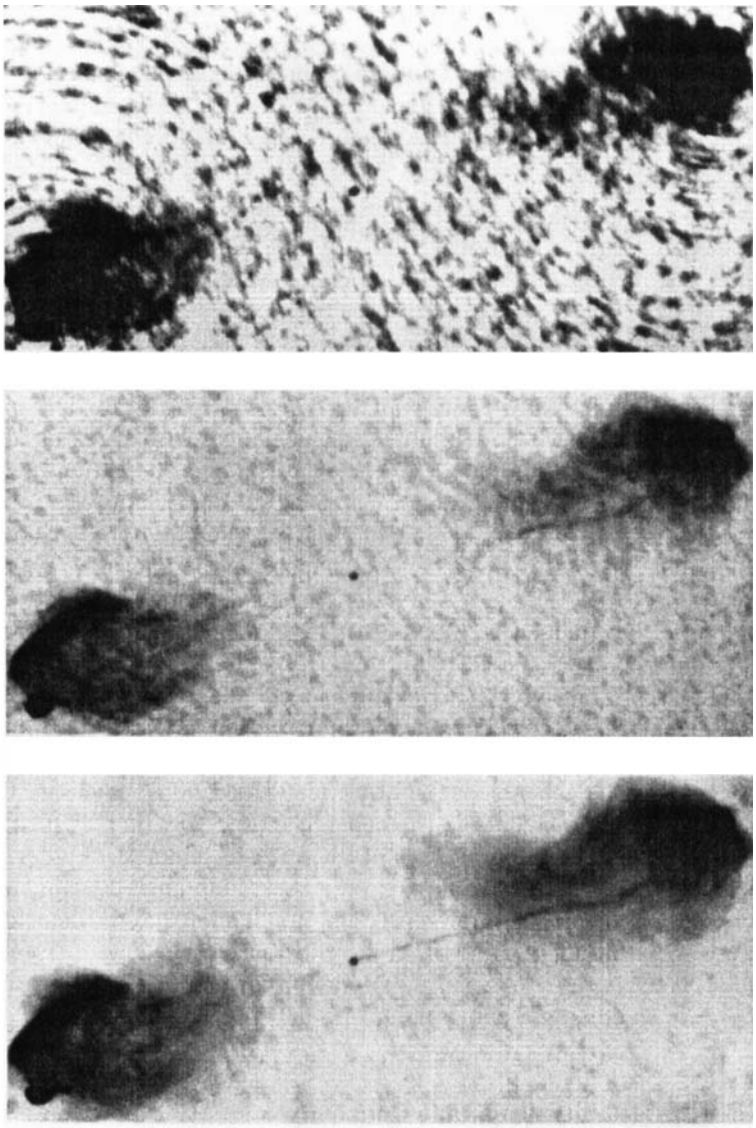


Figure 11.6 Three stages in the reduction of the observation of Cygnus A shown in Fig. 1.18 (Perley, Dreher, and Cowan 1984). The top image is the result of transformation of the calibrated visibility data using the FFT algorithm. The calibration source was approximately 3° from Cygnus A. The center image shows reduction using the maximum entropy algorithm. This compensates principally for the undersampling in the spatial frequencies and thereby removes sidelobes from the synthesized beam. The result is similar to that obtainable using the CLEAN algorithm. The bottom image shows the effect of the self-calibration technique, in which the maximum entropy image is used as the initial model. The final step improves the dynamic range by a factor of 3. In observations where the initial calibration is not as good as in this case, self-calibration usually provides a greater improvement. The long dimension of the field is 2.1 arcmin and contains approximately 1000 pixels. Reproduced by permission of NRAO/AUI.

with the noise when considering complex visibility as a vector quantity, as in self-calibration, than when considering amplitude and phase separately, as in hybrid mapping. The noise combines additively in the vector components resulting in a Gaussian distribution, whereas in the amplitude and phase the more complicated Rice distributions of Eqs. (6.63) result. Cornwell and Wilkinson (1981) have developed a form of adaptive calibration that takes account of the different probability distributions of the amplitude and phase fluctuations, including system noise, for the different antennas. It has been used with the MERLIN array, which incorporates antennas of different sizes and designs (Thomasson 1986). The probability distributions of the antenna-associated errors are legitimate a priori information, which can be empirically determined for an array.

Experience shows that adaptive calibration techniques in many cases converge to a satisfactory result using only a single point source as a starting model, although inaccuracy in the initial model increases the number of iterative cycles required. A point source is a good model for the phase of a symmetrical intensity distribution, but may be a poor model for the amplitude. It must also be remembered that the accuracy of the closure relationships depends on the accuracy of the matching of the frequency responses and polarization parameters from one antenna to another, as discussed in Sections 7.3 and 7.4. In general, any effect that cannot be represented by a single gain factor for each antenna, for example, anomalous behavior of a correlator, degrades the closure accuracy.

In using adaptive calibration techniques, the integration period of the data must not be longer than the coherence time of the phase variations; otherwise the visibility amplitude may be reduced. The coherence time may be governed by the atmosphere, for which the timescale is of the order of minutes. In order for the mapping procedure to work, the field under observation must contain structure fine enough to provide a phase reference, and bright enough to be detected with satisfactory signal-to-noise ratio within the coherence time. Thus adaptive calibration does not solve all problems, and cannot be used for the detection of a very weak source in an otherwise empty field.

Mapping with Visibility Amplitude Data Only

A number of studies have been made concerning the feasibility of producing images using only the amplitude values of the visibility. The Fourier transform of the squared modulus of the visibility is equal to the autocorrelation of the intensity distribution, $I \star\star I$:

$$|\mathcal{V}(u,v)|^2 = \mathcal{V}(u,v)\mathcal{V}^*(u,v) \rightleftharpoons I(l,m) \star\star I(l,m). \tag{11.13}$$

The right-hand side can also be written as a convolution: $I(l,m) * * I(-l,-m)$. The problem of mapping with $|\mathcal{V}|$ only is mainly one of interpreting a map of the autocorrelation of I. Without phase data the position of the center of the field cannot be determined, and there is a 180° rotational ambiguity in the position angle of the map. However, these restrictions are often acceptable.

Examples of studies relevant to mapping without phase data are found in Bates (1969, 1984), Napier (1972), and Fienup (1978). Napier and Bates (1974) review some of the results. The positivity requirement is generally found to be insufficient to provide unique solutions for one-dimensional profiles, but for two-dimensional maps uniqueness is obtained in some cases (Bruck and Sodin 1979). Baldwin and Warner (1978, 1979) considered the case of a two-dimensional distribution of point sources, with some success in deducing a source map from the autocorrelation function. Although these approaches showed promise of providing an advance in the interpretation of radio interferometer data, they have not proved to be of great importance in radio astronomy. No simple, reliable method of interpretation was realized, and, more importantly, the development of techniques that make use of closure relationships has allowed visibility phases to contribute useful data even when not well calibrated.

11.5 MAPPING WITH HIGH DYNAMIC RANGE

The *dynamic range* of an image is usually defined as the ratio of the maximum intensity to the rms level at some part of the field where the background is mainly blank sky. This rms level is assumed to indicate the lowest measurable intensity. The term *image fidelity* is used to indicate the degree to which an image is an accurate representation of a source on the sky. Image fidelity is not directly measurable on an actual source, but simulation of an observation of a model source and reduction of the visibility data allow comparison of the resulting image with the model. This is a way of investigating antenna configurations, processing methods, and other details. The requirements and techniques are discussed in detail by Perley (1989, 1999a).

High dynamic range requires high accuracy in calibration, removal of any erroneous data, and careful deconvolution. That is, it requires high accuracy in the visibility measurements, and very good (u, v) coverage. A phase error $\Delta\phi$ can be regarded as introducing an erroneous component of relative amplitude $\sin \Delta\phi$ into the visibility data, in phase quadrature to the true visibility. An amplitude error of $\varepsilon_a\%$ can be regarded as introducing an error component of relative amplitude $\varepsilon_a\%$ into the visibility. Thus, for example, a phase error of 10° introduces as large an error component as does an amplitude error of 17%. An amplitude error of 17% would be considered unusually large in most cases, except in conditions of strong atmospheric attenuation. However, a 10° phase error would be much more commonly encountered, especially at frequencies where ionospheric or tropospheric irregularities are important. A phase error $\Delta\phi$ (rad) in a correlator output introduces an error component of rms relative amplitude $\Delta\phi/\sqrt{2}$ in the resulting map. With similar errors in $n_a(n_a - 1)/2$ baselines, the dynamic range of a snapshot is limited to $\sim n_a/\Delta\phi$.

Use of self-calibration is an essential step in minimizing gain errors. However, after calibration of the antenna-based gain factors, there remain small baseline-based terms which can also be calibrated. These result from variations, from one antenna to another, in the frequency bandpass or the polarization, as discussed

in Sections 7.3 and 7.4, and similar effects. Note that in arrays with very high sensitivity at the longer wavelengths, the requirement to observe down to the limit set by system noise, in the presence of background sources, places a lower limit on the required dynamic range. A large number of array elements is beneficial in such cases (Lonsdale et al. 2000).

Obtaining the highest possible dynamic range requires attention to details that are specific to particular instruments. For the VLA, the following figures are quoted as a rough guideline for a good observation. Basic calibration results in dynamic range of order 1000 : 1. After self-calibration, dynamic range up to ~20,000 : 1 is possible. After careful correction of baseline-based errors, it may be as high as ~80,000 : 1. If the spectral correlator is used, which avoids errors in the quadrature networks of the continuum correlator and also relaxes the requirement for delay accuracy, ~200,000 : 1 is achievable, with much care, assuming that the signal-to-noise ratio is adequate.

11.6 MOSAICKING

Mosaicking is a technique that allows mapping of an area of sky that is larger than the beam of the array elements. It becomes very important in the millimeter-wavelength range, where antenna beams are relatively narrow. Although radio astronomy antennas for millimeter wavelengths are generally smaller in diameter than are antennas for centimeter wavelengths, their beamwidths are often narrower because the wavelengths are so much shorter.

Consider mapping a square field whose sides are n times the width of the antenna primary beam. One can divide the required area into n^2 subfields, each the size of a beam, and map each such area separately. The n^2 beam-area maps can then be fitted together like mosaic pieces to cover the full field desired. One would anticipate that some difficulty might occur in obtaining uniform sensitivity, particularly near the joins of the mosaic pieces, but clearly the idea is feasible. From the sampling theorem described in Section 5.2, the number of visibility sample points in u and v required in a map covering n^2 beam areas is n^2 times as many as would be required in a map that covers just one beam area. In mosaicking, the increased data are obtained by using n^2 different pointing directions of the antennas. As a result, the sampling of the visibility in u and v must be at an interval $1/n$ of that for a field equal to the beam size, and this interval is usually less than the diameter of the antenna aperture. However, it is possible to determine how the visibility varies on a scale less than the diameter of an antenna, as we now discuss.

Figure 5.9 of Chapter 5 shows two antennas that are tracking the position of a source. The antenna spacing projected normal to the direction of the source is u, and the antenna diameter is d_λ, both quantities being measured in wavelengths. In the u direction the interferometer responds to spatial frequencies from $(u - d_\lambda)$ to $(u + d_\lambda)$, since spacings within this range can be found within the antenna apertures. Measurement of the variation of the visibility over this range of baselines can provide the fine sampling required in mosaicking. The difference in path lengths from the source to the two antenna apertures is w wavelengths, and as

the antennas track, the variation in w gives rise to fringes at the correlator output. Since the apertures of the antennas remain normal to the direction of the source, the path difference w, and its rate of change, are the same for any pair of points of which one is in each aperture plane, regardless of their spacing. Thus, because of the tracking motion, the signals received at any two such points produce a component of the correlator output with the same fringe frequency. Such components cannot, therefore, be separated by Fourier analysis, and information on the variation of the visibility within the spatial frequency range $(u - d_\lambda)$ to $(u + d_\lambda)$ is lost. However, in mosaicking, the antenna beams must be scanned across the field, either by moving periodically between different pointing centers or by continuously scanning, for example, in a raster pattern. The scanning is in addition to the usual tracking motion to follow the source across the sky. In Fig. 5.9 it can be seen that if the antennas are suddenly turned through a small angle $\Delta\theta$, then the position of the point B is changed by $\Delta u\ \Delta\theta$ wavelengths in a direction parallel to that of the source. This results in a phase change of approximately $2\pi\,\Delta u\ \Delta\theta$ in the fringe component corresponding to the spacing $(u + \Delta u)$, of which points A_1 and B are an example. Since this phase change is linearly proportional to Δu, the variation of the visibility within the range $(u - d_\lambda)$ to $(u + d_\lambda)$ can be obtained by Fourier transformation of the correlator output with respect to the pointing offset $\Delta\theta$. Thus the changes in pointing induce variations in the fringe phase that are dependent on the spacing of the incoming rays within the antenna apertures, and this effect allows the information on the variation of the visibility to be retained.

The conclusion given above, that the scanning action of the antennas allows information on a range of visibility values to be retrieved, was first reached by Ekers and Rots (1979), using a mathematical analysis, as follows. Consider a pair of antennas with spacing (u_0, v_0) pointing in the direction (l_p, m_p). As the pointing angle is varied, the effective intensity distribution over the field of interest is represented by $I(l, m)$ convolved with the normalized antenna beam $A_N(l, m)$. The observed fringe visibility is the Fourier transform with respect to u and v of $I(l, m)$ multiplied by the antenna response for the particular pointing:

$$\mathcal{V}(u_0, v_0, l_p, m_p) = \iint A_N(l - l_p, m - m_p) I(l, m) e^{-j2\pi(u_0 l + v_0 m)}\, dl\, dm. \tag{11.14}$$

Assuming that the antenna beam is symmetrical, we can write Eq. (11.14) as

$$\mathcal{V}(u_0, v_0, l_p, m_p) = \iint A_N(l_p - l, m_p - m) I(l, m) e^{-j2\pi(u_0 l + v_0 m)}\, dl\, dm, \tag{11.15}$$

which has the form of a two-dimensional convolution:

$$\mathcal{V}(u_0, v_0, l_p, m_p) = [I(l, m) e^{-j2\pi(u_0 l + v_0 m)}] ** A_N(l, m). \tag{11.16}$$

Now we take the Fourier transform of $\mathcal{V}$ with respect to u and v, which represents the full-field visibility data obtained by means of the ensemble of pointing angles

used:

$$\begin{aligned}\mathcal{V}(u, v) &= \int\int [I(l, m)e^{-j2\pi(u_0 l + v_0 m)}] ** A_N(l, m)e^{j2\pi(ul+vm)} dl\, dm \\ &= [\mathcal{V}(u, v) ** {}^2\delta(u_0 - u, v_0 - v)]\overline{A}_N(u, v).\end{aligned} \tag{11.17}$$

Here $\overline{A}(u, v)$ is the Fourier transform of $A_N(l, m)$, that is, the autocorrelation of the field distribution over the aperture of a single antenna, referred to as the transfer function or spatial sensitivity function of the antenna. The two-dimensional delta function ${}^2\delta(u_0 - u, v_0 - v)$ is the Fourier transform of $e^{-j2\pi(u_0 l + v_0 m)}$. As the final step, Eq. (11.17) becomes

$$\mathcal{V}(u, v) = \mathcal{V}[(u_0 - u), (v_0 - v)]\overline{A}_N(u, v). \tag{11.18}$$

The conclusion from Eq. (11.18) is that if one observes a field of dimensions equal to several beamwidths, obtains the visibility for a number of pointing directions, and then for each antenna pair takes the Fourier transform of the visibility with respect to the pointing direction, the result will be values of the visibility extended over an area of the (u, v) plane as large as the support of the function $\overline{A}_N(u, v)$. For a circular reflector antenna of diameter d, $\overline{A}_N(u, v)$ is nonzero within a circle of diameter $2d$. Thus, if $\overline{A}_N(u, v)$ is known with sufficient accuracy, that is, the beam pattern is sufficiently well calibrated, the visibility can be obtained at the intermediate points required to provide the full-field map.

In the practical reduction of visibility data used in mosaicking, the Fourier transform with respect to pointing is usually not explicitly performed. The importance of the discussion above is that it shows that the information at the required spacing is present in the data if the antenna pointing is scanned with respect to the source, either as a continuous motion or as a series of discrete pointings. The reduction to obtain the intensity distribution is generally based on the use of nonlinear deconvolution algorithms.

Cornwell (1988) has pointed out that the angular spacing required between the pointing centers on the sky can be deduced from the sampling theorem of Fourier transforms (Section 5.2). A more general form of the theorem can be stated as follows: if a function $f(x)$ is nonzero only within an interval of width Δ in the x coordinate, then it is fully specified if its Fourier transform $F(s)$ is sampled at intervals no greater than Δ^{-1} in s. If the sampling is coarser than this, aliasing will occur and the original function will not be reproducible from the samples. Here we consider an antenna beam pointing toward a source that is wide enough to cover most of the reception pattern, that is, the main beam and major sidelobes. As we move the antenna beam to different pointing angles to cover the source, we are effectively sampling the convolution of the source and the antenna beam. The beam pattern is equal to the Fourier transform of the autocorrelation function of the field distribution over the antenna aperture. The field cuts off at the edges of the aperture, which is d_λ wavelengths wide. Thus the autocorrelation function

cuts off at a width $2d_\lambda$. Recall that in the earlier usage of the sampling function (Section 5.2) it was the source width that had the sharp cutoff. In the present case the theorem indicates that the interval between pointings Δl_p should not exceed $1/(2d_\lambda)$ in order to fully sample the source convolved with the beam. In practice, the antenna illumination function is likely to be tapered at the edge, so the autocorrelation function falls to low levels before it reaches the cutoff width $2d_\lambda$. Thus if Δl_p slightly exceeds $1/(2d_\lambda)$, the error introduced may not be large.

Methods of Producing the Mosaic Map

The basic steps in the mosaicking method are as follows:

1. Observe the visibility function for an appropriate series of pointing centers.
2. Reduce the data for each pointing center independently to produce a series of maps, each covering approximately one antenna-beam area.
3. Combine the beam-area maps into the required full-field map.

In step 2 it is desirable also to deconvolve the synthesized beam response from each beam-area map to remove the effects of sidelobes in the response, and this can be done using, for example, CLEAN or MEM. Use of these nonlinear algorithms can fill in some of the frequency components of the intensity that were omitted from the coverage of the antenna array. Cornwell (1988) and Cornwell, Holdaway, and Uson (1993) describe two procedures for mosaic mapping. The first of these, which they refer to as linear mosaicking, is essentially the three steps above with a least-squares procedure for combination of the individual pointing maps in step 3. Although a nonlinear deconvolution is used individually on each beam-area map, the combination of the maps is linear process. The second procedure, which differs in that the deconvolution is performed jointly, is referred to as nonlinear mosaicking and involves a nonlinear algorithm such as MEM. Unmeasured visibility data can best be estimated in the deconvolution process if the full field that is covered by the ensemble of pointing angles contributes simultaneously to the deconvolution, rather than by treating each primary beam area separately. The benefit of a joint deconvolution of the combined beam-area maps is illustrated by consideration of an unresolved component of the intensity distribution located at the edge of a beam area where it occurs in two or more individual beam maps. Being at the beam edge where the response is changing rapidly, the amplitude of the component is more likely to be inaccurately determined, but such errors will tend to average out in the combined data. In the application to mosaicking, maximum entropy can be envisaged as the formation of a map that is consistent with all the visibility data for the various pointings, within the uncertainty resulting from the noise.

Cornwell (1988) discusses use of the maximum entropy algorithm of Cornwell and Evans (1985) in mosaicking. This algorithm is briefly described in Section 11.3 [see Eq. (11.7)]. The procedure is essentially the same as in the application to a single-pointing map, except for a few more steps in determining χ^2 and its gradient. As in Eq. (11.6), χ^2 is the statistic that indicates the deviation of the

model from the measured visibility values and is here expressed as

$$\chi^2 = \sum_p \sum_i \frac{|\mathcal{V}_{ip}^{\text{meas}} - \mathcal{V}_{ip}^{\text{model}}|^2}{\sigma_{vip}^2}, \tag{11.19}$$

where the subscripts i and p indicate the ith visibility value at the pth pointing position, and σ_{vip}^2 is the variance of the visibility. An initial model is required, and the procedure follows a series of steps described by Cornwell (1988), as follows:

1. For the first pointing center, multiply the current trial model with the antenna beam as pointed during the observation, and take the Fourier transform with respect to (l, m) to obtain the predicted visibility values.
2. Subtract the measured visibilities from the model visibilities to obtain a set of residual visibilities. Insert the residual visibilities into the accumulating χ^2 function of Eq. (11.19).
3. By Fourier transformation, convert the residual visibilities, weighted inversely as their variances, into an intensity distribution. Taper this distribution by multiplying it by the antenna beam pattern, and store in a data array of dimensions equal to the full MEM model.
4. Repeat steps 1–3 for each pointing. In step 2, add the value for χ^2 to those for the other pointings in this cycle. In step 3, add the residual intensity values into the data array. The accumulated values in this data array are used to obtain the gradient of χ^2 with respect to the MEM image.

The reason for the additional multiplication of the residual distribution by the beam function in step 3 is that it reduces unwanted responses from sidelobes of the primary beam that fall on adjacent pointing areas. It also weights the data with respect to the signal-to-noise ratio. Completion of the MEM procedure may require several tens of cycles through the steps given above to obtain convergence to the final image. To complete the process, smoothing with a two-dimensional Gaussian beam of width equal to the array resolution is recommended, to reduce the effects of variable resolution across the map.

A slightly different procedure for nonlinear mosaicking is described by Sault, Stavely-Smith, and Brouw (1996). In this case the beam-area maps are combined linearly without the individual deconvolution step, and then the final nonlinear deconvolution is applied to the combined map. In the linear combination each pixel in the combined map is a weighted sum of the corresponding pixels in the individual beam-area maps. As an example, Sault et al. show results for a mosaic of the Small Magellanic Cloud made with the compact configuration of the Australia Telescope using 320 pointings. They demonstrate that the joint deconvolution used in nonlinear mosaicking is superior to the linear combination of the subfield maps, even if these have been individually deconvolved. They also show the deconvolution using both their method and that described by Cornwell (1988), and conclude that the results are of comparable quality.

Some Requirements of Arrays for Mosaicking

In mapping sources wider than the antenna beam it is important to obtain visibility values at increments in u and v that are smaller than the diameter of an antenna. Data equivalent to an essentially continuous coverage in u and v can then be obtained by observing at various pointing positions as discussed above. The minimum spacing of two antennas is limited by mechanical considerations, and there is a gap or region of low sensitivity corresponding to a spacing of about half the minimum spacing between the centers of two antenna apertures. This minimum spacing depends on the antenna design, but in general, unless the range of zenith angles is restricted, two antennas of diameter d cannot be spaced much closer than about $1.4d$, or perhaps $1.25d$ with special design. Otherwise, there is danger of mechanical collision, especially if there is a possibility that the antennas may not always be pointing in the same direction. Total-power observations with a single antenna will, in principle, provide spacings from zero to d/λ, but with some antennas measurements at spatial frequencies greater than $\sim 0.5\mathrm{d}/\lambda$ are unreliable because the spatial sensitivity function of the antenna falls to low levels as a result of the tapered illumination of the reflector. Missing data at low (u, v) values result in broad negative sidelobes of the synthesized beam, such that the beam appears to be situated in a shallow bowl-shaped depression. This effect is most noticeable when the field to be mapped is wide enough that there are several empty (u, v) cells within the central area.

The transfer function $\overline{A}_N(u)$ is the autocorrelation function of the field distribution over the antenna aperture, and depends on the particular design of the antenna, including the illumination pattern of the feed. The solid curve in Fig. 11.7 shows $\overline{A}$ for a uniformly illuminated circular aperture, which can be regarded as an ideal case. Since there is usually some tapering in the illumination of a reflector antenna, $\overline{A}_N$ will generally fall off somewhat more rapidly than the curve shown. The function $\overline{A}_N$ in Fig. 11.7 is proportional to the common area of two overlapping circles of diameter d, and the abscissa is the distance between their centers. In three dimensions this function is sometimes referred to as chat ( ), and its properties are discussed by Bracewell (1995). The dashed curves in Fig. 11.7 show the relative spatial sensitivity for an interferometer using two uniformly illuminated, circular apertures of diameter d. Curve 1 is for a spacing of $1.4d$ between the centers of the apertures; curve 2 for a spacing of $1.25d$. If both total-power and interferometer data are obtained, it can be seen that the minimum sensitivity occurs for spacings of approximately half the antenna spacing.

One solution to increasing the minimum sensitivity in the spatial frequency coverage is the addition of total-power measurements from a larger antenna, of diameter $\sim 2d$ or $\sim 3d$; see, for example, Bajaja and van Albada (1979) or Welch and Thornton (1985). Total-power observations with a diameter-$2d$ antenna will provide spacings from zero to about d and thus cover the gap. However, since the cost of an antenna scales approximately as $d^{2.7}$, the expected cost of an antenna of diameter $2d$ is roughly 6.5 times that of an antenna of diameter d. Furthermore, the large antenna may not achieve the accuracy of surface or pointing of the smaller antennas, so it may have a more restricted range of operating frequency.

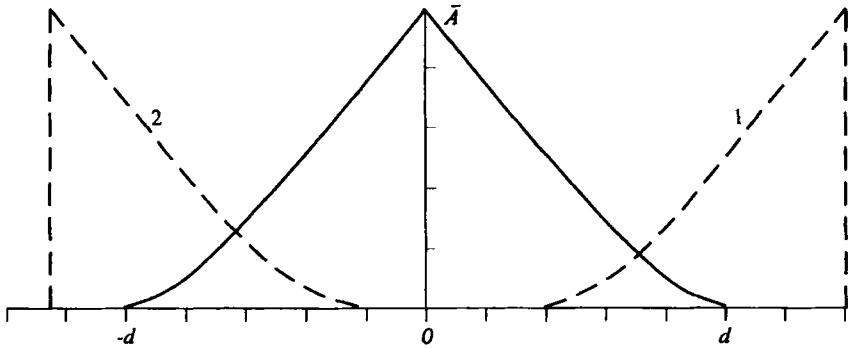


Figure 11.7 The solid curve centered on the origin shows the spatial sensitivity function $\bar{A}$ for a single antenna of diameter d. The curve corresponds to the case of uniform excitation over the aperture. This curve indicates the relative sensitivity to spatial frequencies for total-power observations with a single antenna. The dashed curves show the spatial sensitivity for two antennas of diameter d, with uniform aperture excitation, working as an interferometer. Curve 1 is for a spacing of $1.4d$ between the centers of the antennas, and curve 2 is for a spacing of $1.25d$. If the aperture illumination is tapered the curves will fall off to low values more rapidly than is shown.

Another possibility for covering the missing spatial frequencies is the use of one or more pairs of smaller antennas, say, $d/2$ in diameter, with spacing about $0.7d$. A pair of antennas of diameter $d/2$ have $1/4$ the area, and consequently $1/4$ the sensitivity to fine structure, of a pair of the standard antennas. Since the beam of the smaller antenna has four times the solid angle of a standard antenna, it will require $1/4$ the number of pointing directions, and the integration time for each one can be four times as long. Cornwell, Holdaway, and Uson (1993) present evidence that for mosaicking it is possible to obtain satisfactory performance with a homogeneous array, that is, one in which all antennas are the same size. This requires total-power observation as well as interferometry with some antennas spaced as closely as possible. The deconvolution steps in the data reduction help to fill in remaining (u, v) gaps.

At frequencies of several hundred gigahertz, where antenna beams are of minute-of-arc order, maps of objects of order one degree in size require numbers of pointings in the range 10^2–10^4. Any given pointing cannot be quickly repeated, so dependence on earth rotation to fill in small gaps in the (u, v) coverage may not be practicable. Thus arrays designed for mosaicking of large objects require good instantaneous (u, v) coverage. At such high frequencies it is also desirable to avoid high zenith angles to minimize atmospheric effects.

An alternative to tracking discrete pointing centers is to sweep the beams over the area of sky under investigation in a raster scan motion. This technique has been referred to as "on-the-fly" mosaicking. It has several advantages, as follows:

- The uniformity of the (u, v) coverage for all points in the field is maximized, which results in uniformity of the synthesized beam across the resulting map and thereby simplifies the image processing.
- Each point in the field is observed many times in as rapid succession as possible, so some advantage can be taken of earth rotation to fill in the (u, v) coverage.
- If total-power measurements are made, the scanning motion of the beam can be used to remove atmospheric effects in a similar way to the use of beam switching in large single-dish telescopes.
- Waste of observing time during moves of the antennas from one pointing center to another is eliminated.

The disadvantage of on-the-fly observing is that the real-time integration at the correlator output must be somewhat less than the time taken for the beam to scan over any point in the field, and thus a large number of visibility data, each with a separate pointing position, are generated.

11.7 MULTIFREQUENCY SYNTHESIS

Making observations at several different radio frequencies is an effective way of improving the sampling of the visibility in the (u, v) plane. This technique is referred to as multifrequency synthesis, or bandwidth synthesis. Generally the range of frequencies is about $\pm 15\%$ of the mid-range value. Such a range can be very effective in filling in gaps in the coverage, and since it is not too large, major changes in the source structure with frequency are avoided [see, e.g., Conway, Cornwell, and Wilkinson (1990)]. However, the variation of structure with frequency may be large enough to limit the dynamic range unless some steps are taken to mitigate it, as discussed here. The principal cosmic radio emission mechanisms produce radio spectra that vary smoothly in frequency, and the intensity usually follows a power-law variation with frequency:

$$I(\nu) = I(\nu_0)\left(\frac{\nu}{\nu_0}\right)^{\alpha}, \tag{11.20}$$

where α is the spectral index, which varies with (l, m). If the spectrum does not conform to a power law, then, in effect, we can write

$$\alpha = \frac{\nu}{I}\frac{\partial I}{\partial \nu}. \tag{11.21}$$

If the spectral index were a constant over the source, the spectral effects could be removed. Although this is not the case, the spectral effects are reduced by first correcting the data for a "mean" or "representative" spectral index for the overall structure to be imaged. Thus, from this point, α will represent the spectral index of the deviation of the intensity distribution from this first-order correction.

Consider the case where the intensity variation can be approximated by a linear term:

$$\begin{aligned} I(\nu) &= I(\nu_0) + \frac{\partial I}{\partial \nu}(\nu - \nu_0) = I(\nu_0) + \alpha I(\nu_0)\frac{(\nu - \nu_0)}{\nu} \\ &\simeq I(\nu_0) + \alpha I(\nu_0)\frac{(\nu - \nu_0)}{\nu_0}, \end{aligned} \tag{11.22}$$

where the reference frequency ν_0 is near the center of the range of frequencies used. Equation (11.22) is the sum of a single-frequency term and a spectral term. To determine the synthesized beam of an array working in the multifrequency mode, consider the response to a point source with a spectrum given by Eq. (11.22). The response to the single-frequency term can be obtained by taking the Fourier transform of the spatial transfer function. The transfer function has a delta function of u and v for each visibility measurement. Each frequency used contributes a different set of delta functions. The response to the spectral term is obtained by multiplying the transfer function by $(\nu - \nu_0)/\nu_0$ and taking the Fourier transform. If we call the single-frequency and spectral responses b_0' and b_1', respectively, the synthesized beam is equal to

$$b_0(l, m) = b_0'(l, m) + \alpha(l, m)b_1'(l, m). \tag{11.23}$$

The first component is a conventional synthesized beam, and the second one is an unwanted artifact. The measured intensity distribution obtained as the Fourier transform of the measured visibilities is

$$I_0(l, m) = I(l, m) ** b_0'(l, m) + \alpha(l, m)I(l, m) ** b_1'(l, m), \tag{11.24}$$

where $I(l, m)$ is the true intensity on the sky. Conway, Cornwell, and Wilkinson (1990) and Sault and Wieringa (1994) have both developed deconvolution processes based on the CLEAN algorithm that deconvolve both b_0' and b_1'. In the method used by the first of these groups, components representing each one of the two beams were removed alternately. In the method used by the second group, each component removed represented both beams. These methods provide the distribution of both the source intensity and the spectral index as functions of frequency. Conway et al. also consider a logarithmic rather than a linear form of the frequency offsets from ν_0. These analyses show that for a frequency spread of approximately $\pm 15\%$, the magnitude of the response resulting from the b_1' component is typically 1% and can sometimes be ignored. Removing the b_1' component reduces the spectral effects to $\sim 0.1\%$.

11.8 NON-COPLANAR BASELINES

In Section 3.1 it was shown that, except in the case of east–west linear arrays, the baselines of a synthesis array do not remain in a plane as the earth rotates. It was also shown that for fields of view of small angular size [as given approximately

by Eq. (3.12)], the Fourier transform relationship between visibility and intensity can be expressed satisfactorily in two dimensions. This is the basis of a large part of all synthesis mapping. However, particularly for frequencies less than a few hundred megahertz, the small-field assumption does not always apply. At meter wavelengths the primary beams of the antennas are wide, for example, ~6° for a 25-m diameter antenna at a wavelength of 2 m. Also, the high density of strong sources on the sky at meter wavelengths requires that the full beam be mapped to avoid confusion. We now consider the case where the condition in Eq. (3.12) is not valid, so the two-dimensional solution should not be used. The following treatment follows those of Sramek and Schwab (1989), Cornwell and Perley (1992), and Perley (1999b). We start with the exact result in Eq. (3.7), which is

$$\mathcal{V}(u, v, w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{A_N(l, m) I(l, m)}{\sqrt{1 - l^2 - m^2}} \times \exp\left\{-j2\pi\left[ul + vm + w\left(\sqrt{1 - l^2 - m^2} - 1\right)\right]\right\} dl\, dm. \tag{11.25}$$

Here $\mathcal{V}(u, v, w)$ is the visibility as a function of spatial frequency in three dimensions, $A_N(l, m)$ is the normalized primary beam pattern of an antenna, and $I(l, m)$ is the two-dimensional intensity distribution to be mapped.

The next step is to rewrite Eq. (11.25) in the form of a three-dimensional Fourier transform, which involves the third direction cosine n defined with respect to the w axis. The phase of the visibility $\mathcal{V}(u, v, w)$ is measured relative to the visibility of a (hypothetical) source at the phase reference position for the observation. This introduces a factor $e^{j2\pi w}$ within the exponential term on the right-hand side of Eq. (11.25), as noted in the text following Eq. (3.7). The corresponding phase shift is inserted by the fringe rotation discussed in Section 6.1 under *Delay Tracking and Fringe Rotation*. As a result of this factor, we use $n' = n - 1$ as the conjugate variable of w in order to obtain the three-dimensional Fourier transform. Functions of n' will be indicated by a prime. Thus we rewrite Eq. (11.25) as follows:

$$\mathcal{V}(u, v, w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{A_N(l, m) I(l, m)}{\sqrt{1 - l^2 - m^2}} \delta\left(\sqrt{1 - l^2 - m^2} - n' - 1\right) \times \exp\left\{-j2\pi(ul + vm + wn')\right\} dl\, dm\, dn'. \tag{11.26}$$

The delta function $\delta(\sqrt{1 - l^2 - m^2} - n' - 1)$ is introduced to maintain the condition $n = \sqrt{1 - l^2 - m^2}$, and thereby to allow n' to be treated as an independent variable in the Fourier transformation. In a practical observation $\mathcal{V}$ is measured only at points at which the sampling function $W(u, v, w)$ is nonzero. The Fourier transform of the sampled visibility defines a three-dimensional intensity function I_3' as follows:

$$I_3'(l, m, n') = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W(u, v, w)\mathcal{V}(u, v, w)e^{j2\pi(ul+vm+wn')}\,du\,dv\,dw. \tag{11.27}$$

This is the Fourier transform of the product of the two functions $W(u, v, w)$ and $\mathcal{V}(u, v, w)$, which by the convolution theorem is equal to the convolution of the Fourier transforms of the two functions. Thus

$$I_3'(l, m, n') = \left\{\frac{A_N(l, m)I(l, m)\,\delta\left(\sqrt{1 - l^2 - m^2} - n' - 1\right)}{\sqrt{1 - l^2 - m^2}}\right\} * * * \overline{W}'(l, m, n'). \tag{11.28}$$

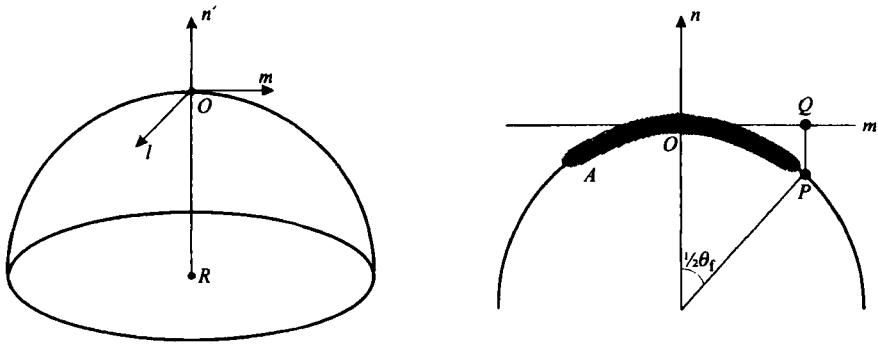
Here $\overline{W}'(l, m, n')$ is the Fourier transform of the three-dimensional sampling function $W(u, v, w)$, and the triple asterisk denotes three-dimensional convolution. Having determined the result of the Fourier transformation, we can now replace n' by $(n - 1)$, and Eq. (11.28) becomes

$$I_3(l, m, n) = \left\{\frac{A_N(l, m)I(l, m)\,\delta\left(\sqrt{1 - l^2 - m^2} - n\right)}{\sqrt{1 - l^2 - m^2}}\right\} * * * \overline{W}(l, m, n). \tag{11.29}$$

The expression in the braces on the right-hand side of Eq. (11.29) is confined to the surface of the unit sphere $n = \sqrt{1 - l^2 - m^2}$, since the delta function is nonzero only on the sphere. The function $\overline{W}$ with which it is convolved is the Fourier transform of the sampling function and is, in effect, a three-dimensional dirty beam. The convolution has the effect of spreading the expression so that I_3 has finite extent in the radial direction of the sphere. Figure 11.8a shows the unit sphere centered on the origin of (l, m, n) coordinates at R. The (l, m) plane in which the results of the conventional two-dimensional analysis lie is tangent to the unit sphere at O, at which point $n = 1$ and $n' = 0$. Note that since l, m, and n are direction cosines, the unit sphere in (l, m, n) is a mathematical concept, not a sphere in real space.

Several ways of obtaining an undistorted wide-field map are possible (Cornwell and Perley 1992), and are discussed as follows.

1. *Three Dimensional Transformation.* $I_3(l, m, n)$ can be deconvolved by means of a three-dimensional extention of the CLEAN algorithm. This is complicated by the fact that the visibility is, in practice, not as well sampled in w as it is in u and v; from Fig. 3.4 the large values of w occur for large zenith angles of the target source. In Fig. 11.8b the width of the angular field is θ_f. The transform must be computed over the range of (l, m) within this field, and over the range PQ in n. Cornwell and Perley suggest using a direct (rather than



(a) (b)

Figure 11.8 (a) One hemisphere of the unit sphere in (l, m, n) coordinates. The point R is the origin of the (l, m, n) coordinates. O is the origin of the (l, m, n') coordinates, which is the phase reference point. (b) Section through the unit sphere in the (m, n) plane. The shaded area represents the extent of the function I_3. A source at point A would not appear, or would be greatly attenuated, in a two-dimensional analysis in the (l, m) plane. The width of the three-dimensional "beam" in the n direction should be comparable to that in l and m, since the range of the sampling function in w is comparable to that in u and v if the observations cover a large range in hour angle. (In the superficially similar case in Fig. 3.5, the intensity function is not confined to the surface of the sphere because the measurements are all made in the $w' = 0$ plane.)

discrete) Fourier transform in the n to w transformation, since otherwise the poor sampling may result in serious sidelobes and aliasing. Thus, two-dimensional FFTs are performed in a series of planes normal to the n axis. The number of planes required is equal to PQ divided by the required sampling interval in the n direction. The range of measured visibility values has a width $2|w|_{\max}$ in the w direction, so, by the sampling theorem, the intensity function is fully specified in the n coordinate if it is sampled at intervals of $(2|w|_{\max})^{-1}$. The distance PQ is approximately equal to $\frac{1}{8}\theta_f^2 \approx \frac{1}{2}|l^2 + m^2|_{\max}$ [note that the angle $POQ = \theta_f/4$, and $(\theta_f/2)^2 = |l^2+m^2|_{\max}$]. Thus the number of planes in which the two-dimensional intensity must be calculated is $|l^2 + m^2|_{\max}|w|_{\max}$. [This result can also be obtained by taking the phase term in Eq. (3.8) that is omitted in going from three to two dimensions and sampling at the Nyquist interval of half a turn of phase.] The maximum possible value of w is $D_{\max}/\lambda$, where $D_{\max}$ is the longest baseline in the array. If θ_f is limited by the beamwidth of antennas of diameter d, for which the angular distance from the beam center to the first null is $\sim\lambda/d$, the required number of planes is $\sim(\lambda/d)^2 \times D_{\max}/\lambda = \lambda D_{\max}/d^2$. Examples of maps made using this method are given by Cornwell and Perley (1992).

2. *Polyhedron Mapping.* The area of the unit sphere for which the map is required can be divided into a number of subfields, which can be mapped individually using the small-field approximation. Each one is mapped in two dimensions

onto a plane that is tangent to the unit sphere at a different point on the sphere. These tangent points are the phase centers for the individual subfields. For each subfield map it is necessary to adjust both the visibility phases and the (u, v, w) coordinates of the whole database to the particular phase center. The subfields can be combined using methods similar to those used in mosaicking, including joint deconvolution. This approach has been referred to as *polyhedron mapping* because the various map planes form part of the surface of a polyhedron. Again examples are given by Cornwell and Perley (1992).

3. *Combination of Snapshots.* In most synthesis arrays the antennas are mounted on an area of approximately level ground and thus lie close to a plane at any given instant. In such cases a long observation can be divided into a series of "snapshots," for each of which the planar baseline condition applies individually. It should therefore be possible to make a map by combining a series of snapshot responses. Each snapshot represents the true intensity distribution convolved with a different dirty beam, since the (u, v) coverage changes progressively as the source moves across the sky. Ideally, deconvolution would thus require optimization of the intensity distribution using the snapshot responses in a combined manner rather than individually. It should be noted that the plane in which the baselines lie for any snapshot is, in general, not normal to the direction of the target source. As a result, the angle at which points on the unit sphere in Fig. 11.8a are projected onto the (l, m) plane is not parallel to the n axis, and varies with the position of the source on the sky. Positions of sources in the snapshot maps suffer an offset in (l, m) that is zero at the phase center but increases with distance from the phase center. Maps should be corrected for this effect before being combined. Since the required correction varies with the hour angle of the source, in long observations the effect can cause smearing of source images in the outer part of the map. Perley (1999b) discusses this effect and its correction. Bracewell (1984) has discussed a method similar to the combination of snapshots described above.

4. *Deconvolution with Variable Point-Source Response.* In cases where the effect of two-dimensional Fourier transformation is principally the distortion of the point-source response in the outer parts of the field, without serious attenuation of the response, then a possible procedure is deconvolution using a point source response (dirty beam) that is varied over the field to match the calculated response (McClean 1984). This approach was used by Waldram and McGilchrist (1990) in analysis of a survey using the Cambridge Low-Frequency Synthesis Telescope, which operates at 151 MHz using earth rotation and baselines that are offset from east–west by 3°. Point source responses were computed for a grid of positions within the field and the response for any particular position could then be obtained by interpolation. The principal requirement was to obtain accurate positions and flux densities for sources identified in maps obtained by two-dimensional transformation. Fitting the appropriate theoretical beam response for each source position allowed distortion of the beam, including any position offset, to be accounted for. The procedure is relatively inexpensive in computer time.

11.9 FURTHER SPECIAL CASES OF IMAGE ANALYSIS

Use of CLEAN and Self-Calibration with Spectral Line Data

A procedure that has been found to provide accurate separation of the continuum from the line features involves use of the deconvolving algorithm CLEAN (van Gorkom and Ekers 1989). However, if CLEAN is applied individually to the maps for the different channels, errors in the CLEAN process appear as differences from channel to channel and may be confused with true spectral features. Such errors can be avoided by subtracting the continuum before applying CLEAN to the line data. First, CLEAN is applied to an average of the continuum-only channels, and the visibility components removed from these channels are also removed from the visibility data for the channels containing line features. When the CLEAN process is terminated, the residuals are also removed from the line data. The resulting line channel maps, which should then contain only line data, can be deconvolved individually. Note that since absorption of the continuum may occur in the line frequency channels, maps of line-minus-continuum may contain negative as well as positive intensity features. Thus algorithms such as maximum entropy that depend on positivity of the intensity may not be easily applicable in such cases.

In applying self-calibration to eliminate phase errors in spectral line data, it can generally be assumed that phase and amplitude differences between channels vary only very little with time, and are removed by the bandpass calibration. This is true for both atmospheric and instrumental effects. Thus the strongest spectral feature in the field under investigation can be used to determine the phase-calibration solution, which is then applied to all channels. This feature might be the continuum emission represented by the average of the line-free channels, or a single channel with a strong maser line.

Low-Frequency Mapping

Synthesis mapping at wavelengths longer than about two meters, that is, at frequencies below about 150 MHz, usually involves wide fields of view because of the widths of the primary antenna beams. Synchrotron emission from radio sources generally becomes stronger as the frequency is reduced, and hence the density of strong sources on the sky increases with decreasing frequency. At low frequencies it is therefore often important to map the whole antenna beam to avoid source confusion resulting from aliasing. Also, the gain of the main beam of a refelector antenna decreases with decreasing frequency, and if phased arrays of dipoles are used, they have to be very large to maintain high gain. As a result, sources in the sidelobes are not so effectively suppressed relative to a source in the main beam as at higher frequencies. In the data analysis, unwanted responses from strong sources with known positions can be subtracted, but in practice the number of sources that can be removed in this way may be limited by the computing involved.

A complication of the wide-field mapping that is required is the problem of non-coplanar baselines, considered in Section 11.8. Another problem is the varia-

tion of ionospheric effects over the beam (Baldwin 1990). The excess path length in the ionosphere is proportional to ν^{-2} [see Section 13.3, Eq. (13.138)], so the resulting phase change is proportional to ν^{-1}. The term *isoplanatic patch* is used to denote an area of the sky over which the variation in the path length for an incoming wave is small compared with the observing wavelength. At centimeter and shorter wavelengths, the beams of reflector antennas used in synthesis arrays are generally smaller than the isoplanatic patch. Thus the effect of an irregularity in the ionosphere (or troposphere) is constant over the beam and can be corrected by a single phase adjustment for each antenna, for example, by self-calibration. However, at meter wavelengths the size of the antenna beam may be several times that of the ionospheric isoplanatic patch. For example, in observations with the VLA in New Mexico, Erickson (1999) estimates that the size of the isoplanatic patch at 74 MHz is $\sim 3°$–$4°$, whereas the beamwidth of a 25-m antenna at the same frequency is $\sim 13°$.

Kassim et al. (1993) describe simultaneous measurements of a number of strong sources at 74 and 330 MHz, using a phase reference procedure to calibrate the phases at the lower frequency. At 74 MHz the phase fluctuations are dominated by the ionosphere, and rates of phase change were found to be as high as one degree per second. These precluded calibration by the usual methods. However, at 330 MHz the rates of phase change were slow enough to allow mapping of strong sources. The resulting 330-MHz phases were scaled to 74 MHz and used to remove the ionospheric component from the 74-MHz data that were recorded simultaneously. The procedure for obtaining maps at 74 MHz was essentially as follows:

1. Simultaneous observations of a strong source were made at 74 and 330 MHz, with periodic observations of a calibrator at 330 MHz.
2. A map of the target source was made at 330 MHz using the standard techniques (i.e., use of a calibrator as at centimeter wavelengths). This was used as a starting model for self-calibration of the 330-MHz data.
3. The self-calibration provided phase calibration for each antenna at 330 MHz. These values were then scaled to 74 MHz, the ionospheric phase changes being inversely proportional to frequency, and used to remove the ionospheric variations from the 74 MHz data.
4. The instrumental phases at 330 and 74 MHz are different at each antenna as a result of dissimilar cable lengths, etc. To calibrate these differences an unresolved calibrator was observed at both 330 and 74 MHz. The ionospheric variations could be removed from the 74-MHz calibrator phases using the phase referencing scheme in step 3. The instrumental phase differences were thereby determined.
5. The 74-MHz map of the target source was made from the calibrated phase data. Self-calibration of the 74-MHz data was used to remove residual phase drifts, and for this the 330-MHz image provided a suitable starting model.

For the strongest sources, for which it was possible to obtain a good signal-to-noise ratio in an averaging time of no more than 10 s, self-calibration at 74 MHz

was sufficient in most cases. Although only eight VLA antennas were equipped for operation at 74 MHz, images with dynamic range of better than 20 dB were obtained for several sources. The problem of non-coplanar baselines did not arise in these measurements because the sources were compact enough for satisfactory two-dimensional mapping. The sources were also strong enough that other sources in the antenna beam and sidelobes could be ignored. For observing weaker sources, where several sources of similar flux density are within the antenna beam, a more complex correction procedure to handle more than one isoplanatic patch would be necessary.

Lensclean

A number of cases are known in which the image of a quasar or radio galaxy is distorted by the gravitational field of a galaxy, following the discovery of this phenomenon by Walsh, Carswell, and Weymann (1979). The line of sight from the lens source intersects, or passes very close to, the galaxy. In some cases the gravitational lensing results in multiple images of a single point-source quasar, and in other cases extended structure is involved: see, for example, Narayan and Wallington (1992). In studies of gravitational lensing, the structure of the gravitational field is of major astrophysical importance. The term *lensclean* has been used to denote a method of analysis, including several variations of the original algorithm, that allow the lensing field to be determined by synthesis imaging. The basis of these methods is analogous to self-calibration, in which the image is sufficiently overdetermined by the visibility measurements that it is possible to determine also the complex gains of the antennas. In lensclean it is the pattern of the gravitational field that is to be determined. An additional constraint is that points in the source of the radiation can each contribute to more than one point in the synthesized image.

The original lensclean procedure (Kochanek and Narayan, 1992) is based on an adaptation of the CLEAN algorithm. The basic principle can be described as follows. Consider the case where the source that is imaged by the lens contains extended structure. An initial model for the lens is chosen. Each point in the source contributes to multiple points in the image, and this mapping from the source to the image is defined by the lens model. For any point in the source, the intensity in the image should ideally be the same at each point at which it appears, since the imaging involves only geometric bending of the radiation from the source, as in an optical system. Suppose that the jth source pixel is mapped into n_j image pixels. In practice, the intensity of these pixels in the image is not equal because of defects in the lens model and noise in the image. The best estimate of the intensity of the pixel in the source is the mean intensity of the corresponding pixels in the image. Thus one can subtract components from the image in the manner of CLEAN and build up a map of the source. For each source pixel for which $n_j > 1$, the mean squared deviation of the intensity of the corresponding image pixels from the mean intensity of the $n_j > 1$ image pixels, σ_j^2, is calculated. For a good lens model the mean value of σ_j^2 over the pixels in the source map should be no greater than the variance of the noise in the image

σ^2_{noise}. If the number of degrees of freedom in the source map is taken to be equal to the number of pixels, then the statistical measure of the quality of the lens model is $\chi^2 = \sum(\sigma_j^2/\sigma^2_{\text{noise}})$, where the sum is taken over the j source pixels. The lens parameters can thus be varied to minimize χ^2. In practice the procedure is more complicated than indicated by the description above. Modifications are included to take account of the finite resolution of the image, which has the effect of spreading the mapping of each source pixel over a number of image pixels. Also, for any unresolved structure in the source, the intensity of the corresponding structure in the image depends on the magnification of the lens.

Ellithorpe, Kochanek, and Hewitt (1996) introduced a *visibility lensclean* procedure in which the CLEAN components are removed from the ungridded visibility values under the constraints of a lens model. The squared deviations of the measured visibility from a model are used to determine a χ^2 statistic. The quality of the fit is judged from the variance of the measured visibility and the number of degrees of freedom is $2N_{\text{vis}} - 3N_{\text{src}} - N_{\text{lens}}$, where N_{vis} is the number of visibility measurements (which each have two degrees of freedom), N_{src} is the number of independent CLEAN components in the source model (three degrees of freedom, from position and amplitude), and N_{lens} is the number of parameters in the lens model. Ellithorpe et al. compared results of the original lensclean with visibility lensclean, and found the best results from the latter, with a further improvement if a self-calibration step is added. The use of the MEM algorithm as an alternative to CLEAN has also been investigated (Wallington, Narayan, and Kochanek 1994).

BIBLIOGRAPHY

Roberts, J. A., Ed., *Indirect Imaging*, Cambridge Univ. Press, Cambridge, UK, 1984.

van Schooneveld, C., Ed., *Image Formation from Coherence Functions in Astronomy*, Reidel, Dordrecht, 1979.

REFERENCES

Ables, J. G., Maximum Entropy Spectral Analysis, *Astron. Astrophys. Suppl.*, **15**, 383–393, 1974.

Bajaja, E., and G. D. van Albada, Complementing Aperture Synthesis Radio Data by Short Spacing Components from Single Dish Observation, *Astron. Astrophys.*, **75**, 251–254, 1979.

Baldwin, J. E., The Design of Large Arrays at Meter Wavelengths, in *Radio Astronomical Seeing*, J. E. Baldwin and Wang, S., Eds., International Academic Publishers and Pergamon Press, Oxford, 1990.

Baldwin, J. E. and P. J. Warner, Phaseless Aperture Synthesis, *Mon. Not. R. Astron. Soc.*, **182**, 411–422, 1978.

Baldwin, J. E. and P. J. Warner, Fundamental Aspects of Aperture Synthesis with Limited or No Phase Information, in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., Reidel, Dordrecht 1979, pp. 67–82.

Bates, R. H. T., Contributions to the Theory of Intensity Interferometry, *Mon. Not. R. Astron. Soc.*, **142**, 413–428, 1969.

Bates, R. H. T., Uniqueness of Solutions to Two-Dimensional Fourier Phase Problems for Localized and Positive Images, *Comp. Vision, Graphics Image Process.*, **25**, 205–217, 1984.

Bracewell, R. N., Inversion of Nonplanar Visibilities, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Univ. Press, 1984, pp. 177–183.

Bracewell, R. N., *Two-Dimensional Imaging*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

Bracewell, R. N. and J. A. Roberts, Aerial Smoothing in Radio Astronomy, *Aust. J. Phys.*, **7**, 615–640, 1954.

Briggs, D. S., *High Fidelity Deconvolution of Moderately Resolved Sources*, Ph.D. thesis, New Mexico Institute of Mining and Technology, 1995.

Bruck, Y. M. and L. G. Sodin, On the Ambiguity of the Image Reconstruction Problem, *Opt. Commun.*, **30**, 304–308, 1979.

Bryan, R. K. and J. Skilling, Deconvolution by Maximum Entropy, as Illustrated by Application to the Jet of M87, *Mon. Not. R. Astron. Soc.*, **191**, 69–79, 1980.

Clark, B. G., An Efficient Implementation of the Algorithm "CLEAN," *Astron. Astrophys.*, **89**, 377–378, 1980.

Clark, B. G., Large Field Mapping, in *Synthesis Mapping, Proc. NRAO Workshop No. 5* (Socorro, NM, June 21–25, 1982), A. R. Thompson and L. R. D'Addario, Eds., National Radio Astronomy Observatory, Green Bank, WV, 1982.

Conway, J. E., T. J. Cornwell, and P. N. Wilkinson, Multi-Frequency Synthesis: a New Technique in Radio Interferometric Imaging, *Mon. Not. R. Astr. Soc.*, **246**, 490–509, 1990.

Cornwell, T. J., Self Calibration, in *Synthesis Mapping, Proc. NRAO Workshop No. 5* (Socorro, NM, June 21–25, 1982), A. R. Thompson and L. R. D'Addario, Eds., National Radio Astronomy Observatory, Green Bank, WV, 1982.

Cornwell, T. J., A Method of Stabilizing the Clean Algorithm, *Astron. Astrophys.*, **121**, 281–285, 1983.

Cornwell, T. J., Radio-interferometric Imaging of Very Large Objects, *Astron. Astrophys.*, **202**, 316–321, 1988.

Cornwell, T. J., The Applications of Closure Phase to Astronomical Imaging, *Science*, **245**, 263–269, 1989.

Cornwell, T. J., Imaging Concepts, in *Very Long Baseline Interferometry and the VLBA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron. Soc. Pacific Conf. Ser. **82**, 39–56, 1995.

Cornwell, T. J., R. Braun, and D. S. Briggs, Deconvolution, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 151–170, 1999.

Cornwell, T. J. and K. F. Evans, A Simple Maximum Entropy Deconvolution Algorithm, *Astron. Astrophys.*, **143**, 77–83, 1985.

Cornwell, T. J., M. A. Holdaway, and J. M. Uson, Radio-interferometric Imaging of Very Large Objects: Implications for Array Design, *Astron. Astrophys.*, **271**, 697–713, 1993.

Cornwell, T. J. and R. A. Perley, Radio-interferometric Imaging of Very Large Fields, *Astron. Astrophys.*, **261**, 353–364, 1992.

Cornwell, T. J. and P. N. Wilkinson, A New Method for Making Maps with Unstable Radio Interferometers, *Mon. Not. R. Astron. Soc.*, **196**, 1067–1086, 1981.

Cotton, W. D., A Method Of Mapping Compact Structure in Radio Sources Using VLBI Observations, *Astron. J.*, **84**, 1122–1128, 1979.

Ekers, R. D. and Rots, A. H., Short Spacing Synthesis from a Primary Beam Scanned Interferometer, in *Image formation from Coherence Functions in Astronomy*, C. van Schoonveld, Ed., Reidel, Dordrecht, 1979, pp. 61–63.

Ellithorpe, J. D., C. S. Kochanek, and J. N. Hewitt, Visibility Lensclean and the Reliability of Deconvolved Radio Images, *Astrophys. J.*, **464**, 556–567, 1996.

Erickson, W. C., Long Wavelength Interferometry, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 601–612, 1999.

Fienup, J. R., Reconstruction of an Object from the Modulus of its Fourier Transform, *Opt. Lett.*, **3**, 27–29, 1978.

Fort, D. N. and H. K. C. Yee, A Method of Obtaining Brightness Distributions from Long Baseline Interferometry, *Astron. Astrophys.*, **50**, 19–22, 1976.

Frieden, B. R., Restoring with Maximum Likelihood and Maximum Entropy, *J. Opt. Soc. Am.*, **62**, 511–518, 1972.

Gull, S. F. and G. J. Daniell, Image Reconstruction from Incomplete and Noisy Data, *Nature*, **272**, 686–690, 1978.

Gull, S. F. and G. J. Daniell, The Maximum Entropy Method, in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., Reidel, Dordrecht, 1979, pp. 219–225.

Högbom, J. A., Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines, *Astron. Astrophys. Suppl.*, **15**, 417–426, 1974.

Högbom, J. A., The Introduction of A Priori Knowledge in Certain Processing Algorithms, in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., Reidel, Dordrecht, 1979, pp. 237–239.

Jaynes, E. T., Prior Probabilities, *IEEE Trans. Syst. Sci. Cyb.*, **SSC-4**, 227–241, 1968.

Jaynes, E. T., On the Rationale of Maximum-Entropy Methods, *Proc. IEEE*, **70**, 939–952, 1982.

Kassim, N. E., R. A. Perley, W. C. Erickson, and K. S. Dwarakanath, Subarcminute Resolution Imaging of Radio Sources at 74 MHZ with the Very Large Array, *Astron. J.*, **106**, 2218–2228, 1993.

Kochanek, C. S. and R. Narayan, Lensclean: An Algorithm for Inverting Extended, Gravitationally Lensed Images with Application to the Radio Ring Lens PKS 1830-211, *Astrophys. J.*, **401**, 461–473, 1992.

Lawson, C. L. and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

Lonsdale, C. J., S. S. Doelman, R. J. Capallo, J. N. Hewitt, and A. R. Whitney, Exploring the Performance of Large-N Radio Astronomical Arrays, in *Radio Telescopes*, H. R. Butcher, Ed., Proc. SPIE, **4015**, 126–134, 2000.

McClean, D. J., A Simple Expansion Method for Wide-Field Mapping, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Univ. Press, Cambridge, UK, 1984, pp. 185–191.

Napier, P. J., The Brightness Temperature Distributions Defined by a Measured Intensity Interferogram, *NZ J. Sci.*, **15**, 342–355, 1972.

Napier, P. J. and R. H. T. Bates, Inferring Phase Information from Modulus Information in Two-Dimensional Aperture Synthesis, *Astron. Astrophys. Suppl.*, **15**, 427–430, 1974.

Napier, P. J., A. R. Thompson, and R. D. Ekers, The Very Large Array: Design and Performance of a Modern Synthesis Radio Telescope, *Proc. IEEE*, **71**, 1295–1320, 1983.

Napier, P. J., D. S. Bagri, B. G. Clark, A. E. E. Rogers, J. D. Romney, A. R. Thompson, and R. C. Walker, The Very Long Baseline Array, *Proc. IEEE*, **82**, 658–672, 1994.

Narayan, R. and R. Nityananda, Maximum Entropy–Flexibility Versus Fundamentalism, in *Indirect Imaging*, J. A. Roberts, Ed., Cambridge Univ. Press, Cambridge, UK, 1984, pp. 281–290.

Narayan, R. and S. Wallington, Introduction to Basic Concepts of Gravitational Lensing, in *Gravitational Lenses*, R. Kayser, T. Schramm, and L. Nieser, Eds., Springer-Verlag, Berlin, 1992, pp. 12–26.

Nityananda, R. and R. Narayan, Maximum Entropy Image Reconstruction—a Practical Non-Information-Theoretic Approach, *J. Astrophys. Astron.*, **3**, 419–450, 1982.

Pearson, T. J. and A. C. S. Readhead, Image Formation by Self Calibration in Radio Astronomy, *Ann. Rev. Astron. Astrophys.*, **22**, 97–130, 1984.

Perley, R. A., High Dynamic Range Imaging, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser., **6**, 287–313, 1989.

Perley, R. A., High Dynamic Range Imaging, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 275–299, 1999a.

Perley, R. A., Imaging with Non-Coplanar Arrays, in *Synthesis Imaging in Radio Astronomy II*, G. B. Taylor, C. L. Carilli, and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **180**, 383–400, 1999b.

Perley, R. A., J. W. Dreher, and J. J. Cowan, The Jet and Filaments in Cygnus A., *Astrophys. J.*, **285**, L35–L38, 1984.

Ponsonby, J. E. B., An Entropy Measure for Partially Polarized Radiation and Its Application to Estimating Radio Sky Polarization Distributions from Incomplete "Aperture Synthesis" Data by the Maximum Entropy Method, *Mon. Not. R. Astron. Soc.*, **163**, 369–380, 1973.

Readhead, A. C. S., R. C. Walker, T. J. Pearson, M. H. Cohen, Mapping Radio Sources with Uncalibrated Visibility Data, *Nature*, **285**, 137–140, 1980.

Readhead, A. C. S. and P. N. Wilkinson, The Mapping of Compact Radio Sources from VLBI Data, *Astrophys. J.*, **223**, 25–36, 1978.

Rogers, A. E. E., Methods of Using Closure Phases in Radio Aperture Synthesis, *Soc. Photo-Opt. Inst. Eng.*, **231**, 10–17, 1980.

Rogers, A. E. E., H. F. Hinteregger, A. R. Whitney, C. C. Counselman, I. I. Shapiro, J. J. Wittels, W. K. Klemperer, W. W. Warnock, T. A. Clark, L. K. Hutton, G. E. Marandino, B. O. Rönnäng, O. E. H. Rydbeck, and A. E. Niell, The Structure of Radio Sources 3C273B and 3C84 Deduced from the "Closure" Phases and Visibility Amplitudes Observed with Three-Element Interferometers, *Astrophys. J.*, **193**, 293–301, 1974.

Sault, R. J. and T. A. Oosterloo, Imaging Algorithms in Radio Interferometry, in *Review of Radio Science 1993–1996*, W. R. Stone, Ed., Oxford Univ. Press, Oxford, UK, 1996, pp. 883–912.

Sault, R. J., L. Stavely-Smith, and W. N. Brouw, An Approach to Interferometric Mosaicing, *Astron. Astrophys. Supp.*, **120**, 375–384, 1996.

Sault, R. J. and M. H. Wieringa, Multi-Frequency Synthesis Techniques in Radio Interferometric Imaging, *Astron. Astrophys. Suppl. Ser.*, **108**, 585–594, 1994.

Schwab, F. R., Adaptive Calibration of Radio Interferometer Data, *Soc. Photo-Opt. Inst. Eng.*, **231**, 18–24, 1980.

Schwab, F. R., Relaxing the Isoplanatism Assumption in Self Calibration: Applications to Low-Frequency Radio Astronomy, *Astron. J.*, **89**, 1076–1081, 1984.

Schwarz, U. J., Mathematical-Statistical Description of the Iterative Beam Removing Technique (Method CLEAN), *Astron. Astrophys.*, **65**, 345–356, 1978.

Schwarz, U. J., The Method "CLEAN"—Use, Misuse, and Variations, in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., Reidel, Dordrecht 1979, pp. 261–275.

Skilling, J. and R. K. Bryan, Maximum Entropy Image Reconstruction: General Algorithm, *Mon. Not. R. Astron. Soc.*, **211**, 111–124, 1984.

Sramek, R. A. and F. R. Schwab, Imaging, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser., **6**, 117–138, 1989.

Subrahmanya, C. R., An Optimum Deconvolution Method, in *Image Formation from Coherence Functions in Astronomy*, C. van Schooneveld, Ed., Reidel, Dordrecht 1979, pp. 287–290.

Thomasson, P., MERLIN, *Q. J. Royal Astron. Soc.*, **27**, 413–431, 1986.

van Gorkom, J. H. and R. D. Ekers, Spectral Line Imaging II: Calibration and Analysis, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pacific Conf. Ser. **6**, 341-353, 1989.

Waldram, E. M. and M. M. McGilchrist, Beam-Sets—A New Approach to the Problem of Wide-Field Mapping with Non-Coplanar Baselines, *Mon. Not. Royal Astron. Soc.*, **245**, 532–541, 1990.

Wallington, S., R. Narayan, and C. S. Kochanek, Gravitational Lens Inversion Using the Maximum Entropy Method, *Astrophys. J.*, **426**, 60–73, 1994.

Walsh, D., R. F. Carswell, and R. J. Weymann, 0957+561A,B: Twin Quasistellar Objects or Gravitational Lens? *Nature*, **279**, 381–384, 1979.

Welch, W. J., and D. D. Thornton, An Introduction to Millimeter and Submillimeter Interferometry and a Summary of the Hat Creek System, *Int. Symp. Millimeter and Submillimeter Wave Radio Astronomy*, International Scientific Radio Union, Institut de Radio Astronomie Millimetrique, Granada, Spain, 1985, pp. 53–64.

Wernecke, S. J., Two-Dimensional Maximum Entropy Reconstruction of Radio Brightness, *Radio Sci.*, **12**, 831–844, 1977.

Wernecke, S. J. and L. R. D'Addario, Maximum Entropy Image Reconstruction, *IEEE Trans. Comput.*, **C-26**, 351–364, 1977.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

12 Interferometer Techniques for Astrometry and Geodesy

The output fringe pattern of an interferometer provides a measure of the scalar product of the baseline and source-position vectors, $\mathbf{D} \cdot \mathbf{s}$. Up to this point we have assumed that these factors are describable by constants that can be specified with high accuracy. However, the measurement of source positions to an accuracy of milliarcseconds (mas) requires that variation in the earth's rotation vector be taken into account. The required baseline accuracy is comparable to that at which variation in the antenna positions resulting from crustal motions of the earth can be detected. The calibration of the baseline and the measurement of source positions can be accomplished in a single observing period of one or more days. Geodetic* data are obtained from repetition of this procedure over intervals of months or years, which reveals the variation in the baseline and earth-rotation parameters.

This chapter is concerned with the techniques by which angular positions can be measured with the greatest possible accuracy, and with the design of interferometers for optimum determination of source-position, baseline, and geodetic parameters.

The redefinition of the meter has an interesting implication for the units of baseline length derived from interferometric data. An interferometer measures the relative time of arrival of the signal wavefront at the two antennas, that is, the geometric delay. Baselines determined from interferometric data are therefore in units of light travel time. Conversion to meters formerly depended on the value chosen for c. However, in 1983 the Conférence Générale des Poids et Mesures adopted a new definition of the meter: "the meter is the length of the path traveled by light in vacuum during a time interval of 1/299,792,458 of a second." The second and the speed of light are now primary quantities, and the meter is a derived quantity. Thus baseline lengths can be given unambiguously in meters. Issues related to fundamental units are discussed by Petley (1983).

12.1 REQUIREMENTS FOR ASTROMETRY

Position measurements of radio sources accurate to no better than a few tens of arcseconds are mainly of historical interest and have been mentioned in Chap-

*For simplicity we use the term *geodetic* to include geodynamic and static phenomena regarding the shape and orientation of the earth.

ter 1. In the earliest studies of this kind, the baseline vector was often established by surveying the antenna locations, the instrumental phase was estimated by calibration of the transmission lines, and the positions of the fringe maxima on the sky were thereby deduced. An informative review of these techniques, including various procedures for minimizing instrumental errors, is given by Smith (1952). In this chapter, we are concerned with more recent procedures for which the precision is of order one milliarcsecond, or better. We begin with a heuristic discussion of how baseline and source position parameters may be determined. A formal discussion is given in Section 12.2.

In the measurement of source declination with a tracking interferometer, it is possible to solve for both the source position and the baseline parameters independently. This can be illustrated simply by the following consideration. The phase of the fringe pattern for an interferometer is $2\pi w$, where w is the spacing component given by Eq. (4.3). The phase can be written

$$\phi = 2\pi D_\lambda[\sin d \sin\delta + \cos d \cos\delta \cos(H - h)] + \phi_{\text{in}}, \tag{12.1}$$

where D_λ is the length of the baseline in wavelengths, H and δ are the hour angle and declination of the source, h and d are the hour angle and declination of the baseline, and ϕ_{in} is an instrumental phase term. For an east–west baseline, $h = -\pi/2$ when measured from the local meridian, $d = 0$, and the phase reduces to

$$\phi = -2\pi D_\lambda \cos\delta \sin H + \phi_{\text{in}}. \tag{12.2}$$

Thus ϕ is proportional to $\cos\delta$, and by observing sources close to the celestial equator, where the dependence on δ is small, D_λ can be established with high accuracy [e.g., Ryle and Elsmore (1973)]. Positions of sources at higher declinations can then be determined, and these can be used to calibrate a north–south baseline for more accurate measurement of low-declination sources.

In the determination of right ascension, interferometer observations provide relative measurements, that is, the differences in right ascension among different sources. The zero of right ascension is defined as the great circle through the pole and through the intersection of the celestial equator and the ecliptic at the vernal equinox. The vernal equinox is the point at which the apparent position of the sun moves from the southern to the northern celestial hemisphere. This direction can be located in terms of the motions of the planets, which are well-defined objects for optical observations. It has been related to the positions of bright stars that provide a reference system for optical measurements of celestial position. Relating the radio measurements to the zero of right ascension is less easy, since solar system objects are generally weak or do not contain sharp enough features in their radio structure. In the 1970s, results were obtained from the lunar occultation of the source 3C273B (Hazard et al. 1971) and from measurements of the weak radio emission from nearby stars such as Algol (β Persei) (Ryle and Elsmore 1973, Elsmore and Ryle 1976).

Techniques have been suggested for determining the direction of the vernal equinox directly from passive radio measurements. For example, radio interferometric observations of the minor planets could be used to determine their posi-

tions in the reference frame of the extragalactic sources (Johnston, Seidelmann, and Wade 1982). Another method is based on a comparison of the positions of pulsars obtained from pulse timing measurements with positions obtained from radio interferometry using extragalactic sources as position calibrators (Fomalont et al. 1992, Taylor et al. 1984, Bartel et al. 1985). Since the timing measurements yield positions relative to the coordinate frame defined by the earth's orbit, while the interferometer measurements refer to the reference frame of the extragalactic sources, the position of the vernal equinox and other dynamical parameters of the earth's orbit can be related to the latter frame.

In the reduction of interferometer measurements in astrometry, the visibility data are interpreted basically in terms of the positions of point sources. The data processing is equivalent, in effect, to model fitting using delta-function intensity components, the visibility function for which has been discussed in Section 4.4. The essential position data are determined from the calibrated visibility phase or, in some VLBI observations, from the geometric delay as measured by maximization of the cross-correlation of the signals (i.e., the use of the bandwidth pattern) and from the fringe frequency. Because the position information is contained in the visibility phase, measurements of closure phase discussed in Section 10.3 are of use in astrometry and geodesy only insofar as they can provide a means of correcting for the effects of source structure. Uniformity of (u, v) coverage is less important than in imaging because high dynamic range is generally not needed. Determination of the position of an unresolved source depends on interferometry with precise phase calibration and a sufficient number of baselines to avoid ambiguity in the position.

Reference Frames

A reference frame based on the positions of distant extragalactic objects can be expected to show greater temporal stability than a frame based on stellar positions, and to approach more closely the conditions of an inertial frame. An inertial frame is one that is at rest or in uniform motion with respect to absolute space, and not in a state of acceleration or rotation [see, e.g., Mueller (1981)]. Newton's first law holds in such a frame. A detailed description of astronomical reference frames is given by Johnston and de Vegt (1999); see bibliography. The International Celestial Reference System (ICRS) adopted by the IAU specifies the zero points and directions of the axes of the coordinate system for celestial positions. The measured positions of a set of reference objects in the coordinates of the reference system provide the International Celestial Reference Frame (ICRF). Thus, the frame provides the reference points with respect to which positions of other objects are measured within the coordinate system.

The most accurate measurements of celestial positions are those of selected extragalactic sources observed by VLBI. Large data bases of such high-resolution observations exist as a result of measurements made for purposes of geodesy and astrometry. These measurements have been made mainly since 1979, using Mark III VLBI systems with dual frequencies of 2.3 and 8.4 GHz to allow calibration of atmospheric effects. The positions are determined mainly by the 8.4 GHz

data. An analysis (Ma et al. 1998) used Mark III measurements through 1995, and included 1.6×10^6 measurements of group delay and phase delay rate. Data for each of 618 sources were examined. Criteria for exclusion of a source included inconsistency in the position measurements, evidence of motion, or presence of extended structure. In this study, 212 sources were found that passed all tests, 294 failed in one criterion, and 102 other sources, including 3C273, failed in several. The 212 sources in the best category were used to define the reference frame. Only 27% of these are in the southern hemisphere. A global solution provides the positions of the sources together with the antenna positions and various geodetic and atmospheric parameters. Position errors of the 212 defining sources are mostly less than 0.5 mas in both right ascension and declination and less than 1 mas in almost all cases. The measurements were adopted by the International Astronomical Union (IAU) in 1998 as the ICRF. Earlier frames have all been based on optical positions of stars, most recently those of the FK5 catalog.

About 50% of sources in the ICRF have redshifts greater than 1.0. The use of such distant objects to define the reference frame provides a level of astrometric uncertainty at least an order of magnitude better than optical measurements of stars. The level of uncertainty in the connection between the radio and optical frames is essentially the uncertainty in optical positions, which will be greatly improved in the future through the use of optical interferometry in space-based programs. Radio measurements of the positions of some of the nearer stars will provide a comparison between the radio and optical frames. Lestrade et al. (1990, 1995) have measured the positions of about 10 stars by VLBI, achieving accuracy in the range 0.5–1.5 mas. These results provide a link between the ICRF and the star positions in the Hipparcos catalog. The visual magnitudes of the known optical counterparts of the reference frame sources are mostly within the range 15–21, and precise positions of objects fainter than 18th magnitude are likely to be very difficult to obtain. Improvements in understanding the limits of accuracy of radio positions continue to be made. Fey and Charlot (1997) have studied corrections to positions for sources such as 3C273 that show resolvable structure, and have defined a structure index to estimate the quality of sources for astrometric measurements.

12.2 SOLUTION FOR BASELINE AND SOURCE-POSITION VECTORS

In determining an interferometer baseline, it is convenient to use calibrators whose angular positions are known with accuracy comparable to that required for the baseline. However, this is not essential, and it is often necessary to solve for source and baseline parameters simultaneously.

Connected-Element Systems

Consider an observation with a tracking interferometer of arbitrary baseline in which the source is unresolved. Let $\mathbf{D}_\lambda$ be the assumed baseline vector, in units of the wavelength, and $(\mathbf{D}_\lambda - \Delta\mathbf{D}_\lambda)$ be the true vector. Similarly,

let $\mathbf{s}$ be a unit vector indicating the assumed position of the source, and let $(\mathbf{s} - \Delta\mathbf{s})$ indicate the true position. Note that the convention used is Δ term = (approximate or assumed value) − (true value). The expected fringe phase, using the assumed positions, is $2\pi \mathbf{D}_\lambda \cdot s$. The observed phase, measured relative to the expected phase, is a function of the hour angle H of the source given by

$$\begin{aligned}\Delta\phi(H) &= 2\pi[\mathbf{D}_\lambda \cdot \mathbf{s} - (\mathbf{D}_\lambda - \Delta\mathbf{D}_\lambda) \cdot (\mathbf{s} - \Delta\mathbf{s})] + \phi_{\text{in}} \\ &= 2\pi(\Delta\mathbf{D}_\lambda \cdot \mathbf{s} + \mathbf{D}_\lambda \cdot \Delta\mathbf{s}) + \phi_{\text{in}}.\end{aligned} \tag{12.3}$$

A second-order term involving $\Delta\mathbf{D}_\lambda \cdot \Delta\mathbf{s}$ has been neglected since we assume that fractional errors in $\mathbf{D}_\lambda$ and $\mathbf{s}$ are small.

The baseline vector can be written in terms of the coordinate system introduced in Section 4.1:

$$\mathbf{D}_\lambda = \begin{bmatrix} X_\lambda \\ Y_\lambda \\ Z_\lambda \end{bmatrix}, \qquad \Delta\mathbf{D}_\lambda = \begin{bmatrix} \Delta X_\lambda \\ \Delta Y_\lambda \\ \Delta Z_\lambda \end{bmatrix}, \tag{12.4}$$

where X, Y, and Z form a right-handed coordinate system with Z parallel to the earth's spin axis and X in the meridian plane of the interferometer. The source-position vector can be specified in the (X, Y, Z) system in terms of the hour angle H and declination δ of the source by using Eq. (4.2):

$$\mathbf{s} = \begin{bmatrix} s_X \\ s_Y \\ s_Z \end{bmatrix} = \begin{bmatrix} \cos\delta\,\cos H \\ -\cos\delta\,\sin H \\ \sin\delta \end{bmatrix}. \tag{12.5}$$

Taking the differential of Eq. (12.5), we can write

$$\Delta\mathbf{s} \simeq \begin{bmatrix} -\sin\delta\cos H\,\Delta\delta + \cos\delta\sin H\,\Delta\alpha \\ \sin\delta\sin H\,\Delta\delta + \cos\delta\cos H\,\Delta\alpha \\ \cos\delta\,\Delta\delta \end{bmatrix}, \tag{12.6}$$

where $\Delta\alpha$ and $\Delta\delta$ are the angular errors in right ascension and declination, and $\Delta\alpha$ has the opposite sign to the corresponding error in hour angle. By substituting Eqs. (12.4)–(12.6) into (12.3), the expression for the measured phase can be written

$$\begin{aligned}\Delta\phi(H) &= \phi_0 + \phi_1 s_X - \phi_2 s_Y \\ &= \phi_0 + \phi_1\cos\delta\cos H + \phi_2\cos\delta\sin H,\end{aligned} \tag{12.7}$$

where

$$\phi_0 = 2\pi(\Delta Z_\lambda \sin\delta + Z_\lambda\cos\delta\,\Delta\delta) + \phi_{\text{in}}, \tag{12.8}$$

$$\phi_1 = 2\pi(\Delta X_\lambda + Y_\lambda\,\Delta\alpha - X_\lambda\tan\delta\,\Delta\delta), \tag{12.9}$$

and

$$\phi_2 = 2\pi(-\Delta Y_\lambda + X_\lambda \, \Delta\alpha + Y_\lambda \tan\delta \, \Delta\delta). \tag{12.10}$$

From Eq. (12.7) it is seen that $\Delta\phi(H)$ is a sinusoid in H with an offset ϕ_0. Thus the three parameters amplitude, phase, and offset can be measured for any source by observing periodically or continuously for approximately 12 h. If n_s sources are observed, $3n_s$ quantities are obtained. The number of unknown parameters required to specify the n_s positions, the baseline, and the instrumental phase (assumed to be constant) is $2n_s + 3$; the right ascension of one source is chosen arbitrarily. Thus, if $n_s \geq 3$, it is possible to solve for all the unknown quantities. Note that the sources should have as wide a range in declination as possible in order to distinguish ΔZ_λ from ϕ_{in} in Eq. (12.8). Least-mean-squares analysis provides simultaneous solutions for the instrumental parameters and the source positions. Usually, many more than three sources are observed, so there is redundant information, and variation of the instrumental phase with time as well as other parameters can be included in the solution. A discussion of the method of least-mean-squares analysis can be found in the appendix to this chapter.

Measurements with VLBI Systems

The use of independent local oscillators at the antennas in VLBI systems does not easily permit calibration of the fringe phase, except in special circumstances. The earliest method used for obtaining positional information in VLBI was the analysis of the fringe frequency (fringe rate). The fringe frequency is the time rate of change of the interferometer phase. Thus, from Eq. (12.1) the fringe frequency is

$$\nu_f = \frac{1}{2\pi}\frac{d\phi}{dt} = -\omega_e D_\lambda \cos d \cos\delta \sin(H - h) + \nu_{\text{in}}, \tag{12.11}$$

where ω_e is the angular velocity of rotation of the earth (dH/dt), and ν_{in} is an instrumental term equal to $d\phi_{\text{in}}/dt$. The component ν_{in} largely results from residual differences in the frequencies of the hydrogen masers which provide the local oscillator references at the antennas.

The quantity $D_\lambda \cos d$ is the projection of the baseline in the equatorial plane, denoted D_E. Thus Eq. (12.11) can be rewritten

$$\nu_f = -\omega_e D_E \cos\delta \sin(H - h) + \nu_{\text{in}}. \tag{12.12}$$

The polar component of the baseline (the projection of the baseline along the polar axis) does not appear in the equation for fringe frequency. An interferometer with a baseline parallel to the spin axis of the earth has lines of constant phase parallel to the celestial equator, and the interferometer phase does not change with hour angle. Therefore, the polar component of the baseline cannot be determined from the analysis of fringe frequency.

The usual practice in VLBI is to refer hour angles to the Greenwich meridian. We follow this convention and use a right-handed coordinate system with X through the Greenwich meridian and with Z toward the north celestial pole. Thus, in terms of the Cartesian coordinates for the baseline, Eq. (12.12) becomes

$$\nu_f = -\omega_e \cos\delta (X_\lambda \sin H + Y_\lambda \cos H) + \nu_{\text{in}}. \tag{12.13}$$

The residual fringe frequency $\Delta\nu_f$, that is, the difference between the observed and expected fringe frequencies, can be calculated by taking the derivatives of Eq. (12.13) with respect to δ, H, X_λ, and Y_λ, and also including the unknown quantity ν_{in}. We thereby obtain

$$\Delta\nu_f = \nu_{\text{in}} + a_1 \cos H + a_2 \sin H, \tag{12.14}$$

where

$$a_1 = \omega_e (Y_\lambda \sin\delta\, \Delta\delta + X_\lambda \cos\delta\, \Delta\alpha - \cos\delta\, \Delta Y_\lambda) \tag{12.15}$$

and

$$a_2 = \omega_e (X_\lambda \sin\delta\, \Delta\delta - Y_\lambda \cos\delta\, \Delta\alpha - \cos\delta\, \Delta X_\lambda). \tag{12.16}$$

Note that $\Delta\nu_f$ is a diurnal sinusoid and that the average value of $\Delta\nu_f$ is the instrumental term ν_{in}. Information about source positions and baselines must come from the two parameters a_1 and a_2. Therefore, unlike the case of fringe phase [Eq. (12.7)] where three parameters per source are available, it is not possible to solve for both source and baseline parameters with fringe-frequency data. For example, from observations of n_s sources, $2n_s + 1$ quantities are obtained. The total number of unknowns (two baseline parameters, $2n_s$ source parameters, and ν_{in}) is $2n_s + 3$. If the position of one source is known, the rest of the source positions and X_λ, Y_λ, and ν_{in} can be determined. Note that the accuracy of the measurements of source declinations is reduced for sources close to the celestial equator because of the $\sin\delta$ factor in Eqs. (12.15) and (12.16).

As an illustration of the order of magnitude of the parameters involved in fringe-frequency observations, consider two antennas with an equatorial component of spacing equal to 1000 km and an observing wavelength of 3 cm. Then $D_E \simeq 3 \times 10^7$ wavelengths, and the fringe frequency for a low-declination source is about 2 kHz. Assume that the coherence time of the independent frequency standards is about 10 min. In this period 10^6 fringe cycles can be counted. If we suppose that the phase can be measured to 0.1 turn, ν_f will be obtained to a precision of 1 part in 10^7. The corresponding errors in D_E and angular position are 10 cm and 0.02 arcsec, respectively.

To overcome the limitations of fringe-frequency analysis, techniques for the precise measurement of the relative group delay of the signals at the antennas were developed. The use of bandwidth synthesis to improve the accuracy of delay measurements has been discussed in Section 9.8. The group delay is equal to the geometric delay τ_g except that, as measured, it also includes unwanted com-

ponents resulting from clock offsets at the antennas and atmospheric differences in the signal paths. The fringe phase measured with a connected-element interferometer observing at frequency ν is $2\pi\nu\tau_g$, modulo 2π. Except for the dispersive ionosphere, the group delay therefore contains the same type of information as the fringe phase, without the ambiguity resulting from the modulo 2π restriction. Thus group delay measurements permit a solution for baselines and source positions similar to that discussed above for connected-element systems, except that clock offset terms also must be included.

In most astrometric experiments, measurements of the group delay and the fringe frequency (or, equivalently, the rate of change of phase delay) are analyzed together. The intrinsic precision with which each of these quantities can be measured is derived in Appendix 12.1 [Eqs. (A12.27) and (A12.34)] and can be written

$$\sigma_f = \sqrt{\frac{3}{2\pi^2}}\left(\frac{T_S}{T_A}\right)\frac{1}{\sqrt{\Delta\nu\tau^3}} \tag{12.17}$$

and

$$\sigma_\tau = \frac{1}{\sqrt{8\pi^2}}\left(\frac{T_S}{T_A}\right)\frac{1}{\sqrt{\Delta\nu\tau}\,\Delta\nu_{\rm rms}}, \tag{12.18}$$

where σ_f and σ_τ are the rms errors in fringe frequency and delay, T_S and T_A are the system and antenna temperatures, $\Delta\nu$ is the IF bandwidth, τ is the integration time, and $\Delta\nu_{\rm rms}$ is the rms bandwidth introduced in Section 9.8 [see also Eq. (A12.32) and related text in Appendix 12.1]. $\Delta\nu_{\rm rms}$ is typically 40% of the spanned bandwidth. For a single rectangular RF band, $\Delta\nu_{\rm rms} = \Delta\nu/\sqrt{12}$. To express the measurement error as an angle, note that the geometric delay is

$$\tau_g = \frac{D}{c}\sin\theta, \tag{12.19}$$

where θ is the angle between the source vector and the plane perpendicular to the baseline. Thus, the sensitivity of the delay to angular changes is

$$\frac{\Delta\tau_g}{\Delta\theta_\tau} = \frac{D}{c}\cos\theta, \tag{12.20}$$

where $\Delta\theta_\tau$ is the increment in θ corresponding to an increment $\Delta\tau_g$ in τ_g. Similarly, the sensitivity of the fringe frequency to angular changes [since $\nu_f = \nu(d\tau_g/dt)$] is (for an east–west baseline)

$$\frac{\Delta\nu_f}{\Delta\theta_f} = D_\lambda\omega_e\cos\theta, \tag{12.21}$$

where $\Delta\theta_f$ is the increment in θ corresponding to an increment $\Delta\nu_f$ in ν_f. Thus by setting $\Delta\nu_f = \sigma_f$ and $\Delta\tau_g = \sigma_\tau$ and ignoring geometric factors, we obtain

the equation

$$\frac{\Delta\theta_\tau}{\Delta\theta_f} \simeq 2\pi \frac{\tau/t_e}{\Delta\nu/\nu}, \tag{12.22}$$

where $t_e = 2\pi/\omega_e$ is the period of the earth's rotation. Equation (12.22) describes the relative precision of delay and fringe-frequency measurements. In practice, measurements of delay are generally more accurate because of the noise imposed by the atmosphere. Measurements of fringe frequency are sensitive to the time derivative of atmospheric path length, and in a turbulent atmosphere this derivative can be large, while the average path length is relatively constant. Note that fringe-frequency and delay measurements are complementary. For example, with a VLBI system of known baseline and instrumental parameters, the position of a source can be found from a single observation using the delay and fringe frequency because these quantities constrain the source position in approximately orthogonal directions. The earliest analyses of fringe-frequency and delay measurements to determine source positions and baselines were made by Cohen and Shaffer (1971) and Hinteregger et al. (1972).

The accuracy with which group delay can be used to measure a source position is proportional to the reciprocal of the bandwidth $1/\Delta\nu$. Similarly, the accuracy with which phase can be used to measure a source position is proportional to the reciprocal of the observing frequency $1/\nu$. Since the proportionality constants are approximately the same, the relative accuracy of these techniques is $\nu/\Delta\nu$. This ratio of the observing frequency to the bandwidth, including effects of bandwidth synthesis, is commonly one to two orders of magnitude. On the other hand, the antenna spacings used in VLBI are one to two orders of magnitude greater than those used in connected-element systems. Thus the accuracy of source positions estimated from group delay measurements with VLBI systems is comparable to the accuracy of those estimated from fringe phase measurements on connected-element systems having much shorter baselines. VLBI position measurements using phase referencing, as described below, are the most accurate of radio methods.

The ultimate limitations on ground-based interferometry are imposed by the atmosphere. Dual-frequency-band measurements effectively remove ionospheric phase noise (see Section 13.3 under *Calibration of Ionospheric Delay*). The rms phase noise of the troposphere increases about as $d^{5/6}$, where d is the projected baseline length, for baselines shorter than a few kilometers [see Eq. (13.101) and Table 13.3]. In this regime, measurement accuracies of angles improve only slowly with increasing baseline length. For baselines greater than $\sim$100 km, the effects of the troposphere above the interferometer elements are uncorrelated, and the measurement accuracy might be expected to improve more rapidly with baseline length. However, for widely spaced elements, the zenith angle can be significantly different, and the atmospheric model becomes very important. The angular accuracy achievable with connected-element systems approaches 10^{-2} arcsec (Kaplan et al. 1982, Perley 1982), and with VLBI it exceeds 10^{-3} arcsec (Fanselow et al. 1984; Herring, Gwinn, and Shapiro 1985; Lestrade 1991; Ma et al. 1998). As an example, the motion of the Sagittarius A* source at the Galactic center has been measured by Backer and Sramek (1999) over a period of 16

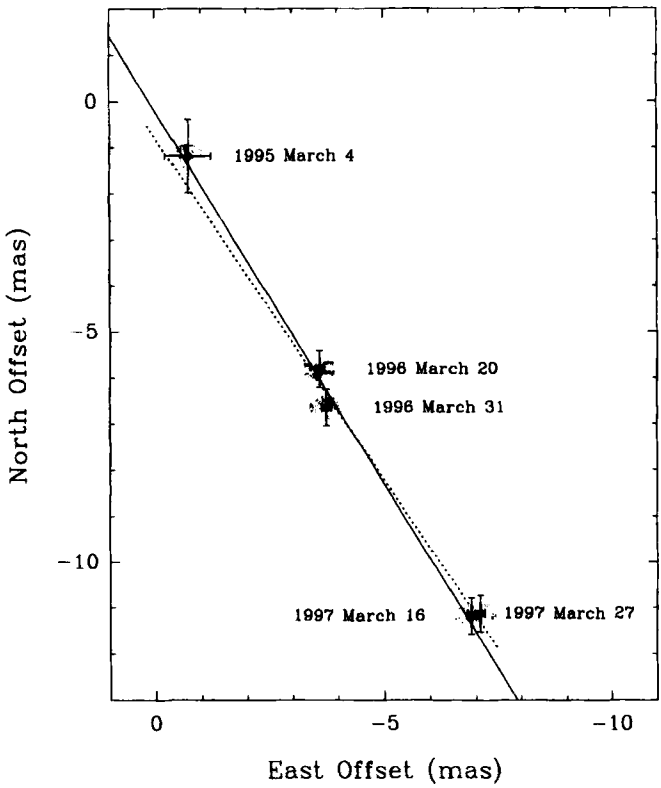


Figure 12.1 Apparent positions of the radio source in the Galactic center, Sgr A*, relative to the extragalactic callibrator J1745-283, measured over a two-year period at 43 GHz with VLBI. The shaded ellipses around the measurement points indicate the scatter-broadened size of Sgr A* (see Fig. 13.25). The one-sigma error bars in the measurements are also shown. The broken line is the variance-weighted least-squares fit to the data, and the solid line indicates the orientation of the Galactic plane. The motion is almost entirely in galactic longitude, attributable to the solar motion around the center of the Galaxy of 219 ± 20 km s^{-1}, for a distance between the sun and Galactic center of 8 kpc. The limit on the residual motion of Sgr A* is nearly two orders of magnitude less than that of the motions of stars lying within a projected distance of about 0.02 pc of Sgr A*. These stellar motions suggest that about 2.6×10^6 $M_\odot$ of matter are contained within 0.02 pc of Sgr A*, and the lack of detected motion of Sgr A* itself suggests that a mass of at least 10^3 $M_\odot$ must be associated with the radio source Sgr A*. From Reid et al. (1999), ©1999 American Astron. Soc.

years using the VLA (connected-element array), and by Reid et al. (1999) over a two-year period using the VLBA (VLBI array). The change in position measured with the VLBA is shown in Fig. 12.1.

Phase Referencing in VLBI

In VLBI measurements of relative positions of closely spaced sources, it is possible to measure the relative fringe phases and thus obtain positional accuracy cor-

responding to the very high angular resolution inherent in the long baselines. The most accurate measurements can be made when the sources are sufficiently close that both fall within the antenna beams [see, e.g., Marcaide and Shapiro (1983)], or when they are no more than a few degrees apart so that tropospheric and ionospheric effects are closely matched (Shapiro et al. 1979, Bartel et al. 1984, Ros et al. 1999). In such cases one source can be used as a calibrator in a manner similar to that for phase calibration in connected-element arrays. In VLBI this procedure is referred to as *phase referencing*. It allows imaging of sources for which the flux densities are too low to permit satisfactory self-calibration. The description here follows reviews of phase referencing procedures by Alef (1989) and Beasley and Conway (1995).

In phase referencing observations measurements are made alternately on the target source and on a nearby calibrator, with periods of a few minutes on each. (Note that the calibrator is also referred to as the phase reference source.) The rate of change of phase during these measurements must be slow enough that, from one calibrator measurement to the next, it is possible to interpolate the phase without ambiguity factors of 2π. It is therefore necessary to use careful modeling to remove geodetic and atmospheric effects, including tectonic plate motions, polar motion, earth tides, and ocean loading, and to make precise corrections for precession and nutation on the source positions. More subtle effects may need to be taken into account; for example, gravitational distortions of antenna structures, which tend to cancel out in connected-element arrays, can affect VLBI baselines because of the difference in elevation angles at widely spaced locations. Phase referencing has become more useful as better models for these effects, together with increased sensitivity and phase stability of receiving systems, have been developed.

Consider the case where we observe the calibration source at time t_1, then the target source at time t_2, and then the calibrator again at time t_3. For any one of these observations the measured phase is

$$\phi_{\text{meas}} = \phi_{\text{vis}} + \phi_{\text{inst}} + \phi_{\text{pos}} + \phi_{\text{ant}} + \phi_{\text{atmos}} + \phi_{\text{ionos}}, \tag{12.23}$$

where the terms on the right-hand side are, respectively, the components of the phase due to the source visibility, instrumental effects (cables, clock errors, etc.), the error in the assumed source position, errors in assumed antenna positions, the effect of the neutral atmosphere, and the effect of the ionosphere. To correct the phase of the target source, we need to interpolate the measurements on the calibrator at t_1 and t_3 to estimate what the calibrator phase would have been if measured at t_2, and then subtract the interpolated phase from the measured phase for the target source. If the positions of the target source and the calibrator are sufficiently close on the sky (not more than a few degrees apart), lines of sight from any antenna to the two sources will pass through the same isoplanatic patch, so the differences in the atmospheric and ionospheric terms can be neglected. We can assume that the instrumental terms do not differ significantly with small position changes, and if the calibrator is unresolved, then its visibility phase is zero. If the calibrator is partially resolved, it should be strong enough to allow imaging by self-calibration, and correction can be made for its phase. Thus the

corrected phase of the target source reduces to

$$\phi^t - \tilde{\phi}^c = \phi^t_{\text{vis}} + (\phi^t_{\text{pos}} - \tilde{\phi}^c_{\text{pos}}), \tag{12.24}$$

where the superscripts t and c refer to the target source and calibrator, respectively, and the tilde indicates interpolated values. The right-hand side of Eq. (12.24) depends only on the structure and position of the target source and the position of the calibrator. Figure 12.2 shows an example of phase referencing in which fringe fitting was performed on the data for the reference source, that is, determination of baseline errors, offsets between time standards at the sites,

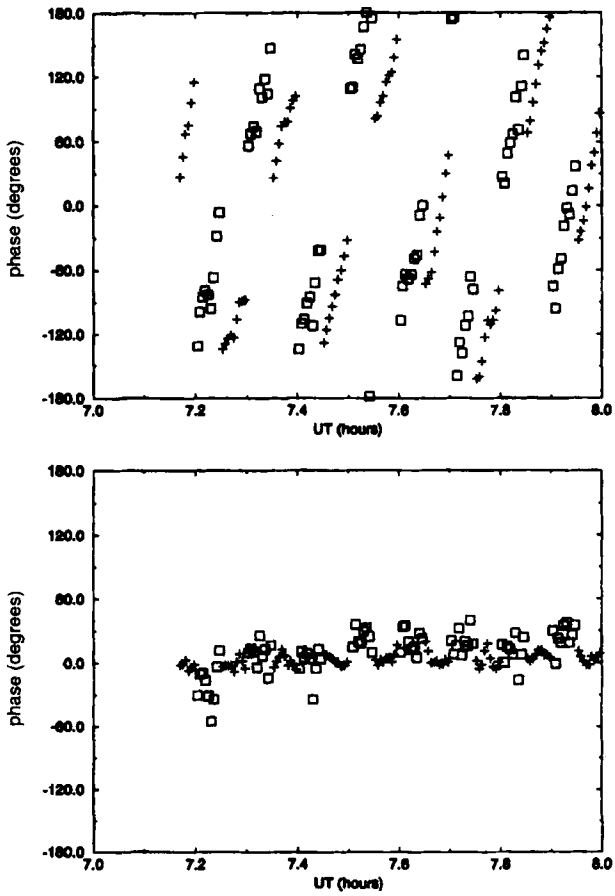


Figure 12.2 An example of phase referencing with the VLBA. The data are from the Brewster–Pie Town baseline with an observing frequency of 8.4 GHz. The top figure shows the uncalibrated data for two sources: 1638+398 (the target source, open squares) and 1641+399 (the phase reference source, crosses). The bottom figure shows the data for 1641 + 399 after fringe fitting, and the data for 1638+398 after phase referencing, using 1641 + 399 as the reference source. From Beasley and Conway (1995), courtesy of the Astron. Soc. Pacific Conf. Ser.

and instrumental phases. The results for the phase reference source (calibrator) are shown as crosses, and the resulting phase and phase rate corrections were interpolated to the times of the data points for the target source, shown as open squares. The corrected phases for the target source are shown in the lower diagram. For fringe fitting it is desirable to have a source that is unresolved and provides a strong signal, so a phase reference source should be chosen for these characteristics when the target source is weak or resolved.

Of the various effects in Eq. (12.23) that are removed by phase referencing, those that vary most rapidly with time are the atmospheric ones, and at frequencies above a few gigahertz they result from the troposphere rather than the ionosphere. Thus at centimeter wavelengths the tropospheric variations limit the time that can be allowed for each cycle of observation of the target and calibrator sources. Variations resulting from a moving-screen model of the troposphere are described in Section 13.1 under *Phase Fluctuations*; the characteristics of the screen are based on turbulence theory (Tartarski 1961). The relative rms variation in phase for the target and calibrator sources, the rays from which pass through the atmosphere a distance d_{tc} apart, is proportional to $d_{tc}^{5/6}$:

$$\sigma = \sigma_0 \, d_{tc}^{5/6}, \tag{12.25}$$

where σ_0 is the phase variation for 1-km ray spacing. In order to be able to interpolate the VLBI phase reference values from one calibrator observation to the next without ambiguity in the number of turns, the rms path length should not change by more than $\sim\lambda/8$ between successive calibrator scans. Then if the scattering screen moves horizontally with velocity v_s, the criterion above results in a limit on the time for one cycle of the target source and calibrator, $t_{\rm cyc}$. To determine this limit we put $d_{tc} = v_s t_{\rm cyc}$, and from Eq. (12.25) obtain

$$t_{\rm cyc} < \left(\frac{\pi}{4\sigma_0}\right)^{6/5} v_s^{-1}. \tag{12.26}$$

This result can be used to illustrate the time limit on the switching cycle. The empirical data in Table 13.3 (of Chapter 13) show that at $\lambda = 6$ cm (5 GHz frequency), the typical rms delay path is about 1 mm for $d_{tc} = 1$ km, at the VLA site. The corresponding value of σ_0 for 6-cm wavelength is 6°, which for $v_s = 0.01$ km s^{-1} yields $t_{\rm cyc} < 19$ min. This result is for typical conditions at the VLA site. For the same location and 1-km ray spacing, but under conditions described as "very turbulent," Sramek (1990) gives a value of 7.5 mm for the rms path deviation. The value of σ_0 for 6-cm wavelength is then 45°, resulting in $t_{\rm cyc} < 1.7$ min. The elevation angle of the source was not less 60° for this last observation, so even shorter switching times could apply at lower elevation angles.

At frequencies below $\sim$1 GHz the ionosphere becomes the limiting factor and medium-scale traveling ionospheric disturbances (MSTIDs), which have velocities of 100–300 m s^{-1} and wavelengths up to several hundred kilometers, become important (Hocke and Schlegel 1996). Phase fluctuations resulting from the ionosphere or troposphere are minimized in the approximate range 5–15 GHz, in which good performance can be obtained by phase referencing in VLBI.

There are also limits on the angular range that should be used in switching to the phase reference source, since even with a static atmosphere phase errors are introduced that increase with switching angle. Phase referencing over 7° with 0.1 mas precision has been demonstrated by Ros et al. (1999). An interesting application of phase referencing to measure the gravitational deflection of the radiation from the source 3C273B by the sun is described by Counselman et al. (1974). In this case the chosen comparison source, 3C279, was spaced approximately 10° from 3C273B, so as to be relatively unaffected by the sun. At each of two VLBI stations, two antennas working with a common frequency standard were used, one tracking each source, to provide a continuous phase comparison and allow precise removal of clock offset effects. Positional accuracy of order 3 mas was achieved at an operating frequency of 8.1 GHz. Measurement of relative position with respect to extragalactic sources has also been used to determine the parallax and proper motion of pulsar PSR B2021+51 (Campbell et al. 1996). The decreasing intensity of the pulsar with increasing frequency limited the observations to a single observing frequency of 2.218 GHz, but use of two reference sources within 2.5° of the pulsar allowed positional accuracy of approximately 0.3 mas to be achieved. As well as enabling imaging by Fourier transformation and deconvolution, phase referencing allows longer integration on sources that would otherwise be too weak to allow fringes to be detected within the coherence time. Self-calibration can then be used for flux densities lower than would otherwise be possible. Lists of sources suitable for phase calibrators can be found, for example, in Patnaik et al. (1992), Johnston et al. (1995), and Ma et al. (1998).

12.3 TIME AND THE MOTION OF THE EARTH

We now consider the effect of changes in the magnitude and direction of the earth's rotation vector on interferometric measurements. These changes cause variations in the apparent celestial coordinates of sources, the baseline vectors of the antennas, and universal time. The variations of the earth's rotation can be divided into three categories.

1. There are variations in the direction of the rotation axis, resulting mainly from precession and nutation of the spinning body. Since the direction of the axis defines the location of the pole of the celestial coordinate system, the result is a variation in the right ascension and declination of celestial objects.
2. The axis of rotation varies slightly with respect to the earth; that is, the positions on the earth at which this axis intersects the earth's surface vary. This effect is known as polar motion. Since the (X, Y, Z) coordinate system of baseline specification introduced in Section 4.1 takes the direction of the earth's axis as the Z axis, polar motion results in a variation of the measured baseline vectors (but not of the baseline length). It also results in a variation in universal time.
3. The rate of rotation varies as a result of atmospheric and crustal effects, and this again results in variation in universal time.

We briefly discuss these effects. Detailed discussions from a geophysical viewpoint can be found in Lambeck (1980).

Precession and Nutation

The gravitational effects of the sun, moon, and planets on the nonspherical earth produce a variety of perturbations in its orbital and rotational motions. To take account of these effects it is necessary to know the resulting variation of the ecliptic, which is defined by the plane of the earth's orbit, as well as the variation of the celestial equator, which is defined by the rotational motion of the earth. The gravitational effects of the sun and moon on the equatorial bulge (quadrupole moment) of the earth result in a precessional motion of the earth's axis around the pole of the ecliptic.

The earth's rotation vector is inclined at about 23.5° to the pole of its orbital plane, the ecliptic. The period of the resulting precession is approximately 26,000 years, corresponding to a motion of the rotation vector of 20 arcsec per year [$2\pi \sin(23.5°)/26{,}000$ radians per year]. The 23.5° obliquity is not constant, but is currently decreasing at a rate of 47 arcsec per century, due to the effect of the planets, which also cause a further component of precession. The luni-solar and planetary precessional effects, together with a smaller relativistic precession, are known as the general precession. Precession results in the motion of the line of intersection of the ecliptic and celestial equator. This line, called the line of nodes, defines the equinoxes and the zero of right ascension, which precess at a rate of 50 arcsec per year. In addition, the time-varying lunisolar gravitation effects cause nutation of the earth's axis with periods of up to 18.6 years and a total amplitude of about 9 arcsec. The principal variations of the ecliptic and equator are those just described, but other smaller effects also occur. The general accuracy within which positional variations can be calculated is better than 1 mas (Herring, Gwinn, and Shapiro 1985). Expressions for precession can be found in Lieske et al. (1977) and for nutation in Wahr (1981). The required procedures are discussed in texts on spherical astronomy, such as Woolard and Clemence (1966), Taff (1981), and Seidelmann (1992).

Since precession and nutation result in variations in celestial coordinates that can be as large as 50 arcsec per year for objects at low declinations, these effects must be taken into account in almost all observational work, whether astrometric or not. Positions of objects in astronomical catalogs are therefore reduced to the coordinates of standard epochs, B1900.0, B1950.0, or J2000.0. These dates denote the beginning of a Besselian year or Julian year, as indicated by the B or J. The positions correspond to the mean equator and equinox for the specified epoch, where "mean" indicates the positions of the equator and equinox resulting from the general precession, but not including nutation. For further explanation and a discussion of a method of conversion between standard epochs, see Seidelmann (1992). Correction is also required for aberration, that is, for the apparent shift in position resulting from the finite velocity of light and the motion of the observer. Two components are involved: annual aberration resulting from the earth's orbital motions, which has a maximum value of about 20 arcsec; and diurnal aber-

ration resulting from the rotational motion, which has a maximum value of 0.3 arcsec. The retarded baseline concept (Section 9.3) used in VLBI data reduction accounts for the diurnal aberration. For the nearer stars, corrections for proper motion (i.e., actual motion of the star through space) are required, and in some cases also for the parallax resulting from the changing position of the earth in its orbit. The impact of radio techniques, particularly VLBI, is resulting in refinement of the classical expressions and parameters. Effects such as the deflection of electromagnetic waves in the sun's gravitational field must also be included in positional work of the highest accuracy (see Section 13.5 under *Refraction*).

Polar Motion

The term *polar motion* denotes the variation of the pole of rotation of the earth (the geographic pole) with respect to the earth's crust. This results in a component of motion of the celestial pole that is distinct from precessional and other motions. Polar motion is largely, but not totally, of geophysical origin. The motion of the geographic pole around the pole of the earth's figure is irregular, but over the last century the distance between these two poles wandered by up to 0.5 arcsec, or 15 m on the earth's surface. In a year's time, the excursion of the figure axis is typically 6 m or less. The motion can be analyzed into several components, some regular and some highly irregular, and not all are understood. The two major components have periods of 12 and 14 months. The 12-month component is a forced motion due to the annual redistribution of water and of atmospheric angular momentum, and is far from any resonance. The 14-month component, known as the Chandler wobble (Chandler 1891), is the motion at a resonance frequency whose driving force is unknown. For a more detailed description, see Wahr (1996).

The motion of the pole of rotation is measured in angle or distance in the x and y directions as shown in Fig. 12.3. The (x, y) origin is the mean pole of 1900–1905, which is referred to as the *conventional international origin* (CIO), and the x axis is in the plane of the Greenwich meridian (Markowitz and Guinot 1968). Since polar motion is a small angular effect, it can often be ignored in mapping observations, especially if the visibility is measured with respect to a calibrator that is only a few degrees from the center of the field being mapped.

Universal Time

Like the motion of the earth, the system of timekeeping based on earth rotation is a complicated subject, and for a detailed discussion one can refer to Smith (1972) or to the texts mentioned in the discussion of precession and nutation above. We shall briefly review some essentials. Solar time is defined in terms of the rotation of the earth with respect to the sun. In practice, the stars present more convenient objects for measurement, so solar time is derived from measurement of the sidereal rotation. The positions of stars or radio sources used for such measurements are adjusted for precession, nutation, and so on, and the resulting time measurements thus depend only on the angular velocity of the earth and on polar motion. When converted to the solar timescale, these measurements provide a form of

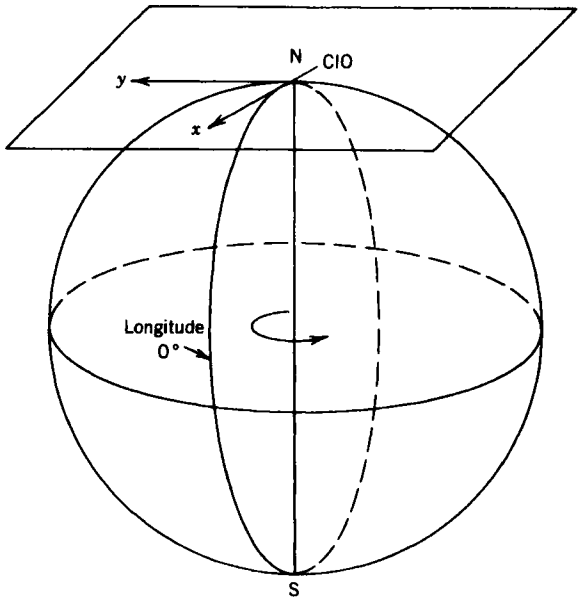


Figure 12.3 Coordinate system for the measurement of polar motion. The x coordinate is in the plane of the Greenwich meridian and the y axis is 90° to the west. CIO is the conventional international origin.

universal time (UT) known as UT0; this is not truly "universal" since the effects of polar motion, which can amount to about 35 ms, depend on the location of the observatory. When UT0 is corrected for polar motion, the result is known as UT1. Since it is a measure of the rotation of the earth relative to fixed celestial objects, UT1 is the form of time required in astronomical observing, including the analysis of interferometric observations, navigation, and surveying. However, UT1 contains the effects of small variations in the earth's rotation rate, attributable largely to geophysical effects such as the seasonal variations in the distribution of water between the surface and atmosphere. Fluctuations in the length of day over the period of a year are typically about 1 ms. To provide a more uniform measure of time, UT2 is derived from UT1 by attempting to remove seasonal variations. UT2 is rarely used. UT1 and UT2 include the effect of the gradual decrease of the rotation rate of the earth. This causes the length of the UT1/UT2 day to increase slightly when compared with International Atomic Time (IAT), which is based on the frequency of the cesium line (see Section 9.5 under *Rubidium and Cesium Standards*). The IAT second is the basis for another form of UT, Coordinated Universal Time (UTC), which is offset from IAT so that $|\mathrm{UT1} - \mathrm{UTC}| < 1$ s. This relationship is maintained by inserting one-second discontinuities (leap seconds) in UTC when required on specified days of the year. Most time services such as Loran C and GPS (see Section 9.5 under *Time Synchronization*) transmit UTC.

The practice at many observatories is to maintain UTC or IAT using an atomic standard and then obtain UT1 from the published values of $\Delta\mathrm{UT1} = \mathrm{UT1} - \mathrm{UTC}$.

Since ΔUT1 is measured rather than computed, in principle it can be determined only after the fact. However, it is possible to predict it by extrapolation with satisfactory accuracy for periods of 1 or 2 weeks, and thus to implement UT1 in real time. Values of ΔUT1 are available from the Bureau International de l'Heure (BIH), which was established in 1912 at the Paris Observatory to coordinate international timekeeping, and from the U.S. Naval Observatory. Rapid service data are available from these institutions with a timeliness suitable for extrapolation.

Measurement of Polar Motion and UT1

The classical optical methods of measuring polar motion and UT1 are by timing the meridian transits of stars of known positions. Observations at different longitudes, using stars at more than one declination, are required to determine all three parameters (x, y, ΔUT1). During the 1970s it became evident that such astrometric tasks can also be performed by radio interferometry (McCarthy and Pilkington 1979).

To specify the baseline components of an interferometer for such measurements, we use the (X, Y, Z) system of Section 4.1, rotated so that the X axis lies in the Greenwich meridian instead of the local meridian. Let ΔX, ΔY, and ΔZ be the changes in the baseline components resulting from polar motion (x, y) (in radians) and a time variation (UT1 − UTC) corresponding to Θ radians. Then we may write

$$\begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} = \begin{bmatrix} 0 & -\Theta & x \\ \Theta & 0 & -y \\ -x & y & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{12.27}$$

where the square matrix is a three-dimensional rotational matrix valid for small angles of rotation. Θ, x and y are the rotation angles about the Z, Y, and X axes, respectively. From Eq. (12.27) we obtain

$$\begin{aligned} \Delta X &= -\Theta Y + xZ, \\ \Delta Y &= \Theta X - yZ, \\ \Delta Z &= -xX + yY. \end{aligned} \tag{12.28}$$

Thus, if one observes a series of sources at periodic intervals and determines the variation in baseline parameters, Eqs. (12.28) can be used to determine UT1 and polar motion. For an interferometer with an east–west baseline ($Z = 0$), one can determine Θ but cannot separate the effects of x and y. An east–west interferometer located on the Greenwich meridian ($X = Z = 0$) would yield measures of Θ and y but not of x. If it had a north–south component of baseline ($Z \neq 0$), one could still measure y but would not be able to separate the effects of x and Θ. In general, one cannot measure all three quantities with a single baseline, since a single direction is specified by two parameters only. Systems suitable for a complete solution might be, for example, two east–west interferometers separated by about 90° in longitude or a three-element non-collinear interferometer.

An example of VLBI measurements of the pole position can be found in Carter, Robertson, and MacKay (1985). Development of the Global Positioning System has also provided a method of making pole-position measurements [see, e.g., Herring (1999)].

The methods just described are applicable to observations using connected-element interferometers in which the phase can be calibrated, and also to VLBI observations in which the bandwidth is sufficient to obtain accurate group delay measurements. If only the fringe frequency is measurable, as in narrow-bandwidth VLBI systems, the result is insensitive to the Z component of the baseline. Then in Eqs. (12.28) one has measurements of ΔX and ΔY only, and in general it is not possible to separate the effects of polar motion and variation of UT1. However, if $Z = 0$ (east–west baseline), UT1 can be derived. A comparison of determinations of UT1 and polar motion by VLBI, satellite laser ranging, and BIH analyses of standard astrometric data is given in Robertson et al. (1983) and Carter et al. (1984).

12.4 GEODETIC MEASUREMENTS

Certain geophysical phenomena, for example, earth tides (Melchior 1978) and movements of tectonic plates, can result in variations in the baseline vector of a VLBI system. Variations in the length of the baseline are clearly attributable to such phenomena, whereas variations in the direction can also result from polar motion and rotational variations. Magnitudes of the effects are of order 1–10 cm year^{-1} for plate motions and 30 cm (diurnal) for earth tides. They are thus measurable using the most accurate techniques of VLBI. Earth tides were first detected by Shapiro et al. (1974), and refined measurements were reported by Herring et al. (1983). In addition to solid-earth tides, displacement of land masses resulting from tidal shifts of water masses, called ocean loading, is believed to be a measurable effect. Early evidence of contemporary motion of tectonic plates was reported by Herring et al. (1986). For reviews of geodetic applications of VLBI, see Shapiro (1976), Counselman (1976), Clark et al. (1985), Carter and Robertson (1993), and Sovers, Fanselow, and Jacobs (1998).

12.5 MAPPING ASTRONOMICAL MASERS

In the envelopes of many newly formed stars, and also those of highly evolved stars, radio emission from molecules such as H_2O and OH is caused by a maser process. The frequency spectrum of the emission is often complicated, containing many spectral features or components caused by clouds of gas moving at different line-of-sight velocities. Maps of strong maser sources reveal hundreds of compact components with brightness temperatures approaching 10^{15} K, angular sizes as small as 10^{-4} arcsec, and flux densities as high as 10^6 Jy. The components are typically distributed over an area of several arcseconds diameter and a Doppler velocity range of 10–300 km s^{-1} (0.7–20 MHz for the H_2O maser

transition at 22 GHz). Individual features have linewidths of about 1 km s^{-1} or less (74 kHz at 22 GHz). The physics and phenomenology of masers are discussed by Reid and Moran (1988) and Elitzur (1992). The processing and analysis of maser data require large correlator systems because the ratio of required bandwidth to spectral resolution is large (10^2–10^4). They also require prodigious amounts of image processing because the ratio of the field of view to the spatial resolution is large (10^2–10^4). As an extreme example, the H_2O maser in W49 has hundreds of features distributed over 3 arcsec (Gwinn, Moran, and Reid 1992). The complete mapping of this source at a resolution of 10^{-3} arcsec with 3 pixels per resolution interval would require the production of 600 maps, each with at least 10^8 pixels. However, most of the map cells would contain no signal. Thus, the usual procedure is to measure the positions of the features crudely by fringe-frequency analysis, and then map small fields around these locations by Fourier synthesis techniques. Examples of maps made by fringe-frequency analysis can be found in Walker, Matsakis, and Garcia-Barreto (1982); by phase analysis in Genzel et al. (1981) and Norris and Booth (1981); and by Fourier synthesis in Reid et al. (1980), Norris, Booth, and Diamond (1982), and Boboltz, Diamond, and Kemball (1997). We shall briefly discuss some of the techniques used in mapping masers and their accuracies. Note that geometric (group) delays cannot be measured accurately because of the narrow bandwidths of the maser lines.

In mapping masers, we must explicitly consider the frequency dependence of the fringe visibility. We assume that a maser source consists of a number of point sources. Furthermore, we assume that the measurements are made with a VLBI system, and that the desired RF band is converted to a single baseband channel. Adapting Eq. (9.23), we can write the residual fringe phase of one maser component at frequency ν as

$$\Delta\phi(\nu) = 2\pi\left[\nu\,\Delta\tau_g(\nu) + (\nu - \nu_{\rm LO})\tau_e + \nu\tau_{\rm at}\right] + \phi_{\rm in} + 2\pi n, \qquad (12.29)$$

where τ_e is the relative delay error due to clock offsets; $\tau_{\rm at}$ is the differential atmospheric delay; $\Delta\tau_g(\nu)$ is the difference between the true geometric delay of the source $\tau_g(\nu)$ and the expected (reference) delay; $\nu_{\rm LO}$ is the local oscillator frequency; $\phi_{\rm in}$ is the instrumental phase, which includes the local oscillator frequency difference and can be a rapidly varying function of time; and $2\pi n$ represents the phase ambiguity. A frequency can usually be found that has only one unresolved maser component, and this component can then be used as a phase reference. The use of a phase reference feature is fundamental to all maser analysis procedures, and it allows maps of the relative positions of maser components to be made with high accuracy. The difference in residual fringe phase between a maser feature at frequency ν and the reference feature at frequency ν_R is

$$\Delta^2\phi(\nu) = \Delta\phi(\nu) - \Delta\phi(\nu_R), \qquad (12.30)$$

which, with the use of Eq. (12.29), becomes

$$\Delta^2\phi(\nu) = 2\pi \left\{\nu\left[\tau_g(\nu) - \tau_g(\nu_R)\right] + (\nu - \nu_R)\left[\tau_g(\nu_R) - \tau_g'(\nu_R)\right] + (\nu - \nu_R)\left[\tau_e + \tau_{\rm at}\right]\right\}, \tag{12.31}$$

where $\tau_g'(\nu_R)$ is the expected delay of the reference feature, and $\tau_g(\nu_R)$ is the true delay. The frequency-independent terms $\phi_{\rm in}$ and $2\pi n$ cancel in Eq. (12.31). However, there are residual terms in Eq. (12.31) that are proportional to the difference in frequency between the feature of interest and the reference feature. These terms arise because phases at different frequencies are differenced in Eq. (12.30). Following the notation of Eq. (12.3), which uses the convention Δ term = (assumed value) − (true value), we can write Eq. (12.31) as

$$\Delta^2\phi(\nu) = \frac{2\pi\nu}{c}\mathbf{D}\cdot\Delta\mathbf{s}_{\nu R} - \frac{2\pi\nu}{c}\Delta\mathbf{D}\cdot\Delta\mathbf{s}_{\nu R} - \frac{2\pi}{c}\left[(\nu - \nu_R)(\Delta\mathbf{D}\cdot\mathbf{s}_R + \mathbf{D}\cdot\Delta\mathbf{s}_R)\right] + 2\pi(\nu - \nu_R)(\tau_e + \tau_{\rm at}), \tag{12.32}$$

where $\mathbf{D}$ is the assumed baseline, $\Delta\mathbf{D}$ is the baseline error, $\mathbf{s}_R$ is the assumed position of the reference feature, and $\Delta\mathbf{s}_R$ is the corresponding position error. $\Delta\mathbf{s}_{\nu R}$ is the separation vector from the feature at frequency ν to the reference feature, and thus the true position of the feature at frequency ν is $\mathbf{s}_R - \Delta\mathbf{s}_R + \Delta\mathbf{s}_{\nu R}$.

The first term on the right-hand side of Eq. (12.32) is the desired quantity from which the position of the feature relative to the reference feature can be determined, and the remaining terms describe the phase errors introduced by uncertainty in baseline, source position, clock offset, and atmospheric delay. These phase error terms can be converted approximately to angular errors by dividing them by $c/2\pi\nu D$. Thus, for example, an error of 0.3 m in a baseline component would cause a delay error of about 1 ns in the term $\Delta\mathbf{D}\cdot\mathbf{s}_R$ in Eq. (12.32) and a phase error of 10^{-3} turns for features separated by 1 MHz. This phase error corresponds to a nominal error of 10^{-6} arcsec on a baseline of 2500 km at 22 GHz, which provides a fringe spacing of 10^{-3} arcsec. Similarly, a clock or atmospheric error of 1 ns would cause the same positional error. The same baseline error also causes additional positional errors, through the $\Delta\mathbf{D}\cdot\Delta\mathbf{s}_{\nu R}$ term, of 10^{-7} arcsec per arcsecond separation of the features. A detailed discussion of mapping errors caused by this calibration method can be found in Genzel et al. (1981).

Another method of calibrating the fringe phase is to scale the phase of the reference feature to the frequency of the feature to be calibrated. That is,

$$\Delta^2\phi(\nu) = \Delta\phi(\nu) - \Delta\phi(\nu_R)\frac{\nu}{\nu_R}. \tag{12.33}$$

This method of calibration is more accurate than the method of Eq. (12.30) because error terms proportional to $\nu - \nu_R$ do not appear. However, there are additional terms involving the phase ambiguity and the instrumental phase. Thus, this

calibration method is applicable only if the fringe phase can be followed carefully enough to avoid the introduction of phase ambiguities.

Maps of lower accuracy and sensitivity than those obtainable from phase data can be made with fringe-frequency data. Suppose that the interferometer is well calibrated. The differential fringe frequency, that is, the difference in fringe frequency between the feature at frequency ν and the reference feature, can then be written [using Eq. (12.14)]

$$\Delta^2 \nu_f(\nu) \simeq \dot{u}\Delta\alpha'(\nu) + \dot{v}\,\Delta\delta(\nu), \tag{12.34}$$

where $\dot{u}$ and $\dot{v}$ are the time derivatives of the projected baseline components, $\Delta\alpha'(\nu)$ and $\Delta\delta(\nu)$ are the coordinate offsets from the reference feature, and $\Delta\alpha'(\nu) = \Delta\alpha(\nu)\cos\delta$. The relative positions of the maser feature can then be found by fitting Eq. (12.34) to a series of fringe-frequency measurements at various hour angles. This technique was first employed by Moran et al. (1968) for the mapping of an OH maser. The errors in fringe-frequency measurements decrease as $\tau^{3/2}$ [see Eq. (A12.27)], where τ is the length of an observation, but for large values of τ the differential fringe frequency $\Delta^2\nu_f$ is not constant, because $\ddot{u}$ and $\ddot{v}$ are not zero. Thus, there is a limited field of view available for accurate mapping with fringe-frequency measurements. This field of view can be estimated by equating the rms fringe-frequency error in Eq. (A12.27) with τ times the derivative of the differential fringe frequency with respect to time. Therefore, for an east–west baseline,

$$D_\lambda \omega_e^2 \Delta\theta\tau\cos\theta \simeq \sqrt{\frac{3}{2\pi^2}}\left(\frac{T_S}{T_A}\right)\frac{1}{\sqrt{\Delta\nu\tau^3}}, \tag{12.35}$$

where $\Delta\theta$ is the field of view. For $\sqrt{2\pi^2/3}\,\cos\theta \simeq 1$, the field of view is

$$\Delta\theta \simeq \frac{T_S}{D_\lambda T_A \omega_e^2 \tau^2 \sqrt{\Delta\nu\tau}}, \tag{12.36}$$

or

$$\Delta\theta \simeq \frac{1}{\mathcal{R}_{\rm sn} D_\lambda \omega_e^2 \tau^2}, \tag{12.37}$$

where $\mathcal{R}_{\rm sn}$ is the signal-to-noise ratio. Let $\mathcal{R}_{\rm sn} = 10$ and $\tau = 100$ s. The field of view is then about equal to 2000 times the fringe spacing. This restriction is often important. Usually when a feature is found, the phase center of the field is moved to the estimated position of the feature, and the position is then redetermined. Only components that are detected in individual observations on each baseline can be mapped with the fringe-frequency mapping technique. Thus, fringe-frequency mapping is less sensitive than synthesis mapping, in which fully coherent sensitivity is achieved.

The fringe-frequency analysis procedure can be extended to handle the case in which there are many point components in one frequency channel. From each observation (i.e., a measurement on one baseline lasting for a few minutes), the

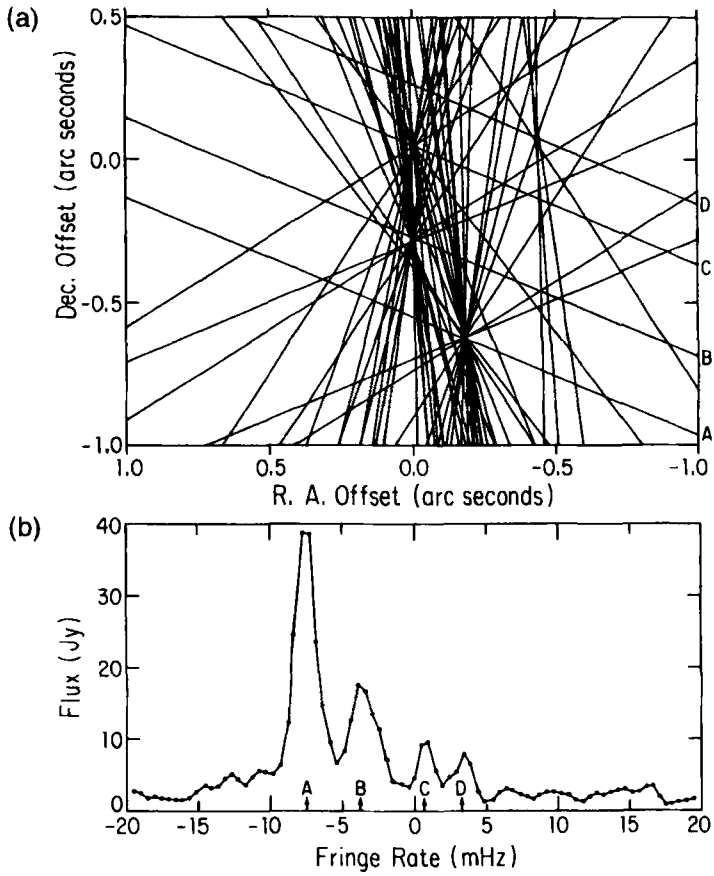


Figure 12.4 Plot (b) is the fringe-frequency spectrum of the water vapor maser in W49N, at one particular hour angle and one frequency in the radio spectrum of the maser. The ordinate is flux density. There are four peaks, each corresponding to a separate feature on the sky. Plot (a) shows such lines from many scans. The peaks in the lower plot and their corresponding lines in the upper plot are labeled A–D. There are at least four separate features at the frequency of these data. Their positions are marked by the locations where many lines intersect. The feature corresponding to line D is sufficiently far from the phase center that its fringe frequency changes enough during the 20-min integration to degrade significantly the estimate of the feature position. The window in which accurate positions can be determined is 0.5 arcsec in right ascension and 2 arcsec in declination. The window can be moved by shifting the phase center of the data. Figure from Walker (1981), courtesy of *The Astronomical Journal.*

fringe-frequency spectrum is calculated. Multiple components will appear as distinct fringe-frequency features, as shown in Fig. 12.4. The fringe frequency of each feature defines a line in $(\Delta\alpha', \Delta\delta)$ space on which a maser component lies. The slope of the line is $\tan^{-1}(\dot{v}/\dot{u})$. As the projected baseline changes, the slopes of the lines change. The intersections of the lines define the source positions (see

Fig. 12.4). For this method to work, the components must be sufficiently separate to produce separate peaks in the fringe-frequency spectrum. The fringe-frequency resolution is about τ^{-1}, which defines an effective beam of width

$$\Delta\theta_f = \frac{1}{D_\lambda \omega_e \tau \cos\theta}. \tag{12.38}$$

Fringe-frequency mapping is discussed in detail, for example, by Walker (1981).

APPENDIX 12.1 LEAST-MEAN-SQUARES ANALYSIS

The principles of least-mean-squares analysis play a fundamental role in astrometry, where the goal is to extract a number of parameters from a set of noisy measurements. We briefly discuss these principles in an elementary way, ignoring mathematical subtleties, and apply them to the problems encountered in interferometry. Detailed discussions of the statistical analysis of data can be found in books such as Bevington and Robinson (1992) and Hamilton (1964). Suppose that we wish to measure a quantity m. We make a set of measurements y_i that are the sum of the desired quantity m and a noise contribution n_i:

$$y_i = m + n_i, \tag{A12.1}$$

where n_i is a Gaussian random variable with zero mean and variance σ_i^2. The probability that the ith measurement will take any specific value of y_i is given by the probability (density) function

$$p(y_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} e^{-(y_i - m)^2/2\sigma_i^2}. \tag{A12.2}$$

If all the measurements are independent, then the probability that an experiment will yield a set of N measurements $y_1, y_2, \ldots, y_N$ is

$$L = \prod_{i=1}^{N} p(y_i), \tag{A12.3}$$

where the $\prod$ denotes the product of the $p(y_i)$ terms. L, viewed as a function of m, is called the likelihood function. The method of maximum likelihood is based on the assumption that the best estimate of m is the one that maximizes L. Maximizing L is the same as maximizing $\ln L$, where

$$\ln L = \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\,\sigma_i} - \frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - m)^2}{\sigma_i^2}. \tag{A12.4}$$

Since the first summation term on the right-hand side of Eq. (A12.4) is a constant and the second summation term is multiplied by $-\frac{1}{2}$, the maximization of L is

equivalent to the minimization of the second summation term in Eq. (A12.4) with respect to m. Thus, we wish to minimize the quantity χ^2 given by

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - m)^2}{\sigma_i^2}. \tag{A12.5}$$

In the more general problem discussed later in this appendix, m is replaced by a function with one or more parameters describing the system model. With this generalization, Eq. (A12.5) becomes the fundamental equation of the method of weighted least squares. In this method the parameters of the model are determined by minimizing the sum of the squared differences between the measurements and the model, weighted by the variances of the measurements. The quantity χ^2, which indicates the goodness of fit, is a random variable whose mean value equals the number of data points less the number of parameters when the model adequately describes the measurements. The method of least squares, appropriate when the noise is a Gaussian random process, is a special case of the more general method of maximum likelihood. Gauss invented the method of least squares, perhaps as early as 1795, using arguments similar to those given here, for the purpose of estimating the orbital parameters of planets and comets (Gauss 1809). The method was independently developed by Legendre in 1806 (Hall 1970).

Returning to Eq. (A12.5) we can estimate m by setting the derivative of χ^2 with respect to m equal to zero. The resulting estimate of m, denoted m_e, is

$$m_e = \frac{\sum \frac{y_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}, \tag{A12.6}$$

where the summation goes from $i = 1$ to N. Using Eq. (A12.2), we note that $\langle y_i \rangle = m$ and $\langle y_i^2 \rangle = m^2 + \sigma_i^2$. Therefore, by calculating the expectation of Eq. (A12.6), it is clear that $\langle m_e \rangle = \langle y_i \rangle = m$, and it is easy to show that

$$\langle m_e^2 \rangle = m^2 + \left(\sum \frac{1}{\sigma_i^2} \right)^{-1}. \tag{A12.7}$$

Hence the variance of the estimate of m_e is

$$\sigma_m^2 = \langle m_e^2 \rangle - \langle m_e \rangle^2 = \left(\sum \frac{1}{\sigma_i^2} \right)^{-1}. \tag{A12.8}$$

Equation (A12.8) shows that when poor quality or noisy data are added to better data, the value of σ_m may be reduced only slightly. If the statistical error σ_i of each of the measurements has the same value, σ, then Eq. (A12.8) reduces to the

well-known result

$$\sigma_m = \frac{\sigma}{\sqrt{N}}, \tag{A12.9}$$

and m_e is the average of the measurements. In many instances σ is not known. An estimate of σ is

$$\sigma_e^2 = \frac{1}{N}\sum(y_i - m)^2. \tag{A12.10}$$

However, m is not known, only its estimate, m_e. If m_e were used in place of m in Eq. (A12.10), the value of σ_e^2 would be an underestimate of σ^2 because of the manner in which m_e was determined in minimizing χ^2. The unbiased estimate of σ^2 is

$$\sigma_e^2 = \frac{1}{N-1}\sum(y_i - m_e)^2. \tag{A12.11}$$

It is easy to show by substitution of Eq. (A12.6) into Eq. (A12.11) that $\langle\sigma_e^2\rangle = \sigma^2$. The term $N - 1$, which is called the number of degrees of freedom, appears in Eq. (A12.11) because there are N data points and one free parameter.

Consider a model described by the function $f(x; p_1, \ldots, p_n)$, where x is the independent variable, which takes values x_i, where $i = 1$ to N, at the sample points, and $p_1, \ldots, p_n$ are a set of parameters. We assume that the values of the independent variable are exactly known. If the function f correctly models the measurement system, the measurement set is given by

$$y_i = f(x_i; p_1, \ldots, p_n) + n_i, \tag{A12.12}$$

where n_i represents the measurement error. The general problem is to find the values of the parameters for which χ^2, given by the generalization of Eq. (A12.5),

$$\chi^2 = \sum \frac{[y_i - f(x_i)]^2}{\sigma_i^2}, \tag{A12.13}$$

is a minimum.

A simple example of this problem is the fitting of a straight line to a data set. Let

$$f(x; a, b) = a + bx, \tag{A12.14}$$

where a and b are the parameters to be found. Minimizing χ^2 is accomplished by solving the equations

$$\frac{\partial\chi^2}{\partial a} = -\sum \frac{2(y_i - a - bx_i)}{\sigma_i^2} = 0, \tag{A12.15a}$$

and

$$\frac{\partial \chi^2}{\partial b} = -\sum \frac{2(y_i - a - bx_i)x_i}{\sigma_i^2} = 0. \tag{A12.15b}$$

In matrix notation we have

$$\begin{bmatrix} \sum \frac{y_i}{\sigma_i^2} \\ \sum \frac{x_i y_i}{\sigma_i^2} \end{bmatrix} = \begin{bmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{bmatrix} \begin{bmatrix} a_e \\ b_e \end{bmatrix}, \tag{A12.16}$$

where we distinguish between the true values of the parameters and their estimates by the subscript e. The solution is

$$a_e = \frac{1}{\Delta}\left[\left(\sum \frac{x_i^2}{\sigma_i^2}\right)\left(\sum \frac{y_i}{\sigma_i^2}\right) - \left(\sum \frac{x_i}{\sigma_i^2}\right)\left(\sum \frac{x_i y_i}{\sigma_i^2}\right)\right] \tag{A12.17}$$

and

$$b_e = \frac{1}{\Delta}\left[\left(\sum \frac{1}{\sigma_i^2}\right)\left(\sum \frac{x_i y_i}{\sigma_i^2}\right) - \left(\sum \frac{x_i}{\sigma_i^2}\right)\left(\sum \frac{y_i}{\sigma_i^2}\right)\right], \tag{A12.18}$$

where Δ is the determinant of the square matrix in Eq. (A12.16), given by

$$\Delta = \left(\sum \frac{1}{\sigma_i^2}\right)\left(\sum \frac{x_i^2}{\sigma_i^2}\right) - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2. \tag{A12.19}$$

Estimates of the errors in the parameters a_e and b_e can be calculated from Eqs. (A12.17) and (A12.18) and are given by

$$\sigma_a^2 = \langle a_e^2 \rangle - \langle a_e \rangle^2 = \frac{1}{\Delta} \sum \frac{x_i^2}{\sigma_i^2} \tag{A12.20}$$

and

$$\sigma_b^2 = \langle b_e^2 \rangle - \langle b_e \rangle^2 = \frac{1}{\Delta} \sum \frac{1}{\sigma_i^2}. \tag{A12.21}$$

Note that a_e and b_e are random variables, and in general $\langle a_e b_e \rangle$ is not zero, so that the parameter estimates are correlated. The error estimates in Eqs. (A12.20) and (A12.21) include the deleterious effects of the correlation between parameters. In this particular example, the correlation can be made equal to zero by adjusting the origin of the x axis so that $\sum (x_i/\sigma_i^2) = 0$.

The above analysis can be used to estimate the accuracy of measurements of fringe frequency and delay made with an interferometer. Fringe frequency, the rate of change of fringe phase with time,

$$\nu_f = \frac{1}{2\pi}\frac{\partial\phi}{\partial t}, \tag{A12.22}$$

can be estimated by fitting a straight line to a sequence of uniformly spaced measurements of phase with respect to time. The fringe frequency is proportional to the slope of this line. Assume that N measurements of phase ϕ_i, each having the same rms error σ_ϕ, are made at times t_i, spaced by interval T, running from time $-NT/2$ to $NT/2$, such that the total time of the observation is $\tau = NT$. From Eq. (A12.21) and the above definitions, including Eq. (A12.22), the error in the fringe-frequency estimate is

$$\sigma_f^2 = \frac{\sigma_\phi^2}{(2\pi)^2 \sum t_i^2}, \tag{A12.23}$$

since $\sum t_i = 0$. The term $\sum t_i^2$ is approximately given by

$$\sum t_i^2 \simeq \frac{1}{T}\int_{-\tau/2}^{\tau/2} t^2\,dt = \frac{1}{T}\frac{\tau^3}{12} = \frac{N\tau^2}{12}. \tag{A12.24}$$

$\tau/\sqrt{12}$ can be thought of as the rms time span of the data. Thus, Eq. (A12.23) becomes

$$\sigma_f^2 = \frac{12\sigma_\phi^2}{(2\pi)^2 N\tau^2}. \tag{A12.25}$$

The expression for σ_ϕ, given in Eq. (6.64) for the case when the source is unresolved and there are no processing losses, is

$$\sigma_\phi = \frac{T_S}{T_A\sqrt{2\,\Delta\nu T}}, \tag{A12.26}$$

where T_S is the system temperature, T_A is the antenna temperature due to the source, and $\Delta\nu$ is the bandwidth. Substitution of Eq. (A12.26) into Eq. (A12.25) yields

$$\sigma_f = \sqrt{\frac{3}{2\pi^2}}\left(\frac{T_S}{T_A}\right)\frac{1}{\sqrt{\Delta\nu\tau^3}}\ \text{(Hz)}. \tag{A12.27}$$

Note that this result does not depend on the details of the analysis procedure, such as the choice of N. Equivalently, one can estimate the fringe frequency by finding the peak of the fringe-frequency spectrum, that is, the peak of the Fourier transform of $e^{j\phi_i}$.

The delay is the rate of change of phase with frequency,

$$\tau = \frac{1}{2\pi}\frac{\partial\phi}{\partial\nu}. \tag{A12.28}$$

Thus, the delay can be estimated by finding the slope of a straight line fitted to a sequence of phase measurements as a function of frequency. For a single band, such data can be obtained from the cross power spectrum, the Fourier transform of the cross-correlation function. Assume that N measurements of phase are made at frequencies ν_i, each with a bandwidth $\Delta\nu/N$ and with an error σ_ϕ. In this calculation only the relative frequencies are important. It is convenient for the purpose of analysis to set the zero of the frequency axis such that $\sum \nu_i = 0$. The error in delay [from Eqs. (A12.19), (A12.21), and (A12.28)] is

$$\sigma_\tau^2 = \frac{\sigma_\phi^2}{(2\pi)^2 \sum \nu_i^2}. \tag{A12.29}$$

Using a calculation for $\sum \nu_i^2$ analogous to the one in Eq. (A12.24), we can write Eq. (A12.29) as

$$\sigma_\tau^2 = \frac{12\sigma_\phi^2}{(2\pi)^2 N\,\Delta\nu^2}. \tag{A12.30}$$

Thus, substitution of Eq. (A12.26) (with an integration time of τ and bandwidth $\Delta\nu/N$) into Eq. (A12.30) yields

$$\sigma_\tau = \sqrt{\frac{3}{2\pi^2}}\left(\frac{T_S}{T_A}\right)\frac{1}{\sqrt{\Delta\nu^3\tau}}. \tag{A12.31}$$

We can define the rms bandwidth as

$$\Delta\nu_{\text{rms}} = \sqrt{\frac{1}{N}\sum \nu_i^2} \tag{A12.32}$$

and obtain from Eqs. (A12.26) and (A12.29) the result quoted in Section 9.8 [Eq. (9.159)],

$$\sigma_\tau = \frac{1}{\zeta}\left(\frac{T_S}{T_A}\right)\frac{1}{\sqrt{\Delta\nu_{\text{rms}}^3\tau}}, \tag{A12.33}$$

where $\zeta = \pi(768)^{1/4}$. (Note that in Section 9.8, σ_ϕ applies to the full bandwidth $\Delta\nu$.) The expressions for σ_τ in Eqs. (A12.30), (A12.31), and (A12.33) incorporate the condition $\Delta\nu_{\text{rms}} = \Delta\nu/\sqrt{12}$ and apply to a continuous passband of width $\Delta\nu$.

In bandwidth synthesis, which is described in Section 9.8, the measurement system consists of N channels of width $\Delta\nu/N$, which are not in general contigu-

ous. The rms delay error is obtained by substituting Eqs. (A12.26) and (A12.32) into Eq. (A12.29), yielding

$$\sigma_\tau = \frac{1}{\sqrt{8\pi^2}} \left(\frac{T_S}{T_A}\right) \frac{1}{\sqrt{\Delta\nu\tau}\,\Delta\nu_{\text{rms}}}, \tag{A12.34}$$

where $\Delta\nu_{\text{rms}}$ is given by Eq. (A12.32) and $\Delta\nu$ is the total bandwidth. $\Delta\nu_{\text{rms}}$ is generally equal to about 40% of the total frequency range spanned.

A general formulation of the linear least-squares solution can be found when the model function f is a linear function of the parameters p_k, that is, when

$$f(x; p_1, \ldots, p_n) = \sum_{k=1}^{n} \frac{\partial f}{\partial p_k} p_k, \tag{A12.35}$$

where n is the number of parameters. For example, the model could be a cubic polynomial

$$f(x; p_0, p_1, p_2, p_3) = p_0 + p_1 x + p_2 x^2 + p_3 x^3, \tag{A12.36}$$

in which case $\partial f/\partial p_k = x^k$ for $k = 0, 1, 2$, and 3. If the parameters appear as linear multiplicative factors, then the minimization of Eq. (A12.13) leads to a set of n equations of the form

$$\frac{\partial \chi^2}{\partial p_k} = 0, \qquad k = 1, 2, \ldots, n. \tag{A12.37}$$

Substitution of Eq. (A12.13) into Eq. (A12.37) and use of Eq. (A12.35) yield the set of n equations

$$D_k = \sum_{j=1}^{n} T_{kj} p_j, \qquad k = 1, 2, \ldots, n, \tag{A12.38}$$

where

$$D_k = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \frac{\partial f(x_i)}{\partial p_k} \tag{A12.39}$$

and

$$T_{jk} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial f(x_i)}{\partial p_j} \frac{\partial f(x_i)}{\partial p_k}, \tag{A12.40}$$

and the summations are carried out over the set of N independent measurements. In matrix notation, the equation set (A12.38) is

$$[D] = [T][P_e], \tag{A12.41}$$

where $[D]$ is a column matrix with elements D_k, $[P_e]$ is a column matrix containing the estimates of the parameters p_{ek}, and $[T]$ is a symmetric square matrix with elements T_{jk}. For obvious reasons, $[T]$ is sometimes called the matrix of the normal equations. Note that Eq. (A12.41) is a generalization of Eq. (A12.16). The matrices $[T]$ and $[D]$ are sometimes written as the product of other matrices (Hamilton 1964, Ch. 4). Let $[M]$ be the variance matrix (size $N \times N$) whose diagonal elements are σ_i^2 and whose off-diagonal elements are zero; let $[F]$ be a column matrix containing the data y_i; and let $[A]$ be the partial derivative matrix (size $n \times N$) whose elements are $\partial f(x_i)/\partial p_k$. Then one can write $[T] = [A]^T[M]^{-1}[A]$ and $[D] = [A]^T[M]^{-1}[F]$, where $[A]^T$ is the transpose of $[A]$ and $[M]^{-1}$ is the inverse of $[M]$. The analysis can be generalized to include the situation where the errors between measurements are correlated. In this case, $[M]$ is modified to include off-diagonal elements $\sigma_i\sigma_j\rho'_{ij}$, where ρ'_{ij} is the correlation coefficient for the ith and jth measurements.

The solution to Eq. (A12.41) is

$$[P_e] = [T]^{-1}[D], \tag{A12.42}$$

where $[T]^{-1}$ is the inverse matrix of $[T]$, and $[P_e]$ is the column matrix containing the parameter estimates. The elements of $[T]^{-1}$ are denoted T'_{jk}. It can be shown by direct calculation that the estimates of the errors of the parameters σ_{ek}^2 are the diagonal elements of $[T]^{-1}$, which is called the covariance matrix. Thus

$$\sigma_{ek}^2 = T'_{kk}. \tag{A12.43}$$

The probability that parameter p_k will be within $\pm\sigma_k$ of its true value is 0.68, which is the integral under the one-dimensional Gaussian probability distribution between $\pm\sigma_k$. The probability that all the n parameters will be within $\pm\sigma$ of their true values (i.e., within the error "box" in the n-dimensional space) is approximately 0.68^n when the correlations are moderate.

The normalized correlation coefficients between parameters are proportional to the off-diagonal elements of $[T]^{-1}$:

$$\rho_{jk} = \frac{\langle (p_{ej} - p_j)(p_{ek} - p_k)\rangle}{\sigma_{ek}\sigma_{ej}} = \frac{T'_{jk}}{\sqrt{T'_{jj}T'_{kk}}}. \tag{A12.44}$$

For any two parameters, there is a bivariate Gaussian probability distribution that describes the distribution of errors

$$p(\epsilon_j, \epsilon_k) = \frac{1}{2\pi\sigma_j\sigma_k\sqrt{1-\rho_{jk}^2}} \exp\left\{-\frac{1}{2(1-\rho_{jk}^2)}\left[\frac{\epsilon_j^2}{\sigma_j^2} + \frac{\epsilon_k^2}{\sigma_k^2} - \frac{2\rho_{jk}\epsilon_j\epsilon_k}{\sigma_j\sigma_k}\right]\right\}, \tag{A12.45}$$

where $\epsilon_k = p_{ek} - p_k$ and $\epsilon_j = p_{ej} - p_j$. The contour of $p(\epsilon_k, \epsilon_j) = p(0,0)e^{-1/2}$ defines an ellipse, shown in Fig. A12.1, which is known as the error ellipse. The

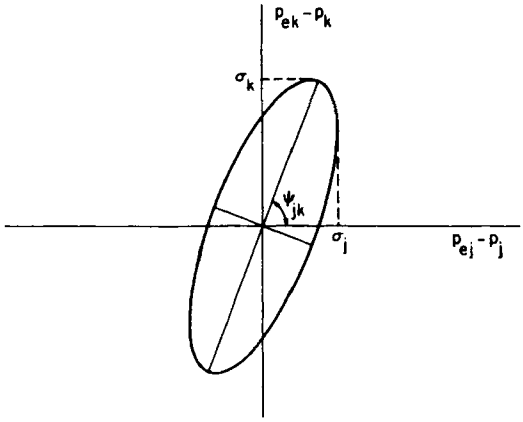


Figure A12.1 The error ellipse, or contour, defining the e^{-1} level of the joint probability function [Eq. (A12.45)] for the estimates of parameters p_k and p_j. The quantities $p_{ek} - p_k$ and $p_{ej} - p_j$ are the parameter estimates minus their true values. The angle ψ_{jk} is defined by Eq. (A12.46).

probability that both parameters will lie within the error ellipse is the integral of Eq. (A12.45) over the area of the error ellipse, which equals 0.46. The orientation of the error ellipse is given by

$$\psi_{jk} = \frac{1}{2}\tan^{-1}\left(\frac{2\rho_{jk}\sigma_j\sigma_k}{\sigma_j^2 - \sigma_k^2}\right). \tag{A12.46}$$

The errors in the parameters p_k are completely determined by the matrix $[T]^{-1}$ through Eqs. (A12.43)–(A12.45). The elements of $[T]^{-1}$ depend only on the partial derivatives of the model function and the values of the measurement errors, which can usually be predicted in advance from the characteristics of the measurement apparatus. Therefore, once an experiment is planned, the errors in the parameters can be predicted from $[T]^{-1}$ without reference to the data. For this reason, $[T]$ is sometimes called the design matrix. Studies of the design matrix for a specific experiment might reveal a very high correlation between two parameters, leading to large errors in their estimated values. It is often possible to modify the experiment to obtain more data that will reduce the correlation. After the data are analyzed, the value of χ^2 can be computed. If the model is a good fit to the data, χ^2 should be approximately equal to $N - n$, the number of measurements minus the number of parameters. If it is not, the difficulty is often that the values of σ_i are estimated incorrectly or that the model does not describe adequately the measurement system, that is, the model has too few parameters or is not correct. Even if $\chi^2 \simeq N - n$, the derived errors in Eq. (A12.43) may not be realistic, and they are referred to as "formal errors." The formal errors describe the *precision* of the parameter estimates. The *accuracy* of the parameter measurements is the deviation between the estimates of the parameters and the true values of the pa-

rameters. The accuracy of the measurements is often difficult to determine. For example, an unknown effect that closely mimics the functional dependence of one of the model parameters may be present in an experiment. The model may appear to be a good one, but the accuracy of the particular model parameter in question will be much poorer than expected because of the systematic error introduced by the unmodeled effect.

The discussion of linear least-mean-squares analysis can be generalized to include nonlinear functions in a straightforward manner. Assume that $f(x;p)$ has one nonlinear parameter p_n. For the purpose of discussion we can separate f into linear and nonlinear parts, $f_L(x;p_1,\ldots,p_{n-1})$ and $f_{NL}(x;p_n)$, and approximate the nonlinear function by the first two terms in a Taylor expansion

$$f_{NL}(x;p_n) \simeq f_{NL}(x,p_{0n}) + \frac{\partial f_{NL}}{\partial p_n}\Delta p_n\,, \tag{A12.47}$$

where p_{0n} is the initial guess of parameter p_n and $\Delta p_n = p_n - p_{0n}$. We assume that the initial parameter guesses are accurate enough for Eq. (A12.47) to be valid. We replace the data with $y_i - f_{NL}(x_i;p_{0n})$ and then compute the elements of the matrices $[D]$ and $[T]$ from the partial derivatives, including $\partial f_{NL}/\partial p_n$. The nth parameter in the matrix $[P_e]$ in Eq. (A12.42) will be the differential parameter Δp_n defined in Eq. (A12.47). The solution must be iterated with a new Taylor expansion centered on the parameter $p_{0n} + \Delta p_n$. Thus, nonlinear functions can be accommodated in the analysis through linearization, but initial guesses of the nonlinear parameters and solution iteration are required. In some cases nonlinear estimation problems can cause difficulties [see, e.g., Lampton, Margon, and Bowyer (1976), Press et al. (1992)].

We can envision how the principles of least-mean-squares analysis are applied to a large astrometric experiment. Consider a hypothetical VLBI experiment made on a three-station array. Suppose that 10 recordings are made of each of 20 sources during observations made over 1 day (an epoch). The observations are repeated 6 times a year for 5 years. The data set would consist of 18,000 measurements (20 sources $\times$10 observations $\times$3 baselines $\times$30 epochs) of delay and fringe frequency, or 36,000 total measurements. The measurements of delay and fringe frequency can be combined in the analysis since, in the least-square method, the relevant quantities are the squares of the measurements divided by their variances, which are dimensionless as in Eq. (A12.13). Now we can count the number of parameters in the analysis model: 39 source coordinates (one right ascension fixed), 9 station coordinates, 90 atmospheric parameters (a zenith excess path length at each station at each epoch), 120 clock parameters (a clock error and clock rate error at two of the stations per epoch), and 90 polar motion and UT1 − UTC parameters, as well as several other parameters to model precession, nutation, solid-earth tides, gravitational deflection by the sun, movement of stations, and other effects such as antenna axis offsets (see Section 4.6). The total number of parameters is about 360. The parameters within each observation epoch are linked because of the common clock and atmosphere parameters. Parameters among epochs are linked because of baseline, precession, and nutation parameters. Naturally, partial solutions from subsets of the data should be

obtained before a grand global solution is attempted. Procedures are available for obtaining global solutions that do not require the inversion of matrices as large as the total number of parameters [see, e.g., Morrison (1969)]. Experiments of the scale described here, and larger ones, have been carried out [e.g., Fanselow et al. (1984), Herring, Gwinn, and Shapiro (1985), Ma et al. (1998)].

One final topic concerns the estimation of the coordinates of a radio source with a well-calibrated interferometer, which has accurately known baselines and instrumental phases. In this case, the differential interferometer phase is, from Eq. (12.1),

$$\begin{aligned}\Delta\phi = 2\pi D_\lambda \{&[\sin d \cos\delta - \cos d \sin\delta \cos(H-h)]\ \Delta\delta \\ &+ \cos d \cos\delta \sin(H-h)\ \Delta\alpha\}\,.\end{aligned} \tag{A12.48}$$

Expressing the geometric quantities in terms of projected baseline components, we can write Eq. (A12.48) as

$$\Delta\phi = 2\pi(u\ \Delta\alpha' + v\ \Delta\delta), \tag{A12.49}$$

where $\Delta\alpha' = \Delta\alpha \cos\delta$. A set of phase measurements from one or more baselines can be analyzed by the method of least squares to determine $\Delta\alpha'$ and $\Delta\delta$. The partial derivatives are $\partial f/\partial p_1 = 2\pi u$ and $\partial f/\partial p_2 = 2\pi v$, where $p_1 = \Delta\alpha'$ and $p_2 = \Delta\delta$. From Eqs. (A12.40) and (A12.49), the normal-equation matrix is

$$[T] = \frac{4\pi^2}{\sigma_\phi^2}\begin{bmatrix}\sum u_i^2 & \sum u_i v_i \\ \sum u_i v_i & \sum v_i^2\end{bmatrix}, \tag{A12.50}$$

where all the measurements are assumed to have the same uncertainty σ_ϕ given by Eq. (A12.26). The inverse of $[T]$ is

$$[T]^{-1} = \frac{1}{\Delta}\begin{bmatrix}\sum v_i^2 & -\sum u_i v_i \\ -\sum u_i v_i & \sum u_i^2\end{bmatrix}, \tag{A12.51}$$

where Δ is the determinant of the matrix in Eq. (A12.50),

$$\Delta = \frac{4\pi^2}{\sigma_\phi^2}\left[\sum u_i^2 \sum v_i^2 - \left(\sum u_i v_i\right)^2\right]. \tag{A12.52}$$

The correlation coefficient defined by Eq. (A12.44) is

$$\rho_{12} = \frac{-\sum u_i v_i}{\sqrt{\sum u_i^2 \sum v_i^2}}\,. \tag{A12.53}$$

The variances of the estimates of the parameters are given by the diagonal elements of Eq. (A12.51),

$$\sigma_{\alpha'}^2 = \frac{\sigma_\phi^2 \sum v_i^2}{4\pi^2 \left[\sum v_i^2 \sum u_i^2 - \left(\sum u_i v_i\right)^2\right]}, \tag{A12.54}$$

and

$$\sigma_\delta^2 = \frac{\sigma_\phi^2 \sum u_i^2}{4\pi^2 \left[\sum v_i^2 \sum u_i^2 - \left(\sum u_i v_i\right)^2\right]}. \tag{A12.55}$$

If the (u, v) loci are long (that is, the observations extend over a large fraction of the day), then $\sum u_i v_i$ will be small compared to $\sum u_i^2$ and $\sum v_i^2$ so that

$$\sigma_{\alpha'} \simeq \frac{\sigma_\phi}{2\pi\sqrt{\sum u_i^2}}, \tag{A12.56}$$

and

$$\sigma_\delta \simeq \frac{\sigma_\phi}{2\pi\sqrt{\sum v_i^2}}. \tag{A12.57}$$

Furthermore, if only one baseline is used on a high-declination source, then $u_i \simeq v_i \simeq D_\lambda$ and both errors reduce to the intuitive result

$$\sigma_{\alpha'} \simeq \sigma_\delta \simeq \frac{\sigma_\phi}{2\pi\sqrt{N}D_\lambda}. \tag{A12.58}$$

Alternatively, the source position can be found by Fourier transformation of the visibility data. This procedure can be thought of as mapping or as multiplying the visibility data by the exponential factors $\exp[2\pi(u_i\,\Delta\alpha' + v_i\,\Delta\delta)]$ and summing over the data. The resulting "function" is maximized with respect to $\Delta\alpha'$ and $\Delta\delta$. In this latter view, it is easy to understand that (basic) mapping (i.e., no tapering or gridding of the data) is a maximum likelihood procedure for finding the position of a point source and therefore formally equivalent to the method of least squares. The synthesized beam b_0 for N measurements is

$$b_0(\Delta\alpha', \Delta\delta) = \frac{1}{N}\sum \cos\left[2\pi(u_i\,\Delta\alpha' + v_i\,\Delta\delta)\right]. \tag{A12.59}$$

The shape of b_0 near its peak can be found by expanding Eq. (A12.59) to second order:

$$b_0(\Delta\alpha', \Delta\delta) \simeq 1 - \frac{2\pi^2}{N}\left(\Delta\alpha'^2 \sum u_i^2 + \Delta\delta^2 \sum v_i^2 - 2\,\Delta\alpha'\,\Delta\delta \sum u_i v_i\right). \tag{A12.60}$$

From Eq. (A12.60) it is easy to see that the contours of the synthesized beam are proportional to the error ellipse defined by Eqs. (A12.45), (A12.46), and (A12.53)–(A12.55). Note that the method of least squares can be applied only in the regime of high signal-to-noise ratio, where phase ambiguities can be resolved. However, the Fourier synthesis method can be applied in any case.

BIBLIOGRAPHY

Calame, O., Ed., *High-Precision Earth Rotation and Earth-Moon Dynamics*, Reidel, Dordrecht, 1982.

Davis, R. J., and R. S. Booth, Eds., *Sub-arcsecond Radio Astronomy*, Cambridge Univ. Press, Cambridge, UK, 1993.

Enge, P. and P. Misra, Eds., *Proc. IEEE*, Special Issue on Global Positioning System, **87**, No. 1, Jan. 1999.

Jespersen, J. and D. W. Hanson, Eds., *Proc. IEEE*, Special Issue on Time and Frequency, **79**, No. 7, July 1991.

Johnston, K. J. and C. de Vegt, Reference Frames in Astronomy, *Ann. Rev. Astron. Astrophys.*, **37**, 97–125, 1999.

McCarthy, D. D. and J. D. H. Pilkington, Eds., *Time and the Earth's Rotation*, IAU Symp. 82, Reidel, Dordrecht, 1979.

NASA, *Radio Interferometry Techniques for Geodesy*, NASA Conf. Pub. 2115, National Aeronautics and Space Administration, Washington, DC, 1980.

Reid, M. J. and J. M. Moran, Eds., *The Impact of VLBI on Astrophysics and Geophysics*, IAU Symp. 129, Kluwer, Dordrecht, 1988.

REFERENCES

Alef, W., Introduction to Phase-Reference Mapping, *Very Long Baseline Interferometry. Techniques and Applications*, M. Felli and R. E. Spencer, Eds., Kluwer, Dordrecht, 1989, pp. 261–274.

Backer, D. C. and R. A. Sramek, Proper Motion of the Compact, Nonthermal Radio Source in the Galactic Center, Sagittarius A*, *Astrophys. J.*, **524**, 805–815, 1999.

Bartel, N., M. I. Ratner, I. I. Shapiro, R. J. Cappallo, A. E. E. Rogers, and A. R. Whitney, Pulsar Astrometry via VLBI, *Astron. J.*, **90**, 318–325, 1985.

Bartel, N., M. I. Ratner, I. I. Shapiro, T. A. Herring, and B. E. Corey, Proper Motion of Components of the Quasar 3C345, in *VLBI and Compact Radio Sources*, R. Fanti, K. Kellermann, and G. Setti, Eds., IAU Symp. 110, Reidel, Dordrecht, 1984, pp. 113–116.

Beasley, A. J. and J. E. Conway, VLBI Phase-Referencing, *Very Long Baseline Interferometry and the VLBA*, J. A. Zensus, P. J. Diamond, and P. J. Napier, Eds., Astron. Soc. Pacific Conf. Ser., **82**, 327–343, 1995.

Bevington, P. R. and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed., McGraw-Hill, New York, 1992.

Boboltz, D. A., P. J. Diamond, and A. J. Kemball, R. Aquarri: First Detection of Circumstellar SiO Maser Proper Motions, *Astrophys. J.*, **487**, L147–L150, 1997.

Campbell, R. M., N. Bartel, I. I. Shapiro, M. I. Ratner, R. J. Capallo, A. R. Whitney, and N. Putnam, VLBI-Derived Trigonometric Parallax and Proper Motion of PSR B2021+51, *Astrophys. J.*, **461**, L95–L98, 1996.

Carter, W. E. and D. S. Robertson, Very-Long-Baseline Interferometry Applied to Geophysics, in *Developments in Astrometry and Their Impact on Astrophysics and Geodynamics*, I. I. Mueller and B. Kolaczek, Eds., Kluwer, Dordrecht, 1993, pp. 133–144.

Carter, W. E., D. S. Robertson, and J. R. MacKay, Geodetic Radio Interferometric Surveying: Applications and Results, *J. Geophys. Res.*, **90**, 4577–4587, 1985.

Carter, W. E., D. S. Robertson, J. E. Pettey, B. D. Tapley, B. E. Schutz, R. J. Eanes, and M. Lufeng, Variations in the Rotation of the Earth, *Science*, **224**, 957–961, 1984.

Chandler, S. C., On the Variation of Latitude, *Astron. J.*, **11**, 65–70, 1891.

Clark, T. A., B. E. Corey, J. L. Davis, G. Elgered, T. A. Herring, H. F. Hinteregger, C. A. Knight, J. I. Levine, G. Lundqvist, C. Ma, E. F. Nesman, R. B. Phillips, A. E. E. Rogers, B. O. Rönnäng, J. W. Ryan, B. R. Schupler, D. B. Shaffer, I. I. Shapiro, N. R. Vandenberg, J. C. Webber, and A. R. Whitney, Precise Geodesy Using the Mark-III Very-Long-Baseline Interferometer System, *IEEE Trans. Geosci. Remote Sensing*, **GE-23**, 438–449, 1985.

Cohen, M. H. and D. B. Shaffer, Positions of Radio Sources from Long-Baseline Interferometry, *Astron. J.*, **76**, 91–100, 1971.

Counselman, C. C., III, Radio Astrometry, *Ann. Rev. Astron. Astrophys.*, **14**, 197–214, 1976.

Counselman, C. C., III, S. M. Kent, C. A. Knight, I. I. Shapiro, T. A. Clark, H. F. Hinteregger, A. E. E. Rogers, and A. R. Whitney, Solar Gravitational Deflection of Radio Waves Measured by Very Long Baseline Interferometry, *Phys. Rev. Lett.*, **33**, 1621–1623, 1974.

Elitzur, M., *Astronomical Masers*, Kluwer, Dordrecht, 1992.

Elsmore, B. and M. Ryle, Further Astrometric Observations with the 5-km Radio Telescope, *Mon. Not. R. Astron. Soc.*, **174**, 411–423, 1976.

Fanselow, J. L., O. J. Sovers, J. B. Thomas, G. H. Purcell, Jr., E. J. Cohen, D. H. Rogstad, L. J. Skjerve, and D. J. Spitzmesser, Radio Interferometric Determination of Source Positions Utilizing Deep Space Network Antennas—1971 to 1980, *Astron. J.*, **89**, 987–998, 1984.

Fey, A. L. and P. Charlot, VLBA Observations of Radio Reference Frame Structures. II. Astrometric Suitability Based on Observed Structure, *Astrophys. J. Suppl.*, **111**, 95–142, 1997.

Fomalont, E. B., W. M. Goss, A. G. Lyne, R. N. Manchester, and K. Justtanont, Positions and Proper Motions of Pulsars, *Mon. Not. R. Astron. Soc.*, **258**, 479–510, 1992.

Gauss, K. F., *Theoria Motus*, 1809; repr. in transl. as *Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*, Dover, New York, 1963, p. 249.

Genzel, R., M. J. Reid, J. M. Moran, and D. Downes, Proper Motions and Distances of H_2O Maser Sources. I. The Outflow in Orion-KL, *Astrophys. J.*, **244**, 884–902, 1981.

Gwinn, C. R., J. M. Moran, and M. J. Reid, Distance and Kinematics of the W49N H_2O Maser Outflow, *Astrophys. J.*, **393**, 149–164, 1992.

Hall, T., *Karl Friedrich Gauss*, MIT Press, Cambridge, MA, 1970, p. 74.

Hamilton, W. C., *Statistics in Physical Science*, Ronald, New York, 1964.

Hazard, C., J. Sutton, A. N. Argue, C. M. Kenworthy, L. V. Morrison, and C. A. Murray, Accurate Radio and Optical Positions of 3C273B, *Nature Phys. Sci.*, **233**, 89–91, 1971.

Herring, T. A., Geodetic Applications of GPS, *Proc. IEEE*, Special Issue on Global Positioning System, **87**, No. 1, 92–110, 1999.

Herring, T. A., B. E. Corey, C. C. Counselman III, I. I. Shapiro, A. E. E. Rogers, A. R. Whitney, T. A. Clark, C. A. Knight, C. Ma, J. W. Ryan, B. R. Schupler, N. R. Vandenberg, G. Elgered, G. Lundqvist, B. O. Rönnäng, J. Campbell, and P. Richards, Determination of Tidal Parameters from VLBI Observations, in *Proc. 9th Int. Symp. Earth Tides*, J. Kuo, Ed., E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, 1983, pp. 205–211.

Herring, T. A., C. R. Gwinn, and I. I. Shapiro, Geodesy by Radio Interferometry: Corrections to the IAU 1980 Nutation Series, in *Proc. MERIT/COTES Symp.*, I. I. Mueller, Ed., Ohio State Univ. Press, Columbus, OH, 1985.

Herring, T. A., I. I. Shapiro, T. A. Clark, C. Ma, J. W. Ryan, B. R. Schupler, C. A. Knight, G. Lundqvist, D. B. Shaffer, N. R. Vandenberg, H. F. Hinteregger, R. B. Phillips, A. E. E. Rogers, J. C. Webber, A. R. Whitney, G. Elgered, B. O. Rönnäng, B. E. Corey, and J. L. Davis, Geodesy by Radio Interferometry: Evidence for Contemporary Plate Motion, *J. Geophys. Res.*, **91**, 8344–8347, 1986.

Hinteregger, H. F., I. I. Shapiro, D. S. Robertson, C. A. Knight, R. A. Ergas, A. R. Whitney, A. E. E. Rogers, J. M. Moran, T. A. Clark, and B. F. Burke, Precision Geodesy Via Radio Interferometry, *Science*, **178**, 396–398, 1972.

Hocke, K. and K. Schlegel, A Review of Atmospheric Gravity Waves and Travelling Ionospheric Disturbances 1982–1995, *Ann. Geophysicae*, **14**, 917–940, 1996.

Johnston, K. J., A. L. Fey, N. Zacharias, J. L. Russell, C. Ma, C. de Vegt, J. E. Reynolds, D. L. Jauncey, B. A. Archinal, M. S. Carter, T. E. Corbin, T. M. Eubanks, D. R. Florkowski, D. M. Hall, D. D. McCarthy, P. M. McCulloch, E. A. King, G. Nicolson, and D. B. Shaffer, A Radio Reference Frame, *Astron. J.*, **110**, 880–915, 1995.

Johnston, K. J., P. K. Seidelmann, and C. M. Wade, Observations of 1 Ceres and 2 Pallas at Centimeter Wavelengths, *Astron. J.*, **87**, 1593–1599, 1982.

Kaplan, G. H., F. J. Josties, P. E. Angerhofer, K. J. Johnston, and J. H. Spencer, Precise Radio Source Positions from Interferometric Observations, *Astron. J.*, **87**, 570–576, 1982.

Lambeck, K., *The Earth's Variable Rotation: Geophysical Causes and Consequences*, Cambridge Univ. Press, Cambridge, UK, 1980.

Lampton, M., B. Margon, and S. Bowyer, Parameter Estimation in X-Ray Astronomy, *Astrophys. J.*, **208**, 177–190, 1976.

Lestrade, J. F., VLBI Phase-Referencing for Observations of Weak Radio Sources, *Radio Interferometry: Theory, Techniques and Applications*, T. J. Cornwell and R. A. Perley, Eds., Astron. Soc. Pacific Conf. Ser., **19**, 289–297, 1991.

Lestrade, J.-F., D. L. Jones, R. A. Preston, R. B. Phillips, M. A. Titus, J. Kovalevsky, L. Lindegren, R. Hering, M. Froeschlé, J.-L. Falin, F. Mignard, C. S. Jacobs, O. J. Sovers, M. Eubanks, and D. Gabuzda, Preliminary Link of the Hipparcos and VLBI Reference Frames, *Astron. Astrophys*, **304**, 182–188, 1995.

Lestrade, J.-F., A. E. E. Rogers, A. R. Whitney, A. E. Niell, R. B. Phillips, and R. A. Preston, Phase-Referenced VLBI Observations of Weak Radio Sources. Milliarcsecond Position of Algol, *Astron. J.*, **99**, 1663–1673, 1990.

Lieske, J. H., T. Lederle, W. Fricke, and B. Morando, Expressions for the Precession Quantities Based upon the IAU (1976) System of Astronomical Constants, *Astron. Astrophys.*, **58**, 1–16, 1977.

Ma, C., E. F. Arias, T. M. Eubanks, A. L. Fey, A.-M. Gontier, C. S. Jacobs, O. J. Sovers, B. A. Archinal, and P. Charlot, The International Celestial Reference Frame as Realized by Very Long Baseline Interferometry, *Astron. J.*, **116**, 516–546, 1998.

McCarthy, D. D. and J. D. H. Pilkington, Eds., *Time and the Earth's Rotation*, IAU Symp. No. 82, Reidel, Dordrecht, 1979 (see papers on radio interferometry).

Marcaide, J. M. and I. I. Shapiro, High Precision Astrometry via Very-Long-Baseline Radio Interferometry: Estimate of the Angular Separation between the Quasars 1038 + 528A and B, *Astron. J.*, **88**, 1133–1137, 1983.

Markowitz, W. and B. Guinot, Eds., *Continental Drift, Secular Motion of the Pole, and Rotation of the Earth*, IAU Symp. No. 32, Reidel, Dordrecht, 1968, pp. 13–14.

Melchior, P., *The Tides of the Planet Earth*, Pergamon Press, Oxford, 1978.

Moran, J. M., B. F. Burke, A. H. Barrett, A. E. E. Rogers, J. A. Ball, J. C. Carter, and D. D. Cudaback, The Structure of the OH Source in W3. *Astrophys. J. (Lett.)* **152**, L97–L101, 1968.

Morrison, N., *Introduction to Sequential Smoothing and Prediction*, McGraw-Hill, New York, 1969, p. 645.

Mueller, I. I., Reference Coordinate Systems for Earth Dynamics: A Preview, in *Reference Coordinate Systems for Earth Dynamics*, E. M. Gaposchkin and B. Kolaczek, Eds., Reidel, Dordrecht, 1981, pp. 1–22.

Norris, R. P., and R. S. Booth, Observations of OH Masers in W3OH, *Mon. Not. R. Astron. Soc.*, **195**, 213–226, 1981.

Norris, R. P., R. S. Booth, and P. J. Diamond, MERLIN Spectral Line Observations of W3OH, *Mon. Not. R. Astron. Soc.*, **201**, 209–222, 1982.

Patnaik, A. R., I. W. A. Browne, P. N. Wilkinson, and J. M. Wrobel, Interferometer Phase Calibration Sources–I. The Region $35° \leq \delta \leq 75°$, *Mon. Not. R. Astron. Soc.*, **254**, 655–676, 1992.

Perley, R. A., The Positions, Structures, and Polarizations of 404 Compact Radio Sources, *Astron. J.*, **87**, 859–880, 1982.

Petley, B. W., New Definition of the Metre, *Nature*, **303**, 373–376, 1983.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes*, 2nd ed., Cambridge U. Press, 1992.

Reid, M. J., A. D. Haschick, B. F. Burke, J. M. Moran, K. J. Johnston, and G. W. Swenson, Jr., The Structure of Interstellar Hydroxyl Masers: VLBI Synthesis Observations of W3(OH), *Astrophys. J.*, **239**, 89–111, 1980.

Reid, M. J., A. C. S. Readhead, R. C. Vermeulen, and R. N. Treuhaft, The Proper Motion of Sagittarius A*. I. First VLBA Results, *Astrophys. J.*, **524**, 816–823, 1999.

Reid, M. J. and J. M. Moran, Astronomical Masers, in *Galactic and Extragalactic Radio Astronomy*, G. L. Verschuur and K. I. Kellermann, Eds., Kluwer, Dordrecht, 1988.

Robertson, D. S., W. E. Carter, R. J. Eanes, B. E. Schutz, B. D. Tapley, R. W. King, R. B. Langley, P. J. Morgan, and I. I. Shapiro, Comparison of Earth Rotation as Inferred from Radio Interferometric, Laser Ranging, and Astrometric Observations, *Nature*, **302**, 509–511, 1983.

Ros, E., J. M. Marcaide, J. C. Guirado, M. I. Ratner, I. I. Shapiro, T. P. Krichbaum, A. Witzel, and R. A. Preston, High Precision Difference Astrometry Applied to the Triplet of S5 Radio Sources B1803+784/Q1928+738/B2007+777, *Astron. Astrophys.*, **348**, 381–393, 1999.

Ryle, M. and B. Elsmore, Astrometry with 5-km Telescope, *Mon. Not. R. Astron. Soc.*, **164** 223–242, 1973.

Seidelmann, P. K., Ed., *Explanatory Supplement to the Astronomical Almanac*, University Science Books, Mill Valley, CA, 1992.

Shapiro, I. I., Estimation of Astrometric and Geodetic Parameters, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 261–276.

Shapiro, I. I., D. S. Robertson, C. A. Knight, C. C. Counselman III, A. E. E. Rogers, H. F. Hinteregger, S. Lippincott, A. R. Whitney, T. A. Clark, A. E. Niell, and D. J. Spitzmesser, Transcontinental Baselines and the Rotation of the Earth Measured by Radio Interferometry, *Science*, **186**, 920–922, 1974.

Shapiro, I. I., J. J. Wittels, C. C. Counselman III, D. S. Robertson, A. R. Whitney, H. F. Hinteregger, C. A. Knight, A. E. E. Rogers T. A. Clark, L. K. Hutton, and A. E. Niell, Submilliarcsecond Astrometry via VLBI. I. Relative Position of the Radio Sources 3C345 and NRAO512, *Astron. J.*, **84**, 1459–1469, 1979.

Smith, F. G., The Determination of the Position of a Radio Star, *Mon. Not. R. Astron. Soc.*, **112**, 497–513, 1952.

Smith, H. M., International Time and Frequency Coordination, *Proc. IEEE*, **60**, 479–487, 1972.

Sovers, O. J., J. L. Fanselow, and C. S. Jacobs, Astrometry and Geodesy with Radio Interferometry: Experiments, Models, Results, *Rev. Mod. Phys.*, **70**, 1393–1454, 1998.

Sramek, R. A., Atmospheric Phase Stability at the VLA, in *Radio Astonomical Seeing*, J. E. Baldwin and Wang Shouguan, Eds., International Academic Publishers and Pergamon Press, Oxford, 1990, pp. 21–30.

Taff, L. G., *Computational Spherical Astronomy*, Wiley, New York, 1981.

Tartarski, V. I., *Wave Propagation in a Turbulent Medium*, transl. by R. A. Silverman, McGraw-Hill, New York, 1961.

Taylor, J. H., C. R. Gwinn, J. M. Weisberg, and L. A. Rawley, Pulsar Astrometry, in *VLBI and Compact Radio Sources*, R. Fanti, K. Kellermann, and G. Setti, Eds., IAU Symposium 110, Reidel, Dordrecht, 1984, pp. 347–353.

Wahr, J. M., The Forced Nutations of an Elliptical, Rotating, Elastic and Oceanless Earth, *Geophys. J. R. Astron. Soc.*, **64**, 705–727, 1981.

Wahr, J. M., *Geodesy and Gravity*, Samezdot Press, Golden, CO, 1996.

Walker, R. C., 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, 1981.

Walker, R. C., D. N. Matsakis, and J. A. Garcia-Barreto, H_2O Masers in W49N. I. Maps. *Astrophys. J.*, **255**, 128–142, 1982.

Woolard, E. W. and G. M. Clemence, *Spherical Astronomy*, Academic Press, New York, 1966.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

13 Propagation Effects

The neutral and ionized media lying between a radio source and the surface of the earth often have profound effects on the radiation fields traversing them. The most important of these media are the neutral lower atmosphere, or troposphere, the ionosphere, the ionized interplanetary medium, and the ionized interstellar medium. We are concerned with three types of effects of these media. First, the large-scale structures in the media give rise to refractive effects. These effects, which can be analyzed in terms of geometric optics and Fermat's principle, are the deflection of the radio waves, the change of the propagation velocity, and the rotation of the plane of polarization. Second, radiation can be absorbed. Finally, radiation can be scattered by the turbulent structure in the media. The phenomenon of scattering results in scintillation, or seeing.

In the troposphere, water vapor plays a particularly important role in radio propagation. The refractivity of water vapor is about 20 times greater in the radio range than in the near-infrared or optical regimes. The phase fluctuations in radio interferometers at centimeter, millimeter, and submillimeter wavelengths are caused predominantly by fluctuations in the distribution of water vapor. Water vapor is poorly mixed in the atmosphere, and the total column density of water vapor cannot be accurately sensed from surface meteorological measurements. Uncertainties in the water vapor content are a fundamental limitation to the accuracy of VLBI measurements. Small-scale (<1 km) fluctuations in water vapor distribution limit the angular resolution of connected-element interferometers. Furthermore, spectral lines of water vapor cause substantial absorption at frequencies above 100 GHz and usually render the troposphere completely opaque at frequencies between 1 and 10 THz (300 and 30 μm). Thus, any discussion of the neutral atmosphere must be primarily concerned with the effects of water vapor. Propagation in the neutral atmosphere from the point of view of radio communications is discussed by Crane (1981) and Bohlander, McMillan, and Gallagher (1985).

Above the neutral atmosphere, radiation encounters three morphologically distinct plasmas: the ionosphere, the interplanetary medium, and the interstellar medium. Most plasma effects that concern us scale as ν^{-2}. Therefore, detrimental effects can be mitigated by carrying out investigations at the highest frequency possible. However, the effects of the ionosphere can easily be detected in VLBI measurements at frequencies up to at least 10 GHz. Furthermore, because of as-

trophysical requirements, many observations must be made at frequencies where plasma effects cause problems.

Our interest in the propagation media arises because the media degrade interferometric measurements of radio sources. Alternatively, observations of radio sources can be used to probe the characteristics of the propagation media. Radio interferometric measurements have been used widely for this purpose.

13.1 NEUTRAL ATMOSPHERE

In the lowest part of the atmosphere the temperature decreases monotonically from the surface at a rate of about 6.5 K km^{-1}, except for an occasional low-level inversion, until it reaches about 218 K at an altitude of approximately 11 km. This lowermost layer is called the troposphere. Above 11 km the temperature is constant for a distance of about 10 km in the region called the tropopause. Above the tropopause the temperature begins to rise with altitude in the stratosphere. Within the neutral atmosphere, the propagation of radio waves is most affected by the troposphere. Before discussing the refraction, absorption, and scattering of radio waves in the troposphere in detail, we introduce some basic physical concepts.

Basic Physics

Consider a plane wave propagating along the y direction in a uniform dissipative dielectric medium, as represented by the equation

$$\mathbf{E}(y, t) = \mathbf{E}_0 e^{j(kny - 2\pi\nu t)}, \tag{13.1}$$

where k is the propagation constant in free space and is equal to $2\pi\nu/c$, c is the velocity of light, and $\mathbf{E}_0$ is the electric field amplitude. n is the complex index of refraction, equal to $n_R + jn_I$. If the imaginary part of the index of refraction is positive, the wave will decay exponentially. The power absorption coefficient is defined as

$$\alpha = \frac{4\pi\nu}{c} n_I. \tag{13.2}$$

The propagation constant in the atmosphere is k multiplied by the real part of the index of refraction, which can be written

$$kn_R = \frac{2\pi n\nu}{c} = \frac{2\pi\nu}{v_p}, \tag{13.3}$$

where $n = n_R$ is the index of refraction when absorption is neglected, and v_p is the phase velocity. The phase velocity of the wave, c/n, is less than c by about 0.03% in the lower atmosphere. The extra time required to traverse a medium with index of refraction $n(y)$ compared with the time necessary to traverse the

same distance in free space is

$$\Delta t = \frac{1}{c}\int (n-1)\,dy, \tag{13.4}$$

where we assume that the effect of the difference in physical length between the actual ray path and the straight-line path is negligible. The *excess* path length is defined as $c\Delta t$, or

$$\mathcal{L} = 10^{-6}\int N(y)\,dy, \tag{13.5}$$

where we have introduced the refractivity N, defined by $N = 10^6(n-1)$. Note that the concept of excess path length, which is used extensively in this chapter, does not represent an actual physical path.

The refractivity of moist air in the radio range is given by the empirical formula (see discussion in this section under *Smith–Weintraub Equation*)

$$N = 77.6\frac{p_D}{T} + 64.8\frac{p_V}{T} + 3.776\times 10^5\frac{p_V}{T^2}, \tag{13.6}$$

where T is the temperature in kelvins, p_D is the partial pressure of the dry air, and p_V is the partial pressure of water vapor in millibars (1 mb = 100 newtons per square meter = 100 pascals; 1 atmosphere = 1013 mb). The first two terms on the right-hand side of Eq. (13.6) arise from the displacement polarizations of the gaseous constituents of the air. The third term is due to the permanent dipole moment of water vapor. Equation (13.6) is accurate to better than 1% for frequencies below 100 GHz. The contributions of dispersive components of refractivity associated with resonances are very small (see discussion in this section under *Origin of Refraction*).

The refractivity can be expressed in terms of gas density, using the ideal gas law

$$p = \frac{\rho RT}{\mathcal{M}}, \tag{13.7}$$

where p and ρ are the partial pressure and density of any constituent gas, R is the universal gas constant, equal to 8.314 J mol^{-1} K^{-1}, and $\mathcal{M}$ is the molecular weight, which for dry air in the troposphere is $\mathcal{M}_D = 28.96$ g mol^{-1} and for water vapor is $\mathcal{M}_V = 18.02$ g mol^{-1}. Thus, $p_D = \rho_D RT/\mathcal{M}_D$ and $p_V = \rho_V RT/\mathcal{M}_V$, where ρ_D and ρ_V are the densities of dry air and water vapor, respectively. Since the total pressure P is the sum of the partial pressures, and the total density ρ_T is the sum of the constituent densities, Eq. (13.7) can be written $P = \rho_T RT/\mathcal{M}_T$, where

$$\mathcal{M}_T = \left(\frac{1}{\mathcal{M}_D}\frac{\rho_D}{\rho_T} + \frac{1}{\mathcal{M}_V}\frac{\rho_V}{\rho_T}\right)^{-1}. \tag{13.8}$$

Substitution of the appropriate forms of Eq. (13.7) and the equation $\rho_D = \rho_T - \rho_V$ into Eq. (13.6) yields

$$N = 0.2228\rho_T + 0.076\rho_V + 1742\frac{\rho_V}{T}, \tag{13.9}$$

where ρ_T and ρ_V are in g m^{-3}. Since the second term on the right-hand side of Eq. (13.9) is small with respect to the third term, it can be combined with the third term to give, for $T = 280$ K,

$$N \simeq 0.2228\rho_T + 1763\frac{\rho_V}{T} = N_D + N_V. \tag{13.10}$$

Equation (13.10) defines the dry and wet refractivities, N_D and N_V, respectively. These definitions are not universally followed in the literature. Note that N_D is proportional to the total density and therefore has a contribution due to the induced dipole moment of water vapor. Graphs of mean values of N_V around the world are shown in Fig. 13.1.

The atmosphere obeys the equation of hydrostatic equilibrium to a high degree of accuracy (Humphreys 1940). A parcel of gas in static equilibrium between pressure and gravity obeys the equation

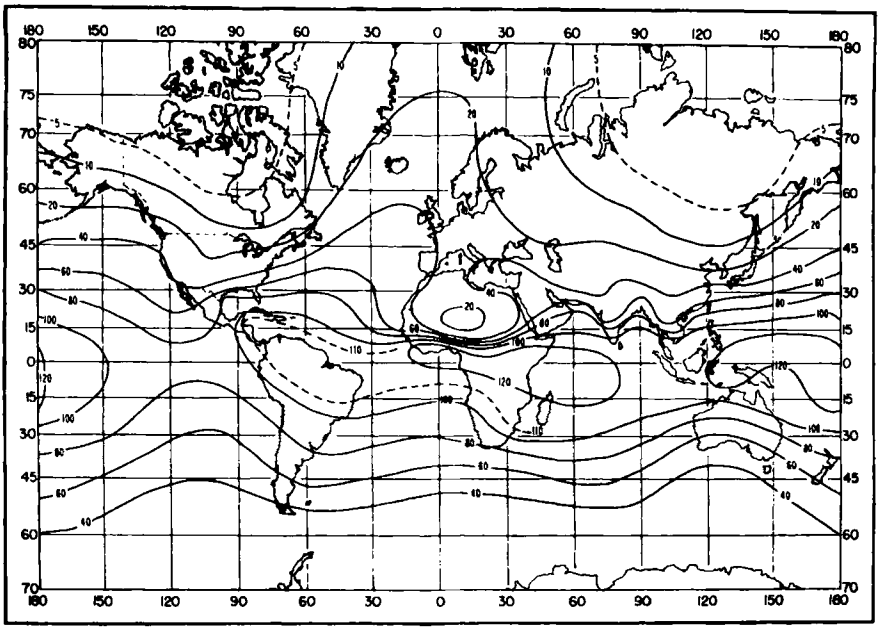
$$\frac{dP}{dh} = -\rho_T \mathrm{g}, \tag{13.11}$$

where g is the acceleration due to gravity, approximately equal to 980 cm s^{-2}, and h is the height above the earth's surface. Using the ideal gas law, Eq. (13.7), we can integrate Eq. (13.11) assuming specific forms for the temperature profile and mixing ratio. If an isothermal atmosphere with constant mixing ratio is assumed, then ρ_T is an exponential function with a scale height of $RT/\mathcal{M}g \simeq 8.5$ km for 290 K, which is close to the observed scale height. Other models are described by Hess (1959). The excess path length caused by the dry component of refractivity does not depend on the height distribution of total density or temperature, but only on the surface pressure P_0, under conditions of hydrostatic equilibrium. If g is assumed to be constant with height, the surface pressure can be obtained by integrating Eq. (13.11),

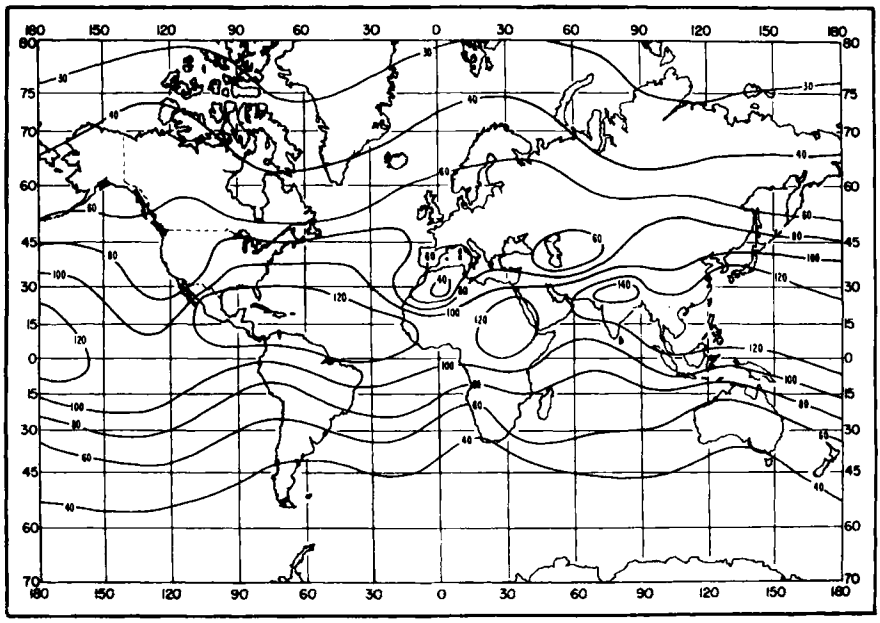
$$P_0 = \mathrm{g}\int_0^\infty \rho_T(h)\,dh. \tag{13.12}$$

From Eqs. (13.5), (13.10), and (13.12), the dry excess path length in the zenith direction is

$$\mathcal{L}_D = 10^{-6}\int_0^\infty N_D\,dh = AP_0, \tag{13.13}$$



(a)



(b)

Figure 13.1 (a) Worldwide distribution of the mean sea-level value of the wet refractivity N_V for February; (b) N_V for August. Note the seasonal variation in mean water vapor content. From Bean et al. (1966).

where $A = 77.6R/g\mathcal{M}_D = 0.228$ cm mb^{-1}. Under standard conditions for which $P_0 = 1013$ mb, the value of $\mathcal{L}_D$ is 231 cm.

Water vapor is not well mixed in the atmosphere and therefore is not well correlated with ground-based meteorological parameters (Reber and Swope 1972). On average, water vapor density has an exponential distribution with a scale height of 2 km. The partial pressure and density of water vapor from Eq. (13.7) are related by

$$\rho_V = \frac{217 p_V}{T} (\text{g m}^{-3}). \tag{13.14}$$

The partial pressure of water vapor for saturated air, p_{VS}, at temperature T, obtained from the Clausius–Clapeyron equation (Hess 1959), can be approximated to an accuracy of better than 1% within the temperature range 240–310 K by the formula (Crane 1976)

$$p_{VS} = 6.11 \left(\frac{T}{273}\right)^{-5.3} e^{25.2(T-273)/T} \text{ (mb)}. \tag{13.15}$$

The relative humidity is p_V/p_{VS}. The component of the path length resulting primarily from the permanent dipole moment of water vapor is

$$\mathcal{L}_V = 1763 \times 10^{-6} \int_0^\infty \frac{\rho_V(h)}{T(h)} dh, \tag{13.16}$$

where the units of $\mathcal{L}_V$ are the same as those of h. If we assume that the atmosphere is isothermal and that p_V decreases exponentially with a scale height of 2 km, then from Eqs. (13.14) and (13.16)

$$\mathcal{L}_V = 7.6 \times 10^4 \frac{p_{V0}}{T^2} \text{ (cm)}, \tag{13.17}$$

where p_{V0} is the partial pressure of water vapor at the surface of the earth. Hence, at ambient air temperature, $\mathcal{L}_V$ in centimeters is approximately equal to p_{V0} in millibars. For an exponential distribution of density with a scale height of 2 km and temperature of 280 K, the path length is given by $\mathcal{L}_V = 1.26\rho_{V0}$, where ρ_{V0} is the water vapor density at the surface.

The integrated water vapor density, or the height of the column of water condensed from the atmosphere, is given by

$$w = \frac{1}{\rho_w} \int_0^\infty \rho_V(h)\, dh, \tag{13.18}$$

where ρ_w is the density of water, 10^6 g m^{-3}. Hence, from Eq. (13.16) for an isothermal atmosphere at 280 K

$$\mathcal{L}_V \simeq 6.3w. \tag{13.19}$$

This formula, which is widely used in the literature, is an excellent approximation for frequencies below 100 GHz. In the windows above 100 GHz, the ratio $\mathcal{L}_V/w$ can vary from 6.3 to about 8 (see Fig. 13.8 and associated discussion). The values of $\mathcal{L}_V$ under extreme conditions for a temperate, sea-level site can be calculated from the equations above. With $T = 303$ K (30°C) and relative humidity = 0.8, we have $p_{V0} = 34$ mb, $\rho_{V0} = 24$ g m^{-3}, $w = 4.9$ cm, and $\mathcal{L}_V = 28$ cm. With $T = 258$ K (−15°C) and relative humidity = 0.5, we have $p_{V0} = 1.0$ mb, $\rho_{V0} = 0.8$ g m^{-3}, $w = 0.15$ cm, and $\mathcal{L}_V = 1.1$ cm. The total zenith excess path length through the atmosphere is $\mathcal{L} \simeq \mathcal{L}_D + \mathcal{L}_V$, which from Eqs. (13.13) and (13.19) is

$$\mathcal{L} \simeq 0.228 P_0 + 6.3w \text{ (cm)}, \tag{13.20}$$

where P_0 is in millibars, and w is in centimeters. Equation (13.20) is reasonably accurate for estimation purposes because the fractional variation in the temperature of the lower atmosphere, and in the scale height of water vapor, is usually less than 10%. However, it is usually not accurate enough to predict the path length to a small fraction of a wavelength at millimeter wavelengths.

Refraction and Propagation Delay

If the vertical distributions of temperature and water vapor pressure are known, then precise estimates of the angle of arrival and excess propagation time for a ray impinging on the atmosphere at an arbitrary angle can be computed by ray tracing. Here we consider a few elementary cases in order to derive some simple analytic expressions. The simplest case is that of an interferometer in a uniform or plane-parallel atmosphere, as shown in Fig. 13.2. The refraction of the ray is

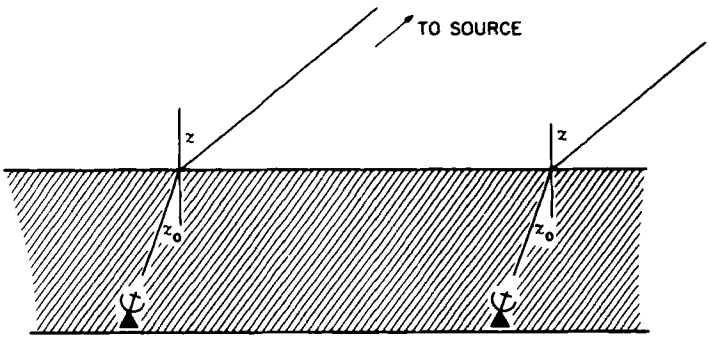


Figure 13.2 Two-element interferometer with the atmosphere modeled as a uniform flat slab. The geometric delay is the same as it would be if the interferometer were in free space.

governed by Snell's law, which is

$$n_0 \sin z_0 = \sin z, \tag{13.21}$$

where z is the zenith angle at the top of the atmosphere (where $n = 1$), and z_0 is the zenith angle at the surface (where $n = n_0$). The geometric delay for an interferometer, as defined in Chapter 2, is

$$\tau_g = \frac{n_0 D}{c} \sin z_0 = \frac{D}{c} \sin z. \tag{13.22}$$

τ_g can be calculated from the angle of arrival z_0 and the velocity of light at the earth's surface c/n_0, or from z and the velocity of light in free space. Thus, if earth curvature is neglected and the atmosphere is uniform, then the resulting geometric delay is the same as the free-space value. The angle of refraction need only be calculated to ensure that the antennas track the source properly. The angle of refraction, $\Delta z = z - z_0$, can be written, using Eq. (13.21), as

$$\Delta z = z - \sin^{-1}\left(\frac{1}{n_0} \sin z\right). \tag{13.23}$$

This equation can be expanded in a Taylor series in $n_0 - 1$, which to first order gives

$$\Delta z \simeq (n_0 - 1) \tan z. \tag{13.24}$$

Since $n_0 - 1 \simeq 3 \times 10^{-4}$ at the surface of the earth, Eq. (13.24) can be written

$$\Delta z \text{ (arcmin)} \simeq \tan z. \tag{13.25}$$

The angle of refraction can also be calculated for more realistic cases. Ignore the curvature of the earth and consider the atmosphere to consist of a large number of plane-parallel layers numbered 0 through m, as shown in Fig. 13.3. Let the index of refraction at the surface be n_0, and at the top layer, $n_m = 1$. Applying Snell's law to the various layers gives the following set of equations:

$$\begin{aligned} n_0 \sin z_0 &= n_1 \sin z_1 \\ n_1 \sin z_1 &= n_2 \sin z_2 \\ \vdots \qquad & \quad \vdots \\ n_{m-1} \sin z_{m-1} &= \sin z, \end{aligned} \tag{13.26}$$

where $z = z_m$. From these equations, we see that $n_0 \sin z_0 = \sin z$. This result is identical to that for the homogenous case. Thus, regardless of the vertical distribution of the index of refraction, the angle of refraction is given by Eq. (13.21), where n_0 is the surface value of the index of refraction. This result can also be

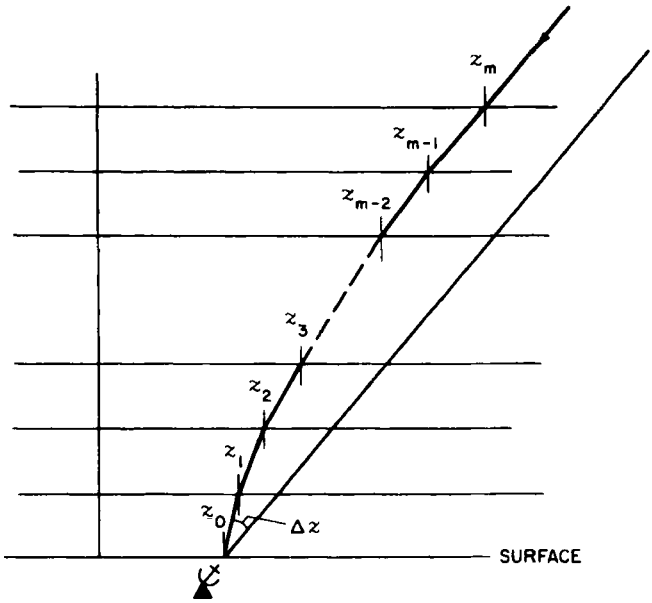


Figure 13.3 The atmosphere modeled as a set of thin, uniform slabs. The angle of incidence on the topmost slab is z_m, which is equal to the free-space zenith angle z, and the angle of incidence at the surface is z_0. The total bending is $\Delta z = z - z_0$.

obtained by an elementary application of Fermat's principle. An interesting application of this result is that if $n_0 = 1$, as would be the case if the measuring device were in a vacuum chamber at the surface of the earth, then there would be no net refraction; that is, $z_0 = z$.

For an atmosphere consisting of spherical layers, the angle of refraction is given by the formula (Smart 1977)

$$\Delta z = r_0 n_0 \sin z_0 \int_1^{n_0} \frac{dn}{n\sqrt{r^2 n^2 - r_0^2 n_0^2 \sin^2 z_0}}, \tag{13.27}$$

where r is the distance from the center of the earth to the layer where the index of refraction is n and r_0 is the radius of the earth. This result is derivable from Snell's law in spherical coordinates: $nr \sin z =$ constant (Smart 1977). For small zenith angles, expansion of Eq. (13.27) gives

$$\Delta z \simeq (n_0 - 1) \tan z_0 - a_2 \tan z_0 \sec^2 z_0, \tag{13.28}$$

where a_2 is a constant. Equation (13.28) can also be written

$$\Delta z \simeq a_1 \tan z_0 - a_2 \tan^3 z_0, \tag{13.29}$$

where $a_1 \simeq 56$ arcsec and $a_2 \simeq 0.07$ arcsec for a dry atmosphere under standard conditions (COESA 1976). The refraction at the horizon is about 0.46° (see Fig. 13.5). See Saastamoinen (1972a) for a more detailed treatment.

The differential delay induced in an interferometer by a horizontally stratified troposphere results from the difference in zenith angle of the source at the antennas. Consider two closely spaced antennas. If the excess path in the zenith direction is $\mathcal{L}_0$, then the excess path in other directions is approximately $\mathcal{L}_0 \sec z$. This approximation becomes inaccurate at large zenith angles. The difference in excess paths, $\Delta\mathcal{L}$, is

$$\Delta\mathcal{L} \simeq \mathcal{L}_0 \, \Delta z \frac{\sin z}{\cos^2 z}, \tag{13.30}$$

where Δz is the difference in zenith angles at the two antennas.

If the antennas are on the equator and the source has a declination of zero, then Δz is equal to the difference in longitudes of D/r_0, where D is the separation between antennas. For this case,

$$\Delta\mathcal{L} \simeq \frac{\mathcal{L}_0 D}{r_0} \frac{\sin z}{\cos^2 z}. \tag{13.31}$$

If $D = 10$ km, $\mathcal{L}_0 = 230$ cm, $r_0 = 6370$ km, and $z = 80°$, then $\Delta\mathcal{L}$ is 12 cm. The calculation of the difference in excess paths can be easily generalized as follows. Let $\mathbf{r}_1$ and $\mathbf{r}_2$ be vectors from the center of the earth to each antenna. The geometric delay is $(\mathbf{r}_1 \cdot \mathbf{s} - \mathbf{r}_2 \cdot \mathbf{s})/c$, where $\mathbf{s}$ is the unit vector in the direction of the source. Since $\cos z_1 = (\mathbf{r}_1 \cdot \mathbf{s})/r_0$ and $\cos z_2 = (\mathbf{r}_2 \cdot \mathbf{s})/r_0$, where z_1 and z_2 are the zenith angles at the two antennas, the geometric delay can be written

$$\tau_g = \frac{r_0}{c}(\cos z_1 - \cos z_2) \simeq \frac{r_0}{c} \Delta z \sin z. \tag{13.32}$$

Substitution of Δz from Eq. (13.32) into Eq. (13.30) yields an expression for the difference in excess path lengths, valid for short-baseline interferometers and moderate values of zenith angle:

$$\Delta\mathcal{L} \simeq \frac{c\tau_g \mathcal{L}_0}{r_0} \sec^2 z. \tag{13.33}$$

For very-long-baseline interferometers, the expression in Eq. (13.30) is not appropriate. The difference in excess path lengths is approximately $\Delta\mathcal{L} = \mathcal{L}_1 \sec z_1 - \mathcal{L}_2 \sec z_2$, where $\mathcal{L}_1$, $\mathcal{L}_2$, z_1, and z_2 are the excess zenith path lengths and the zenith angles at the two antennas. We now derive a more accurate expression for the excess path length to each antenna. Assume the index of refraction to be exponentially distributed with a scale height h_0. The geometry is shown in Fig. 13.4. The excess path length is

$$\mathcal{L} = 10^{-6} N_0 \int_0^\infty \exp\left(-\frac{h}{h_0}\right) dy, \tag{13.34}$$

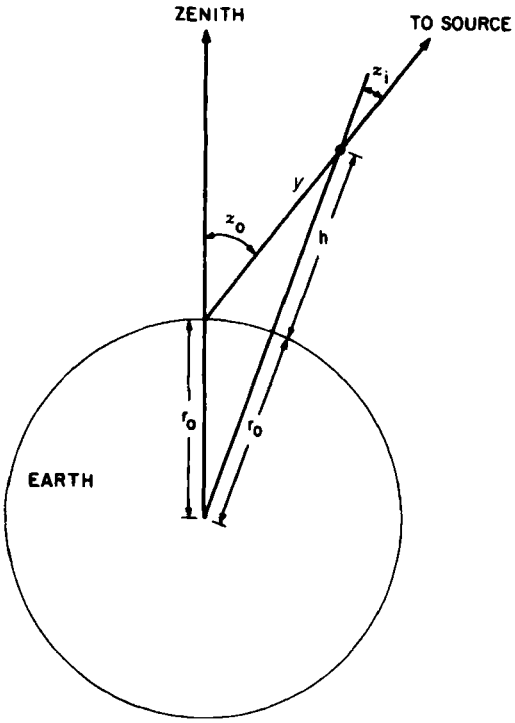


Figure 13.4 Geometry for calculating the propagation delay, taking into account the sphericity of the earth. The ray path along the y coordinate is assumed to be straight. The angle z_i is the zenith angle of the ray at height h. This angle is needed in the calculation of the excess path length through the ionosphere [Eqs. (13.139) and (13.140)].

where N_0 is the refractivity at the earth's surface, h is the height above the surface, and dy is the differential length along the ray path. Bending of the ray is neglected. From the geometry of Fig. 13.4, one can show that

$$h \simeq y \cos z + \frac{y^2}{2r_0} \sin^2 z. \tag{13.35}$$

Therefore

$$\mathcal{L} \simeq 10^{-6} N_0 \int_0^\infty \exp\left(-\frac{y}{h_0} \cos z\right) \exp\left(-\frac{y^2}{2r_0 h_0} \sin^2 z\right) dy. \tag{13.36}$$

The argument of the rightmost exponential function in Eq. (13.36) is small, and this exponential function can be expanded in a Taylor series so that

$$\mathcal{L} \simeq 10^{-6} N_0 \int_0^\infty \exp\left(-\frac{y}{h_0}\cos z\right) \times \left(1 - \frac{y^2}{2r_0 h_0}\sin^2 z \cdots\right) dy. \quad (13.37)$$

Integration of Eq. (13.37) yields

$$\mathcal{L} \simeq 10^{-6} N_0 h_0 \sec z \left(1 - \frac{h_0}{r_0}\tan^2 z\right). \quad (13.38)$$

Equation (13.38) can also be written

$$\mathcal{L} \simeq 10^{-6} N_0 h_0 \left[\left(1 + \frac{h_0}{r_0}\right)\sec z - \frac{h_0}{r_0}\sec^3 z\right]. \quad (13.39)$$

Thus, $\mathcal{L}$ is a function of odd powers of $\sec z$, whereas the bending angle, given in Eq. (13.29), is a function of odd powers of $\tan z$. Equations (13.38) and (13.39) both diverge as z approaches 90°. For $z = 90°$, Eq. (13.35) shows that $h \simeq y^2/2r_0$. Hence, for Eq. (13.34), the excess path at the horizon is

$$\mathcal{L} \simeq 10^{-6} N_0 \sqrt{\frac{\pi r_0 h_0}{2}} \simeq 70\mathcal{L}_0 \simeq 14 N_0 \text{ (cm)} \quad (13.40)$$

for $r_0 = 6370$ km and $h_0 = 2$ km. A model incorporating both the dry atmosphere with a scale height $h_D = 8$ km and the wet atmosphere with a scale height $h_V = 2$ km can be obtained by applying Eq. (13.38) to both the dry and wet components using Eqs. (13.13) and (13.17). This result is

$$\mathcal{L} \simeq 0.228 P_0 \sec z\,(1 - 0.0013\tan^2 z) + \frac{7.5 \times 10^4 p_{V0} \sec z}{T^2}(1 - 0.0003\tan^2 z). \quad (13.41)$$

More sophisticated models have been derived by Marini (1972), Saastamoinen (1972b), Davis et al. (1985), Niell (1996), and others. A comparison of the approximate formula of Eq. (13.41) and a ray-tracing solution is given in Fig. 13.5.

Absorption

When the sky is clear, the principal sources of atmospheric attenuation are the molecular resonances of water vapor, oxygen, and ozone. The resonances of water vapor and oxygen are pressure broadened and cause attenuation far from the resonance frequencies. A plot of the absorption versus frequency is shown in Fig. 13.6. Below 30 GHz absorption is dominated by the weak 6_{16}–5_{23} transition of H_2O at 22.2 GHz (Liebe 1969). Absorption by this line rarely exceeds 20% in the zenith direction. The oxygen lines in the band 50–70 GHz are considerably stronger, and no astronomical observations can be made from the ground in this band. An isolated oxygen line at 118 GHz makes observations impossible in

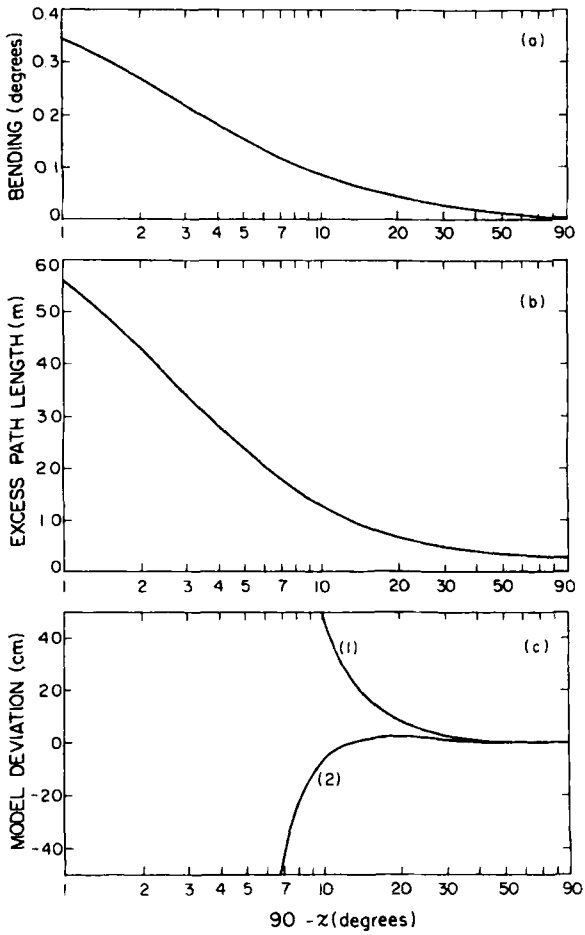


Figure 13.5 (a) The bending angle versus $90° - z$, where z is the zenith angle that the ray would have in the absence of refraction, calculated by a ray-tracing algorithm for a standard dry atmosphere (COESA 1976). (b) The excess path length versus $90° - z$ calculated by a ray-tracing algorithm. The zenith excess path is 2.31 m. (c) Deviation between the excess path length and (1) the $\mathcal{L}_0 \sec z$ model and (2) the model of Eq. (13.41); in both cases $\rho_{V0} = 0$ and the zenith excess path is the same as in (b).

the band 116–120 GHz. At higher frequencies there is a series of strong water vapor lines at 183, 325, 380, 448, 475, 557, 621, 752, 988, and 1097 GHz and higher (Liebe 1981). Observations can be made in the windows between these lines at dry locations, usually found at high altitudes. The physics of atmospheric absorption is discussed in detail by Waters (1976), and a model of absorption at frequencies below 1000 GHz is given by Liebe (1981, 1985, 1989). We are concerned here only with the phenomenology of absorption and its calibration. The absorption coefficient depends on the temperature, gas density, and total pressure.

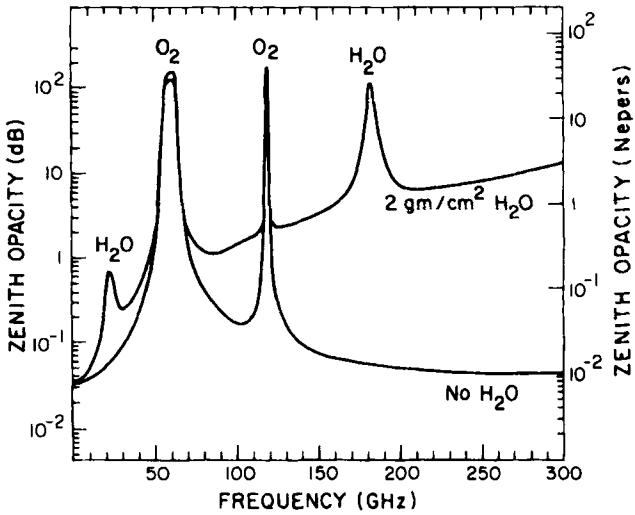


Figure 13.6 Atmospheric zenith opacity. The absorption from narrow ozone lines has been omitted. Adapted from Waters (1976). For zenith opacity at frequencies above 300 GHz, see Liebe (1981, 1989). Note that 2 g cm^{-2} of H_2O corresponds to $w = 2$ cm.

For example, the absorption coefficient for the 22 GHz H_2O line can be written (Staelin 1966)

$$\begin{aligned}\alpha = {} & \left(3.24 \times 10^{-4} e^{-644/T}\right) \frac{\nu^2 P \rho_V}{T^{3.125}} \left(1 + 0.0147 \frac{\rho_V T}{P}\right) \\ & \times \left[\frac{1}{(\nu - 22.235)^2 + \Delta\nu^2} + \frac{1}{(\nu + 22.235)^2 + \Delta\nu^2}\right] \\ & + 2.55 \times 10^{-8} \rho_V \nu^2 \frac{\Delta\nu}{T^{3/2}} \qquad (\mathrm{cm}^{-1}) .\end{aligned} \tag{13.42}$$

Here $\Delta\nu$ is approximately the half width at half maximum of the line in gigahertz, given by the equation

$$\Delta\nu = 2.58 \times 10^{-3} \left(1 + 0.0147 \frac{\rho_V T}{P}\right) \frac{P}{(T/318)^{0.625}} , \tag{13.43}$$

where ν is the frequency in gigahertz, T is the temperature in kelvins, P is the total pressure in millibars, and ρ_V is the water vapor density in grams per cubic meter. The lineshape specified by Eq. (13.42), the Van Vleck–Weisskopf profile, appears to fit the empirical data better than other theoretical profiles (Hill 1986). Other line parameterizations of the line profile are available, for example, Pol, Ruf, and Keihm (1998).

The intensity of a ray passing through an absorbing medium obeys the radiative transfer equation. We assume that the medium is in local thermodynamic equilibrium at temperature T and that scattering is negligible. In the domain where the Rayleigh–Jeans approximation to the Planck function is valid, so that the intensity is proportional to the brightness temperature, the equation of radiative transfer can be written (Rybicki and Lightman 1979)

$$\frac{dT_B}{dy} = -\alpha(T_B - T), \tag{13.44}$$

where T_B is the brightness temperature and α is the absorption coefficient defined in Eqs. (13.2) and (13.42). The solution to Eq. (13.44) for radiation propagating along the y axis is

$$T_B(\nu) = T_{B0}(\nu)e^{-\tau_\nu} + \int_0^\infty \alpha(\nu, y)T(y)e^{-\tau'_\nu}\,dy, \tag{13.45}$$

where T_{B0} is the brightness temperature in the absence of absorption, including the cosmic background component,

$$\tau'_\nu = \int_0^y \alpha(\nu, y')\,dy', \tag{13.46}$$

and

$$\tau_\nu = \int_0^\infty \alpha(\nu, y')\,dy'. \tag{13.47}$$

Here y is the distance measured from the observer. τ_ν is called the *optical depth*, or *opacity*. The first term on the right-hand side of Eq. (13.45) describes the absorption of the signal, and the second describes the emission contribution of the atmosphere. Equation (13.45) illustrates the fundamental law that an absorbing medium must also radiate. If $T(y)$ is constant throughout the medium, then Eq. (13.45) can be written

$$T_B(\nu) = T_{B0}(\nu)e^{-\tau_\nu} + T(1 - e^{-\tau_\nu}). \tag{13.48}$$

The presence of absorption can have a very significant effect on system performance. If the receiver temperature is T_R, then the system temperature, which is the sum of T_R and the atmospheric brightness temperature (the effects of ground radiation being neglected), is

$$T_S = T_R + T_{\rm at}(1 - e^{-\tau_\nu}), \tag{13.49}$$

where $T_{\rm at}$ is the temperature of the atmosphere. In the absence of a source, the antenna temperature is taken as equal to the brightness temperature of the sky.

Furthermore, if the brightness temperature scale is referenced to a point outside the atmosphere by multiplying the measurements of brightness temperature [see Eq. (13.48)] by e^{τ_ν}, then the effective system temperature is $T_S e^{\tau_\nu}$, or

$$T_S' = T_R e^{\tau_\nu} + T_{\text{at}}(e^{\tau_\nu} - 1). \tag{13.50}$$

In effect, the atmospheric loss is modeled by an equivalent attenuator at the receiver input. Suppose that $T_R = 30$ K, $T_{\text{at}} = 290$ K, and $\tau_\nu = 0.2$; then the effective system temperature is 100 K. In such a situation the atmosphere would degrade the system sensitivity by a factor of more than 3. Note that the loss in sensitivity results primarily from the increase in system temperature rather than from the attenuation of the signal, which is only 20%. The emission from the atmosphere induces signals in spaced antennas that are uncorrelated and thus contributes only to the noise in the output of an interferometer.

The absorption can be estimated directly from measurements made with a radio telescope. In one technique, called the tipping-scan method, the opacity is determined from the atmospheric emission. If the antenna is scanned from the zenith to the horizon, the observed brightness temperature, in the absence of background sources, will depend on the zenith angle, since the opacity is proportional to the path length through the atmosphere, which varies approximately as $\sec z$. Thus, the atmospheric brightness temperature is

$$T_B = T_{\text{at}}(1 - e^{-\tau_0 \sec z}), \tag{13.51}$$

where τ_0 is the zenith opacity. τ_0 is the negative of the slope of the curve of $\ln(T_{\text{at}} - T_B)$ plotted against $\sec z$, since

$$\ln\left(1 - \frac{T_B}{T_{\text{at}}}\right) = -\tau_0 \sec z. \tag{13.52}$$

The accuracy of this method is affected by ground pickup through the sidelobes, which varies as a function of zenith angle. The opacity can also be estimated from measurements of the absorption suffered by a radio source over a range of zenith angles. The observed antenna temperature on-source minus the antenna temperature off-source at the same zenith angle is

$$\Delta T_B = T_{S0} e^{-\tau_0 \sec z}, \tag{13.53}$$

where T_{S0} is the component of antenna temperature due to the source in the absence of the atmosphere. From Eq. (13.53)

$$\ln \Delta T_B - \ln T_{S0} = -\tau_0 \sec z. \tag{13.54}$$

Thus, τ_0 can be found without knowledge of T_{S0} if a sufficient range in $\sec z$ is covered. This method is affected by changes in antenna gain as a function of zenith angle.

Another technique, called the chopper-wheel method, is commonly used at millimeter wavelengths. A wheel consisting of alternate open and absorbing sections is placed in front of the feed horn. As the wheel rotates, the radiometer alternately views the sky and the absorbing sections, and synchronously measures the difference in temperature between the sky and the chopper wheel at temperature T_0. Thus, the on-source and off-source antenna temperatures are

$$\Delta T_{\text{on}} = T_{S0}e^{-\tau_\nu} + T_{\text{at}}(1 - e^{-\tau_\nu}) - T_0 \tag{13.55}$$

and

$$\Delta T_{\text{off}} = T_{\text{at}}(1 - e^{-\tau_\nu}) - T_0. \tag{13.56}$$

These measurements can be combined to obtain T_{S0} and thereby eliminate the effect of atmospheric absorption. In the case where $T_0 = T_{\text{at}}$,

$$T_{S0} = \left(\frac{\Delta T_{\text{off}} - \Delta T_{\text{on}}}{\Delta T_{\text{off}}}\right) T_0. \tag{13.57}$$

When sensitivity is critical, the chopper wheel is used only to calibrate the output in the off-source position. $\Delta T_{\text{off}} - \Delta T_{\text{on}}$ in the numerator of Eq. (13.57) is then replaced by $T_{\text{off}} - T_{\text{on}}$. Measurement of T_{S0} provides the flux density of the source, which determines the visibility at the origin of the (u, v) plane.

The opacity can be estimated also from surface meteorological measurements when other data are not available. This method is not as accurate as the direct radiometric measurement techniques described above, but has the advantage of not expending observing time. Waters (1976) has analyzed data on absorption versus surface water vapor density for a sea-level site at various frequencies by fitting them to an equation of the form $\tau_0 = \alpha_0 + \alpha_1 \rho_{V0}$. The coefficients α_0 and α_1 are listed in Table 13.1.

TABLE 13.1 Empirical Coefficients for Estimating Opacity from Surface Absolute Humidity[a]

ν (GHz)	α_0 (nepers)	α_1 (nepers m^3 g^{-1})
15	0.013	0.0009
22.2	0.026	0.011
35	0.039	0.0030
90	0.039	0.0090

Source: Waters (1976).

[a]From the equation $\tau_0 = \alpha_0 + \alpha_1 \rho_{V0}$ fitted to opacity data derived from radiosonde measurements and measurements of surface absolute humidity, ρ_{V0} g m^{-3}.

Origin of Refraction

For practical reasons, we have discussed separately the effects of the propagation delay and the absorption in the neutral atmosphere. However, the delay and the absorption are intimately related because they are derived from the real and imaginary parts of the dielectric constant of the gas in the atmosphere. The real and imaginary parts of the dielectric constant are not independent but are related by the Kramers–Kronig relation, which is similar to the Hilbert transform (Van Vleck, Purcell, and Goldstein 1951). We now discuss this relationship from the physical viewpoint of the classical theory of dispersion. From this analysis it will become clear why the atmospherically induced delay is essentially independent of frequency, even in the vicinity of spectral lines that cause significant absorption.

A dilute gas of molecules can be modeled as bound oscillators. In each molecule an electron with mass m and charge $-e$ is harmonically bound to the nucleus, and the electron's motion is characterized by a resonance frequency ν_0 and damping constant $2\pi\Gamma$. The equation of motion with a harmonic driving force $-eE_0e^{-j2\pi\nu t}$ caused by the electric field of an electromagnetic wave is

$$m\ddot{x} + 2\pi m\Gamma\dot{x} + 4\pi^2 m\nu_0^2 x = -eE_0e^{-j2\pi\nu t}, \tag{13.58}$$

where x is the displacement of the bound electron, E_0 and ν are the amplitude and frequency of the applied electric field, and the dots denote time derivatives. The steady-state solution has the form $x = x_0e^{-j2\pi\nu t}$, where

$$x_0 = \frac{eE_0/4\pi^2 m}{\nu^2 - \nu_0^2 + j\nu\Gamma}. \tag{13.59}$$

The magnitude of the dipole moment per unit volume, $\mathbf{P}$, is equal to $-n_m e x_0$, where n_m is the density of gas molecules. The dielectric constant* ε is equal to $1 + \mathbf{P}/(\epsilon_0\mathbf{E})$, so that

$$\varepsilon = 1 - \frac{n_m e^2/4\pi^2 m\epsilon_0}{\nu^2 - \nu_0^2 + j\nu\Gamma}. \tag{13.60}$$

This classical model predicts neither the resonance frequency nor the absolute amplitude of the oscillation. A full treatment of the problem requires the application of quantum mechanics. The proper quantum-mechanical calculation for a system with many resonances yields a result that closely resembles Eq. (13.60) [e.g., Loudon (1983)]:

*In this section and in Section 13.3 we use SI (System International) units, also known as rationalized MKS units. In this system the constitutive relation between the displacement vector $\mathbf{D}$, the electric field vector $\mathbf{E}$, and the polarization vector $\mathbf{P}$ is $\mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P} = \epsilon\mathbf{E}$, where ϵ_0 is the permittivity of free space and ϵ is the permittivity of the medium. The dielectric constant ε is ϵ/ϵ_0. A comparison of various systems of units and equations in electricity and magnetism can be found in Jackson (1999).

$$\varepsilon = 1 - \frac{n_m e^2}{4\pi^2 m\epsilon_0} \sum_i \frac{f_i}{\nu^2 - \nu_{0i}^2 + j\nu\Gamma_i}, \tag{13.61}$$

where f_i is the so-called oscillator strength of the ith resonance. The f_i values obey the sum rule, $\sum f_i = 1$.

The dielectric constant ($\varepsilon = \varepsilon_R + j\varepsilon_I$) and index of refraction ($n = n_R + jn_I$) are connected by Maxwell's relation:

$$n^2 = \varepsilon. \tag{13.62}$$

Thus, $\varepsilon_R = n_R^2 - n_I^2$ and $\varepsilon_I = 2n_I n_R$. Since for a dilute gas $n_R \simeq 1$ and $n_I \ll 1$, we have $n_R \simeq \sqrt{\varepsilon_R}$ and $n_I \simeq \varepsilon_I/2$. Therefore, for a gas with a single resonance

$$n_R \simeq 1 - \frac{n_m e^2(\nu^2 - \nu_0^2)/8\pi^2 m\epsilon_0}{(\nu^2 - \nu_0^2)^2 + \nu^2\Gamma^2} \tag{13.63}$$

and

$$n_I \simeq \frac{n_m e^2 \nu\Gamma/8\pi^2 m\epsilon_0}{(\nu^2 - \nu_0^2)^2 + \nu^2\Gamma^2}. \tag{13.64}$$

The resonance is usually sharp, that is, $\Gamma \ll \nu_0$, and the expressions for n_R and n_I can be simplified by considering their behavior in the vicinity of the resonance frequency ν_0, in which case

$$\nu^2 - \nu_0^2 = (\nu + \nu_0)(\nu - \nu_0) \simeq 2\nu_0(\nu - \nu_0). \tag{13.65}$$

Thus

$$n_R \simeq 1 - \frac{2b(\nu - \nu_0)}{(\nu - \nu_0)^2 + \Gamma^2/4}, \tag{13.66}$$

and

$$n_I \simeq \frac{b\Gamma}{(\nu - \nu_0)^2 + \Gamma^2/4}, \tag{13.67}$$

where $b = n_m e^2/32\pi^2 m\epsilon_0\nu_0$.

Equation (13.67) defines an unnormalized Lorentzian profile for n_I that is symmetric about frequency ν_0 and has a full width at half maximum of Γ and a peak amplitude of $4b/\Gamma$. The function $n_R - 1$ is antisymmetric about frequency ν_0 and has extreme values of $\pm 2b/\Gamma$ at frequencies $\nu_0 \pm \Gamma/2$, respectively. The functions n_R and n_I are plotted in Fig. 13.7. Note that the peak deviation from unity in the real part of the index of refraction, Δn, is equal to one-half the peak value of n_I,

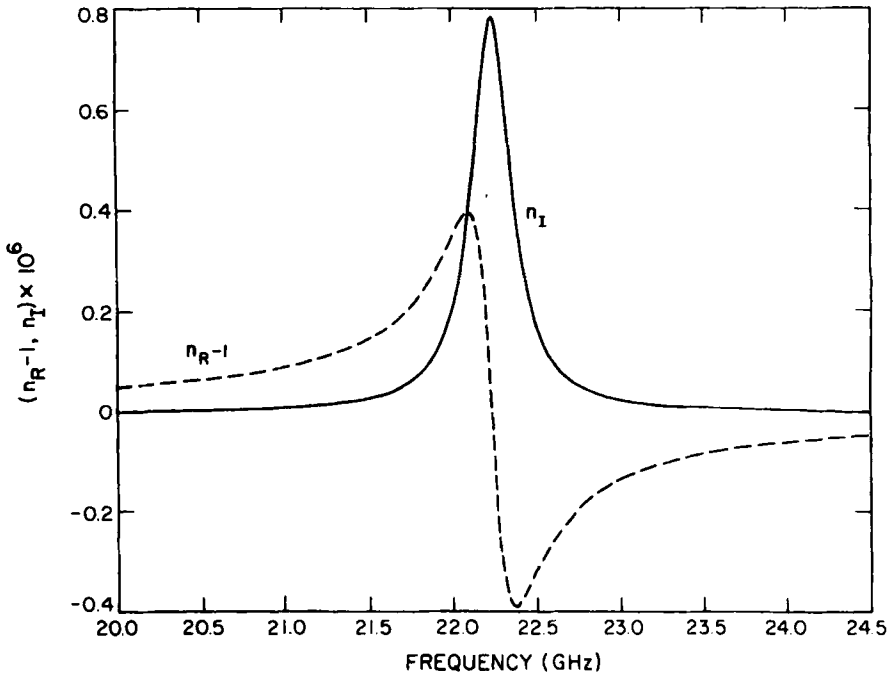


Figure 13.7 Real and imaginary parts of the index of refraction versus frequency for a single resonance given by Eqs. (13.63) and (13.64). The case shown is for the 6_{16}–5_{23} transition in *pure* water vapor with $\rho_V = 7.5$ g m^{-3}. In the atmosphere at the standard sea-level pressure of 1013 mb, the line is broadened to about 2.6 GHz (Liebe 1969). For the curve $n_R - 1$ the peak deviation is Δn [see Eq. (13.68)] and the change in level passing through the line is δn [see Eq. (13.69)].

denoted n_{Imax}. Thus, from Eq. (13.2) we see that the peak absorption coefficient, $\alpha_m = 4\pi n_{\text{Imax}} \nu_0/c$, is related to Δn by the formula

$$\Delta n = \frac{\alpha_m \lambda_0}{8\pi}, \tag{13.68}$$

where λ_0 is the wavelength of the resonance, c/ν_0. The magnitude of the real part of the index of refraction is equal to the peak absorption over a distance of $\lambda_0/8\pi$. In addition, Eq. (13.66) shows that the real part of the index of refraction is not exactly symmetric about ν_0; that is, n_R tends to unity as ν tends to ∞, and n_R tends to $1 + 2b/\nu_0 = 1 + \Delta n\Gamma/\nu_0 = 1 + (\lambda_0\alpha_m/8\pi)(\Gamma/\nu_0)$ as ν tends to zero. Hence, the change δn in the asymptotic value of the index of refraction on passing through a resonance is given by

$$\delta n = \frac{\alpha_m \Gamma \lambda_0^2}{8\pi c}. \tag{13.69}$$

Thus, $\delta n/\Delta n = \gamma/\nu_0$, but unless the resonance is extremely strong, Δn and δn are both negligible. Consider the 22-GHz water vapor line. The attenuation in the atmosphere when $\rho_V = 7.5$ g m^{-3} is 0.15 dB km^{-1}, so $\alpha_m = 3.5 \times 10^{-7}$ cm^{-1}. Equation (13.68) then predicts that $\Delta n = 1.9 \times 10^{-8}$, or $\Delta N = 0.019$, which agrees with the value measured in the laboratory (Liebe 1969). For the same value of ρ_V, the contribution of all transitions of water vapor to the value of the index of refraction at low frequencies ($10^{-6}N_V$), from Eq. (13.10), is equal to 4.4×10^{-5}. Thus, the fractional change in refractivity near the 22 GHz line is only 1 part in 2500. The change in asymptotic level is even smaller. At sea level $\Gamma = 2.6$ GHz and $\delta n = 2.2 \times 10^{-8}$. The water vapor line at 557 GHz (the 1_{10}–1_{01} transition) has an absorption coefficient of 29,000 dB km^{-1}, or 0.069 cm^{-1}. The values of Δn and δn are 1.44×10^{-6} and 0.7×10^{-6}, respectively. In the atmospheric windows above 400 GHz, where radio astronomical observations are possible only from very dry sites, the refractive index can be noticeably different from the value at lower frequencies. The normalized refractivity is shown in Fig. 13.8.

Equation (13.68) is an important result of very general validity. We derived it from a specific model [Eq. (13.58)] that led to an approximately Lorentzian profile for the absorption spectrum. In practice, line profiles are found to differ slightly from the Lorentzian form, and more sophisticated models are needed to fit them exactly. However, Eqs. (13.68) and (13.69) could be derived from the Kramers–Kronig relation.

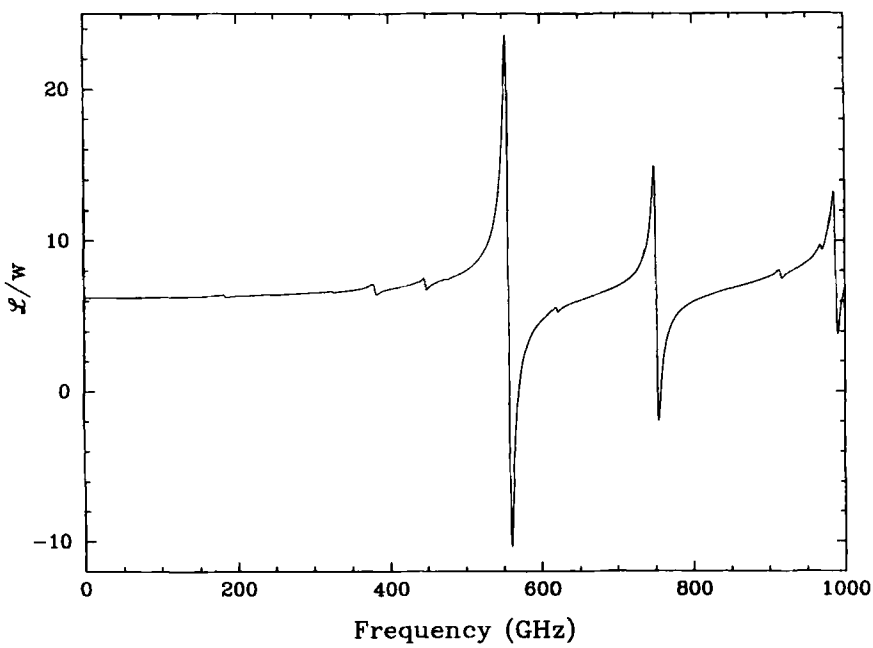


Figure 13.8 The predicted excess path length due to water vapor per unit column density versus frequency, from formulas by Liebe (1989). From Sutton and Hueckstaedt (1996), courtesy of *Astron. Astrophys. Suppl.*

The low-frequency value of the index of refraction, as given by the Smith–Weintraub equation [Eq. (13.9)], results from the contributions of all transitions at higher frequencies. Summing the contributions [see Eq. (13.69)] of many lines, each characterized by parameters Δn_1, Γ_i, α_{mi}, and ν_{0i}, we obtain the low-frequency value of the index of refraction:

$$n_s = 1 + \sum_i \frac{\alpha_{mi}\lambda_{0i}^2\Gamma_i}{8\pi c} = 1 + \sum_i \frac{\Delta n_i \Gamma_i}{\nu_{0i}}. \tag{13.70}$$

The water vapor molecule has a large number of strong rotational transitions in the band from 30 μm to 0.3 mm (from 10 THz to 1000 GHz). The atmosphere is opaque through most of this region because of these lines, which contribute about 98% of the low-frequency refractivity. The remainder comes from the 557-GHz line. The refractivity due to water vapor is small in the optical region and is greater by a factor of 22 in the radio region. Therefore, whereas the effects of water vapor are small in the optical region, they are very important in the radio region. The dry-air refractivity, due primarily to resonances of oxygen and nitrogen in the ultraviolet, is nearly the same in the optical and radio regions.

Smith–Weintraub Equation

Detailed discussions of the radio refractivity equation can be found in Bean and Dutton (1966), Thayer (1974), and Hill, Lawrence, and Priestley (1982). From the classic work of Debye (1929), it can be shown that the refractivity of molecules with induced dipole transitions varies as pressure and T^{-1}, and the refractivity of molecules with permanent dipole moments varies as pressure and T^{-2}. The principal constituents of the atmosphere, oxygen molecules, O_2, and nitrogen molecules, N_2, being homonuclear, have no permanent electric dipole moments. However, molecules such as H_2O and other minor trace constituents have permanent dipole moments. Thus, the general form of the refractivity equation is

$$N = \frac{K_1 p_D}{T\mathcal{Z}_D} + \frac{K_2 p_V}{T\mathcal{Z}_V} + \frac{K_3 p_V}{T^2\mathcal{Z}_V}, \tag{13.71}$$

where p_D and p_V are the partial pressures of the dry air and water vapor; K_1, K_2, and K_3 are constants; and $\mathcal{Z}_D$ and $\mathcal{Z}_V$ are compressibility factors for dry-air gases and water vapor, which correct for non-ideal gas behavior and deviate from unity in atmospheric conditions by less than 1 part in 10^3. These compressibility factors are given by the equations (Owens 1967)

$$\mathcal{Z}_D^{-1} = 1 + p_D\left[57.90 \times 10^{-8}\left(1 + \frac{0.52}{T}\right) - 9.4611 \times 10^{-4}\frac{(T - 273)}{T^2}\right] \tag{13.72a}$$

and

$$\mathcal{Z}_V^{-1} = 1 + 1650\frac{p_V}{T^3}[1 - 0.01317(T - 273) + 1.75 \times 10^{-4}(T - 273)^2 + 1.44 \times 10^{-6}(T - 273)^3], \tag{13.72b}$$

where p_D and p_V are in millibars. The first and second terms in Eq. (13.71) are due to ultraviolet electronic transitions of the induced dipole type for dry-air molecules and water vapor, respectively, and the third term is due to the permanent dipole infrared rotational transitions of water vapor. If we neglect terms other than unity in the $\mathcal{Z}$ factors, Eq. (13.71) becomes

$$N = 77.6\frac{p_D}{T} + 64.8\frac{p_V}{T} + 3.776 \times 10^5\frac{p_V}{T^2}. \tag{13.73}$$

We can rewrite Eq. (13.73) in terms of the total pressure as

$$N = 77.6\frac{P}{T} - 12.8\frac{p_V}{T} + 3.776 \times 10^5\frac{p_V}{T^2}. \tag{13.74}$$

For temperatures around 280 K, the last two terms on the right-hand side of Eq. (13.74) can be combined to give

$$N \simeq \frac{77.6}{T}\left(P + 4810\frac{p_V}{T}\right). \tag{13.75}$$

Equation (13.75) is the well-known *Smith–Weintraub equation* (Smith and Weintraub 1953). This equation is accurate to about 1%, or about $\pm 1N$–unit, at frequencies below 100 GHz. The accuracy of Eqs. (13.74) and (13.75) can be improved by adding a small term that increases monotonically with frequency to account for the effect of the wings of the infrared transitions (see Fig. 13.8). Hill and Clifford (1981) show that because of this effect the wet refractivity increases by about 0.5% at 100 GHz, and 2% at 200 GHz, over its value at low frequencies.

To obtain the optical refractivity, we omit the permanent dipole term from Eq. (13.73) and obtain

$$N_{\text{opt}} \simeq 77.6\frac{p_D}{T} + 64.8\frac{p_V}{T}. \tag{13.76}$$

For precise work Cox (2000) provides more accurate values for N_{opt} that include small terms having wavelength dependence to account for the effects of the wings of ultraviolet transitions. The ratio of the wet refractivity in the radio and optical regions is obtained by omitting the dry-air terms from Eqs. (13.73) and (13.76): $N_{V\text{rad}}/N_{V\text{opt}} \simeq 1 + 5830/T$. For $T \simeq 280$ K, this ratio is about equal to 22, as mentioned in connection with the discussion following Eq. (13.70).

Phase Fluctuations

In the radio region, the most important nonuniformly distributed quantity in the troposphere is the water vapor density. Variations in water vapor distribution in the troposphere that move across an interferometer cause phase fluctuations that degrade the measurements. In the optical region, variations in temperature, rather than in water vapor content, are the principal cause of phase fluctuations. The situation is depicted in Fig. 13.9. A critical dimension is the size of the first Fresnel zone, $\sqrt{\lambda h}$, where h is the distance between the observer and the screen. For $\lambda = 1$ cm and $h = 1$ km, the Fresnel scale is about 3 m. The atmospherically induced phase fluctuations on this scale are very small ($\ll 1$ rad). In this case the phase fluctuation can cause image distortion but not amplitude fluctuation (i.e., scintillation). This is known as the regime of weak scattering. Plasma scattering in the interstellar medium belongs to the regime of strong scattering, where the phenomena are considerably more complex.

The fluctuations along an initially plane wavefront that has traversed the atmosphere can be characterized by a so-called structure function of the phase. This function is defined as

$$\mathcal{D}_\phi(d) = \langle[\Phi(x) - \Phi(x - d)]^2\rangle, \tag{13.77}$$

where $\Phi(x)$ is the phase at point x, $\Phi(x - d)$ is the phase at point $x - d$, and the angle brackets indicate an ensemble average. We assume that $\mathcal{D}_\phi$ depends

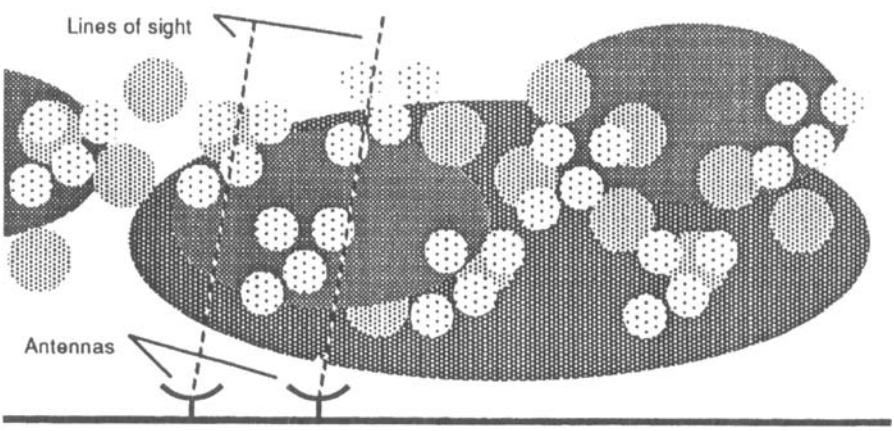


Figure 13.9 A cartoon of a two-element interferometer beneath a tropospheric screen of water vapor irregularities of various sizes. The screen moves over the interferometer at a velocity component v_s parallel to the baseline. The distribution of these irregularities is important in designing the phase compensation schemes discussed in Section 13.2. Note that fluctuations with scale sizes larger than the baseline cover both antennas and do not affect the interferometer phase significantly. From Masson (1994a), courtesy of the Astron. Soc. Pacific Conf. Ser.

only on the magnitude of the separation between the measurement points, that is, the projected baseline length d of the interferometer. The rms deviation in the interferometer phase is

$$\sigma_\phi = \sqrt{\mathcal{D}_\phi(d)}. \tag{13.78}$$

For the sake of illustration, we assume a simple functional form for σ_ϕ given by

$$\sigma_\phi = \frac{2\pi a d^\beta}{\lambda}, \qquad d \le d_m \tag{13.79a}$$

and

$$\sigma_\phi = \sigma_m, \qquad d > d_m, \tag{13.79b}$$

where a is constant, and $\sigma_m = 2\pi a d_m^\beta/\lambda$. The form of Eqs. (13.79) is shown in Fig. 13.10a. This form can be derived by assuming a multiple-scale power-law

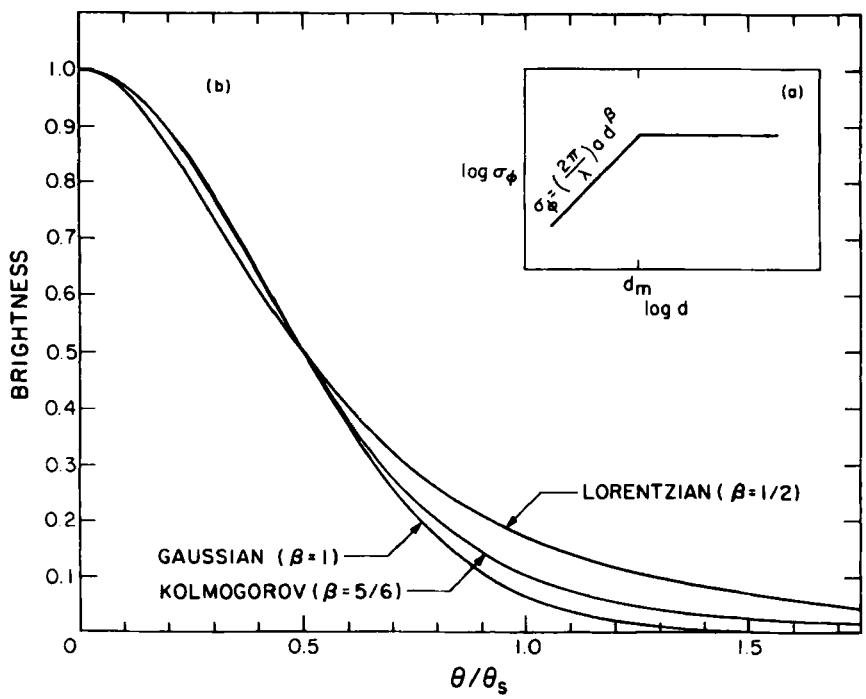


Figure 13.10 (a) Simple model for the rms phase fluctuation induced by the troposphere in an interferometer of baseline length d given by Eqs. (13.79). (b) The point-source response function $\overline{w}_a(\theta)$ for various power-law models obtained by taking the Fourier transform of the visibility in the regime $d < d_m$. The values of θ_s, the full width at half maximum of $\overline{w}_a(\theta)$, for each model are: Gaussian ($\beta = 1$), $\sqrt{8\ln 2}\,a$; modified Lorentzian ($\beta = \frac{1}{2}$), $1.53\pi\lambda^{-1}a^2$; Kolmogorov ($\beta = \frac{5}{6}$), $2.75\lambda^{-1/5}a^{6/5}$. λ is the wavelength and a is the constant defined in Eq. (13.79a).

model for the spectrum of the phase fluctuations. There must be a limiting distance d_m beyond which fluctuations do not increase noticeably; otherwise VLBI would not work. This limit, called the outer scale length of the fluctuations, is probably about a few kilometers. Beyond this dimension the fluctuations in the path lengths become uncorrelated.

First, consider an interferometer that operates in the domain of baselines shorter than d_m. The measured visibilities $\mathcal{V}_m$ are related to the true visibilities by the equation

$$\mathcal{V}_m = \mathcal{V}e^{j\phi}, \tag{13.80}$$

where $\phi = \Phi(x) - \Phi(x-d)$ is a random variable describing the phase fluctuations introduced by the atmosphere. If we assume ϕ is a Gaussian random variable with zero mean, then the expectation of the visibility is

$$\langle \mathcal{V}_m \rangle = \mathcal{V}\langle e^{j\phi} \rangle = \mathcal{V}e^{-\sigma_\phi^2/2} = \mathcal{V}e^{-\mathcal{D}_\phi/2}. \tag{13.81}$$

Consider the conceptually useful case where $\beta = 1$. It would arise in an atmosphere consisting of inhomogeneous wedges of scale size larger than the baseline. In this case σ_ϕ is proportional to d, and the constant a is dimensionless. Substituting Eq. (13.79a) into Eq. (13.81) yields

$$\langle \mathcal{V}_m \rangle = \mathcal{V}e^{-2\pi^2 a^2 q^2}, \tag{13.82}$$

where $q = \sqrt{u^2 + v^2} = d/\lambda$. On average, therefore, the measured visibility is the true visibility multiplied by an atmospheric weighting function $w_a(q)$ given by

$$w_a(q) = e^{-2\pi^2 a^2 q^2}. \tag{13.83}$$

In the image plane, the derived map is the convolution of the true source distribution and the Fourier transform of $w_a(q)$, which is

$$\overline{w}_a(\theta) \propto e^{-\theta^2/2a^2}, \tag{13.84}$$

where θ is here the conjugate variable of q. The full width at half maximum of $\overline{w}_a(\theta)$ is θ_s, given by

$$\theta_s = \sqrt{8 \ln 2}\, a. \tag{13.85}$$

Thus, the resolution is degraded because the derived map is convolved with a Gaussian beam of width θ_s (in addition to the effects of any other weighting functions, as described in Section 10.2 under *Weighting of the Visibility Data*). θ_s is the seeing angle. Images with finer resolution than θ_s can sometimes be obtained by use of adaptive calibration procedures described in Section 11.4. Now, from

Eq. (13.79a), we obtain

$$a = \frac{\sigma_\phi \lambda}{2\pi d} = \frac{\sigma_d}{d}, \tag{13.86}$$

where $\sigma_d = \sigma_\phi \lambda / 2\pi$ is the rms uncertainty in path length. Thus we obtain

$$\theta_s = 2.35 \frac{\sigma_d}{d} \text{ (radians)}. \tag{13.87}$$

Since σ_d/d is constant, θ_s is independent of wavelength. This independence results from the condition $\beta = 1$ in Eq. (13.79a). For the radio regime, σ_d is about 1 mm on a baseline of 1 km, so $a \simeq 10^{-6}$, and $\theta_s \simeq 0.5$ arcsec. Let d_0 be the baseline length for which $\sigma_\phi = 1$ rad. From Eq. (13.86) we see that Eq. (13.85) can be written in the form

$$\theta_s = \frac{\sqrt{2 \ln 2}}{\pi} \frac{\lambda}{d_0} \simeq 0.37 \frac{\lambda}{d_0}. \tag{13.88}$$

For the case where β is arbitrary, we find $\overline{w}_a(\theta)$ by substituting Eq. (13.79a) into Eq. (13.81) and writing the two-dimensional Fourier transform as a Hankel transform (Bracewell 2000). Thus

$$\overline{w}_a(\theta) \propto \int_0^\infty \exp\left[-2\pi^2 a^2 \lambda^{2(\beta-1)} q^{2\beta}\right] J_0(2\pi q \theta) q \, dq, \tag{13.89}$$

where J_0 is the Bessel function of order zero and a has dimensions $\text{cm}^{(1-\beta)}$. In general, $\overline{w}_a(\theta)$ cannot be evaluated analytically. However, by making appropriate substitutions in Eq. (13.89), it is easy to show that $\theta_s \propto a^{1/\beta} \lambda^{(\beta-1)/\beta}$. A case that can be treated analytically is the one for which $\beta = 1/2$. In this case we obtain (Bracewell 2000, p. 338)

$$\overline{w}_a(\theta) \propto \frac{1}{\left[\theta^2 + (\pi a^2/\lambda)^2\right]^{3/2}}, \tag{13.90}$$

which represents a Lorentzian profile raised to the 3/2 power and has very broad skirts. The full width at half maximum of $\overline{w}(\theta)$ is

$$\theta_s = \frac{1.53 \pi a^2}{\lambda}, \tag{13.91}$$

or

$$\theta_s = \frac{0.77}{2\pi} \frac{\lambda}{d_0} \simeq 0.12 \frac{\lambda}{d_0}. \tag{13.92}$$

In the case of Kolmogorov turbulence, which is discussed later in this section, $\beta = 5/6$. Numerical integration of Eq. (13.89) yields

$$\theta_s \simeq 2.75 a^{6/5} \lambda^{-1/5} \simeq 0.30 \frac{\lambda}{d_0}. \tag{13.93}$$

Plots of $\overline{w}_a(\theta)$ for various power-law models of phase fluctuations are shown in Fig. 13.10b.

Now consider the case of an interferometer operating in the domain of baselines greater than d_m, where σ_ϕ is a constant equal to σ_m. This case is most applicable to VLBI arrays or to large connected-element arrays. If the timescale of the fluctuation is short with respect to the measurement time, then, on average, all the visibility measurements are reduced by a constant factor $e^{-\sigma_m^2/2}$. Thus, this type of atmospheric fluctuation does not reduce the resolution. However, on average the measured flux density is reduced from the true value by the factor $e^{-\sigma_m^2/2}$. If the timescale of the fluctuations is long with respect to the measurement time, then each visibility measurement suffers a phase error $e^{j\phi}$. Assume that K visibility measurements are made of a point source of flux density S. The map of the source, considering only one dimension for simplicity, is

$$\overline{w}_a(\theta) = \frac{S}{K}\sum_{i=1}^{K} e^{j\phi_i} e^{j2\pi u_i \theta}. \tag{13.94}$$

The expectation of $\overline{w}_a(\theta)$ at $\theta = 0$ is

$$\langle \overline{w}_a(0)\rangle = S e^{-\sigma_m^2/2}. \tag{13.95}$$

The measured flux density is less than S. The missing flux density is scattered around the map. This is immediately evident from Parseval's theorem:

$$\sum_i |\,\overline{w}_a(\theta_i)\,|^2 = \frac{1}{K}\sum_i |\mathcal{V}(u_i)|^2 = S^2. \tag{13.96}$$

Thus, the total flux density could be obtained by integrating the square of the image-plane response. The rms deviation in the flux density, measured at the peak response for a source at $\theta = 0$, is $\sqrt{\langle \overline{w}_a^2(\theta)\rangle - \langle \overline{w}_a(\theta)\rangle^2}$, which we call σ_S. This quantity can be calculated from Eq. (13.94) and is given by

$$\sigma_S = \frac{S}{\sqrt{K}}\sqrt{1 - e^{-\sigma_m^2}}, \tag{13.97}$$

which reduces to $\sigma_S \simeq S\sigma_m/\sqrt{K}$ when $\sigma_m \ll 1$.

Kolmogorov Turbulence

The theory of propagation through a turbulent neutral atmosphere has been treated in detail in the seminal publications of Tatarski (1961, 1971). This theory has been developed and applied extensively to problems of optical seeing [e.g., Roddier (1981), Woolf (1982), Coulman (1985)] and to infrared interferometry (Sutton, Subramanian, and Townes 1982). We confine the discussion here to a few central ideas concerning the structure function of phase, and indicate how it is related to other functions that are used to characterize atmospheric turbulence.

When the Reynolds number (a dimensionless parameter that involves the viscosity, a characteristic scale size, and the velocity of a flow) exceeds a critical value, the flow becomes turbulent. In the atmosphere the Reynolds number is nearly always high enough that turbulence is fully developed. In the Kolmogorov model for turbulence, the kinetic energy associated with large-scale turbulent motions is transferred to smaller and smaller scale sizes of turbulence until it is finally dissipated into heat by viscous friction. If the turbulence is fully developed and isotropic, then the two-dimensional power spectrum of the phase fluctuations (or the refractive index) varies as $q_s^{-11/3}$, where q_s (cycles per meter) is the spatial frequency (q_s, the conjugate variable of d, is analogous to q, the conjugate variable of θ). The structure function for the refractive index $\mathcal{D}_n(d)$ is defined in a fashion similar to the structure function of phase in Eq. (13.77); that is, $\mathcal{D}_n(d)$ is the mean-square deviation of the difference in the refractive index at two points a distance d apart, or $\mathcal{D}_n(d) = \langle [n(x) - n(x-d)]^2 \rangle$. For the conditions stated above, $\mathcal{D}_n$ can be shown to be given by the equation

$$\mathcal{D}_n(d) = C_n^2 d^{2/3}, \qquad L_{\text{inner}} \ll d \ll L_{\text{outer}}, \tag{13.98}$$

where L_{inner} and L_{outer} are called the inner and outer scales of turbulence, which may be less than a centimeter and a few kilometers, respectively. The parameter C_n^2 characterizes the strength of the turbulence. Note that water vapor, which is the dominant cause of fluctuation in the index of refraction, is poorly mixed in the troposphere and therefore may be only an approximate tracer of the mechanical turbulence.

The structure function of phase for an atmosphere where C_n^2 varies with height from the surface to an overall height L is given by [Tatarski 1961, Eq. (6.65)]

$$\mathcal{D}_\phi(d) = 2.91\left(\frac{2\pi}{\lambda}\right)^2 d^{5/3} \int_0^L C_n^2(h)\, dh, \tag{13.99}$$

which is valid in the range $\sqrt{L\lambda} < d < L_{\text{outer}}$. Note that the factor 2.91 is a dimensionless constant and C_n^2 has units of length$^{-2/3}$. The lower limit on d is equivalent to the requirement that diffraction effects be negligible. If C_n^2 is constant with height, then Eq. (13.99) reduces to

$$\mathcal{D}_\phi(d) = 2.91\left(\frac{2\pi}{\lambda}\right)^2 C_n^2 L\, d^{5/3}. \tag{13.100}$$

Thus, from Eq. (13.78) the rms phase deviation is

$$\sigma_\phi = 1.71\left(\frac{2\pi}{\lambda}\right)\sqrt{C_n^2 L}\, d^{5/6}. \tag{13.101}$$

We can calculate d_0, the baseline length for which $\sigma_\phi = 1$ rad, by setting $\mathcal{D}_\phi$ in Eq. (13.100) equal to 1 rad^2. The expression for d_0 is then

$$d_0 = 0.058\lambda^{6/5}(C_n^2 L)^{-3/5}. \tag{13.102}$$

Another scale length that is proportional to d_0 is the Fried length, r_0† (Fried 1966). This scale is particularly useful for discussion of the effects of turbulence in telescopes with circular apertures and is widely used in the the optical literature. The structure function of phase can be written $D_\phi = 6.88(d/r_0)^{5/3}$, where the factor 6.88 is an approximation of $2[(24/5)\Gamma(6/5)]^{5/6}$ (Fried 1967). Hence, from (13.100) and (13.102), $r_0 = 3.18\,d_0$. The Fried length is defined such that the effective collecting area of a large circular aperture with uniform illumination in the presence of Kolmogorov turbulence is $\pi r_0^2/4$. Hence for an aperture of a diameter small with respect to r_0 the resolution is dominated by diffraction at the aperture. With an aperture large with respect to r_0 the resolution is set by the turbulence and is approximately λ/r_0. The exact resolution in this latter case can be derived from Eq. (13.93), with the result $\theta_s = 0.97\lambda/r_0$. In addition, the rms phase error over an aperture of diameter r_0 is 1.01 rad. The reason that r_0 is larger than d_0 is related to the downweighting of long baselines in two-dimensional apertures [see Eq. (14.13) and related discussion]. For an aperture of diameter r_0, the ratio of the collecting area to the geometric area, which is called the Strehl ratio in the optical literature, is equal to 0.45 (Fried 1965).

Equation (13.102) shows that d_0 is proportional to $\lambda^{6/5}$, and thus the angular resolution or seeing limit ($\sim\lambda/d_0$) is proportional to $\lambda^{-1/5}$ [see Fig. 13.10 and Eq. (13.93)]. This relationship may hold over broad wavelength ranges when C_n^2 is constant. In the optical range C_n^2 is related to temperature fluctuations, whereas in the radio range C_n^2 is dominated by turbulence in the water vapor. It is an interesting coincidence that the seeing angle is about 1 arcsec at both optical and radio wavelengths, for good sites. The important difference is the timescale of fluctuations, τ_c. If the critical level of fluctuation is 1 radian, then $\tau_c \simeq d_0/v_s$, where v_s is the velocity component of the screen parallel to the baseline. Any adaptive optics compensation must operate on a timescale short with respect to τ_c. From Eq. (13.93), τ_c can be expressed as

$$\tau_c \simeq 0.3\frac{\lambda}{\theta_s v_s}. \tag{13.103}$$

For $v_s = 10$ m s^{-1} and $\theta_s = 1$ arcsec, $\tau_c = 3$ ms at 0.5 μm wavelength and 60 s at 1 cm wavelength.

The two-dimensional power spectrum of phase, $\mathcal{S}_2(q_x, q_y)$, is the Fourier transform of the two-dimensional autocorrelation function of phase, $R_\phi(d_x, d_y)$. If R_ϕ is only a function of d, where $d^2 = d_x^2 + d_y^2$, then $\mathcal{S}_2$ is a function of q_s, where $q_s^2 = q_x^2 + q_y^2$, and $\mathcal{S}_2(q_s)$ and $R_\phi(d)$ form a Hankel transform pair. Since $\mathcal{D}_\phi(d) = 2[R_\phi(0) - R_\phi(d)]$, we can write

$$\mathcal{D}_\phi(d) = 4\pi \int_0^\infty [1 - J_0(2\pi q_s d)]\,\mathcal{S}_2(q_s) q_s\, dq_s, \tag{13.104}$$

†In this paragraph we follow Fried's use of this symbol. Elsewhere in this chapter r_0 represents the radius of the earth.

where J_0 is the Bessel function of order zero. When $\mathcal{D}_\phi(d)$ is given by Eq. (13.100), $\mathcal{S}_2(q_s)$ is

$$\mathcal{S}_2(q_s) = 0.0097\left(\frac{2\pi}{\lambda}\right)^2 C_n^2 L q_s^{-11/3}. \tag{13.105}$$

It is often useful to study temporal variations caused by atmospheric turbulence. In order to relate the temporal and spatial variations, we invoke the frozen-screen hypothesis, sometimes attributed to Taylor (1938). In this approximation the turbulent eddies are assumed to remain fixed during the time it takes for the layer to move across the baseline d. The one-dimensional temporal spectrum of the phase fluctuations $\mathcal{S}'_\phi(f)$ (the two-sided spectrum) can be calculated from $\mathcal{S}_2(q_s)$ by

$$\mathcal{S}'_\phi(f) = \frac{1}{v_s}\int_{-\infty}^{\infty} \mathcal{S}_2\left(q_x = \frac{f}{v_s}, q_y\right) dq_y, \tag{13.106}$$

where v_s is in meters per second. Substitution of Eq. (13.105) into Eq. (13.106) yields

$$\mathcal{S}'_\phi(f) = 0.016\left(\frac{2\pi}{\lambda}\right)^2 C_n^2 L v_s^{5/3} f^{-8/3}\ (\text{rad}^2\ \text{Hz}^{-1}). \tag{13.107}$$

Examples of the temporal spectra of water vapor fluctuations can be found in Hogg, Guiraud, and Sweezy (1981) and Masson (1994a) (see Fig. 13.15). The temporal structure function $\mathcal{D}_\tau(\tau) = \langle[\phi(t) - \phi(t-\tau)]^2\rangle$ is related to the spatial structure function by $\mathcal{D}_\tau(\tau) = \mathcal{D}_\phi(d = v_s\tau)$. Hence, for Kolmogorov turbulence, we obtain from Eq. (13.100)

$$\mathcal{D}_\tau(\tau) = 2.91\left(\frac{2\pi}{\lambda}\right)^2 C_n^2 L v_s^{5/3}\tau^{5/3}. \tag{13.108}$$

$\mathcal{D}_\tau(\tau)$ and $\mathcal{S}'_\phi(f)$ are related by a Fourier transformation. The use of temporal structure functions to estimate the effects of fluctuations on interferometers is discussed by Treuhaft and Lanyi (1987) and by Lay (1997a).

The Allan variance $\sigma_y^2(\tau)$, or fractional frequency stability for time interval τ, associated with $\mathcal{S}'_\phi(f)$ has been defined in Section 9.5 under *Analysis of Phase Fluctuations*. It can be calculated by substituting Eq. (9.99) into Eq. (9.111), which gives

$$\sigma_y^2(\tau) = \left(\frac{2}{\pi\nu_0\tau}\right)^2 \int_0^\infty \mathcal{S}'_\phi(f)\sin^4(\pi\tau f)\, df. \tag{13.109}$$

By substituting Eq. (13.107) into Eq. (13.109), and noting that

$$\int_0^\infty [\sin^4(\pi x)]/x^{8/3} dx = 4.61,$$

TABLE 13.2 Power Law Relations for Turbulence

Quantity			Exponent: 3D Turbulence ($\alpha = 11/3$)	Exponent: 2D Turbulence ($\alpha = 8/3$)
2D, 3D power spectrum	$\mathcal{S}_2(q_s)$, $\mathcal{S}(q_s)$	$-\alpha$	$-11/3$	$-8/3$
Structure function (3D)	$\mathcal{D}_\phi(d)$	$\alpha - 2$	$5/3$	$2/3$
Temporal phase spectrum	$\mathcal{S}'_\phi(f)$	$1 - \alpha$	$-8/3$	$-5/3$
Allan variance	$\sigma_y^2(\tau)$	$\alpha - 4$	$-1/3$	$-4/3$
Temporal structure function	$\mathcal{D}_\tau(\tau)$	$\alpha - 2$	$5/3$	$2/3$

Adapted from Wright (1996, p. 526), with permission from the Astronomical Society of the Pacific.

we obtain

$$\sigma_y^2(\tau) = 1.3 \times 10^{-17} C_n^2 L v_s^{5/3} \tau^{-1/3}. \tag{13.110}$$

Armstrong and Sramek (1982) give general expressions for the relations among $\mathcal{S}_2$, $\mathcal{S}'_\phi$, $\mathcal{D}_\phi$, and σ_y for an arbitrary power-law index. If $\mathcal{S}_2 \propto q^{-\alpha}$, then $\mathcal{D}_\phi(d) \propto d^{\alpha-2}$, $\mathcal{S}'_\phi \propto f^{1-\alpha}$, and $\sigma_y^2 \propto \tau^{\alpha-4}$. These relations are summarized in Table 13.2.

A reasonable model for tropospheric turbulence can be made in the following way. For baselines less than some value d_{trans}, the turbulence is three-dimensional and $\mathcal{D}_\phi \propto d^{5/3}$. d_{trans} is small with respect to the scale height of the water vapor, ~2 km. For $d > d_{\text{trans}}$ the turbulence is two-dimensional and $\mathcal{D}_\phi \propto d^{2/3}$. Beyond some baseline d_{outer}, the atmospheric fluctuations become uncorrelated and $\mathcal{D}_\phi \propto d^0$, in other words, independent of baseline length. The value of d_{outer} is of the order of the size of clouds, a few kilometers (Hamaker 1978). These three regions are clearly evident in the data shown in Fig. 13.11. For this particular data set, $d_{\text{trans}} = 1.2$ km, $d_{\text{outer}} = 6$ km, and the slopes are close to the predictions of Kolmogorov turbulence. Table 13.3 gives a compilation of structure function data from many sources. The data include a variety of observations and atmospheric conditions and do not provide an accurate comparison of the quality of different sites. Nevertheless, they show that, as expected, the fluctuations in delay are independent of frequency and tend to decline with increasing site elevation. For Mauna Kea, the values indicate the night-to-day range of the 50% curve in the lower left panel of Fig. 13.13 in Section 13.2. The values for Chajnantor represent a similar diurnal range. Rogers and Moran (1981) and Rogers et al. (1984) discuss the effects of the troposphere on VLBI measurements, and their plot of the Allan variance of the atmosphere is shown as the curve for VLBI data in Fig. 9.14. A general comparison of the measured and theoretical results is given by Coulman (1990). Other site comparisons can be found in Masson (1994b).

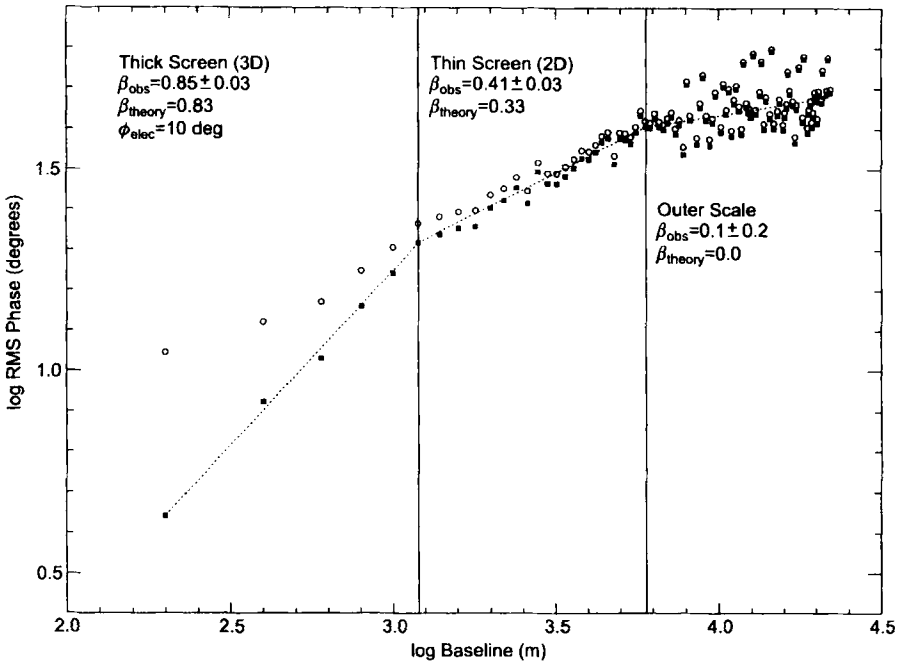


Figure 13.11 The root phase structure function (rms phase) from observations with the VLA at 22 GHz. The open circles show the rms phase variation versus baseline length measured on the source 0748+240 over a period of 90 min. The filled squares show the data after removal of a constant receiver-induced noise component of rms amplitude 10°. The three regimes of the phase structure function are indicated by vertical lines (at 1.2 and 6 km). From Carilli and Holdaway (1999), ©1999 by the American Geophys. Union.

The strength of the phase fluctuations, characterized by the parameter σ_d [defined in Eq. (13.86)], or $C_n^2 L$, is difficult to predict. Measurements at the VLA show that $C_n^2 L$ is not well correlated with surface absolute humidity. The dominant correlation is probably with solar-induced convection. The strong time-of-day dependence of phase stability is described in Section 13.2 (Fig. 13.13). The remarkably good stability of meter-wavelength interferometers under conditions of overcast skies has long been known [see, e.g., Hinder (1972)].

Anomalous Refraction

The beamwidths of many millimeter radio telescopes are sufficiently small that the effects of atmospheric phase fluctuations can be detected. Typically, the apparent positions of unresolved sources have been observed to wander by about 5 arcsec on timescales of a few seconds under certain meteorological conditions [see, e.g., Altenhoff et al. (1987), Downes and Altenhoff (1990)]. The magnitudes of these effects are largely independent of wavelength, as expected, if they are caused by water vapor irregularities in the troposphere. These irregularities may

TABLE 13.3 Structure Function Measurements

Location	Baseline (km)	Altitude (m)	Frequency (GHz)	σ_{d0}^a (mm)	β^b	$10^7\sqrt{C_n^2 L}$ ($m^{1/6}$)	Reference[c]
Cambridge	1.6	17	5	0.7–2.6	1.3	13–50	1
Green Bank	2.5	840	2.7	0.4–4	—	7–70	2
Hat Creek	0.006–0.1	1043	86	0.7–1.0	1.1–1.4	13–18	3
Hat Creek	0.006–0.85	1043	86	0.8–2.2	0.8–1.3	15–41	4
Hat Creek	0.01–0.15	1043	86	1.2	1–2	22	5
Hat Creek	1–1200	1043	100	0.7	0.3–0.6	13	3
NRO	0.035	1350	19	1.9	1.2	35	6
NRO	0.03–0.54	1350	22	0.5–0.9	1.6	9–17	7
VLA Site	0.1–3	2124	22	0.6	0.72	11	8
VLA Site	0.1–35	2124	22	0.65	0.85	12	9
VLA Site	1–35	2124	5	1.0	1.4	18	10
VLA Site	0.05–35	2124	5/15	0.6–1.6	0.6–0.8	11–30	11
Plateau de Bure	0.02–0.3	2552	86	0.3–0.7	1.1–1.9	6–13	12
Mauna Kea[d]	0.1	4070	12	0.4–2.7	0.75	7–49	13
Chajnantor[d]	0.3	5000	11	0.3–1.5	—	5–29	14

Source: Adapted from Wright (1996, p. 524), with permision from the Astronomical Society of the Pacific.

[a] σ_{do} = rms path length deviation on a 1000-m baseline. Hence, $\sigma_d = \sigma_{do}(d/1000\ \text{m})^\beta$. If measurements were not made at $d = 1000$ m, then $\beta = 5/6$ is assumed.

[b] Power-law exponent for baseline dependence of σ_{d0}. For 2D and 3D Kolmogorov turbulence, $\beta = 0.33$ and 0.83, respectively.

[c] *References:* (1) Hinder (1970), Hinder and Ryle (1971), (2) Baars (1967), (3) Wright and Welch (1990), (4) Wright (1996), (5) Bieging et al. (1984), (6) Ishiguro, Kanzawa, and Kasuga (1990), (7) Kasuga, Ishiguro, and Kawabe (1986), (8) Sramek (1983), (9) Carilli and Holdaway (1999), (10) Armstrong and Sramek (1982), (11) Sramek (1990), (12) Olmi and Downes (1992), (13) Masson (1994a), (14) NRAO (1998).

[d] The minimum and maximum values represent the median rms phase fluctuations for all seasons for nighttime (~6 h local time) and daytime (~15 h local time), respectively. (See Fig. 13.13 for diurnal and seasonal variations on Mauna Kea.)

be part of, or closely related to, those produced by Kolmogorov turbulence, which affect interferometric observations (see Fig. 13.11). Hence the term "anomalous refraction" is not particularly appropriate. If these irregularities are assumed to be wedge-shaped, their dominant effect would be a tilting of the wavefront (linear change in phase with position) for a time corresponding to their passage through the antenna beam. A differential excess path length of 0.5 mm for a scale size of 300 m would produce a wavefront tilt of 6 arcsec, and a timescale of 30 s for a wind speed of 10 m s^{-1}. The apparent shift in position of the source is independent of wavelength. There is no effect on the signal amplitude because the scattering is weak (phase fluctuations are small on the Fresnel scale, which is typically a few meters) and no effect on the apparent angular size of the source because the dominant wavefront perturbation is a tilt.

Water Vapor Radiometry

The excess propagation path in a particular direction due to water vapor can be estimated from measurements of the brightness temperature in the same direction at frequencies near water vapor resonances, or in the windows between them. This method was first investigated by Westwater (1967) and Schaper, Staelin, and Waters (1970). To appreciate the degree of correlation between wet path length and brightness temperature, we need to examine the dependence of these quantities on pressure, water vapor density, and temperature. We consider here the interpretation of measurements near the 22.2 GHz resonance. The absorption coefficient given by Eqs. (13.42) and (13.43) is complicated, but at line center it can be approximated by

$$\alpha_m \simeq 0.36 \frac{\rho_V}{PT^{1.875}} e^{-644/T}, \tag{13.111}$$

where T is in kelvins, and we have neglected all except the leading terms in Eq. (13.42). We assume that the opacity given by Eq. (13.47) is small, so that the brightness temperature defined by Eq. (13.45) can be written

$$T_B \simeq 17.8 \int_0^\infty \frac{\rho_V}{PT^{0.875}} e^{-644/T}\, dh, \tag{13.112}$$

when we neglect the background temperature T_{B0} and any contributions from clouds. Recall that Eq. (13.16) shows that

$$\mathcal{L}_V \simeq 0.001763 \int_0^\infty \frac{\rho_V}{T}\, dh. \tag{13.113}$$

Thus, if P and T were constant with height and equal to 1013 mb and 280 K, respectively, we could use Eq. (13.19), $\mathcal{L}_V \simeq 6.3w$, to obtain from Eq. (13.112) the relation $T_B \simeq 12.7w$, where w is the column height of water vapor [see Eq. (13.18)]. Hence, to the degree of approximation used above, we obtain

$$T_B\,(22.2\ \text{GHz})(K) \simeq 2.1\mathcal{L}_V\ (\text{cm}). \tag{13.114}$$

Note that this approximation is valid at sea level. Since, because of pressure broadening, the brightness temperature scales inversely with total pressure [see Eq. (13.112)], the coefficient in Eq. (13.114) is increased to 3.9 for a site at 5000 m elevation where the pressure is approximately 540 mb. Measurements of brightness temperature and path length estimated from radiosonde profiles show that Eq. (13.114) is a good approximation [see, e.g., Moran and Rosen (1981)]. Recall that ρ_V is approximately exponentially distributed with a scale height of 2 km. The temperature, on average, decreases by about 2% per kilometer. This change affects the proportionality between T_B and $\mathcal{L}_V$ only through the exponential factor in Eq. (13.112) and the slight difference in the power law for temperature. Thus, temperature has a small effect. The pressure decreases by 10% per

kilometer, so that water vapor at higher altitudes contributes more heavily to T_B than is desirable for estimation by radiometry. The sensitivity of T_B to pressure is decreased by moving off the resonance frequency to a frequency near the half-power point of the transition. The reason for this is that as pressure increases, the line profile broadens while the integrated line profile is constant. Therefore, the absorption at line center decreases and the absorption in the line wings increases. Westwater (1967) showed that at 20.6 GHz the absorption is nearly invariant with pressure. This particular frequency is called the *hinge point*. The opacity at this frequency is less than at the line center, so the nonlinear relationship between T_B and opacity is less important.

The foregoing discussion assumes that measurements of T_B are made in clear sky conditions. The water droplet content in clouds or fog causes substantial absorption, but small change in the index of refraction compared to that of water vapor. Fortunately, the effect of clouds can be eliminated by combining measurements at two frequencies. In nonprecipitating clouds, the sizes of the water droplets are generally less than 100 μm, and at wavelengths greater than a few millimeters, the scattering is small and the attenuation is due primarily to absorption. The absorption coefficient is given by the empirical formula (Staelin 1966)

$$\alpha_{\text{clouds}} \simeq \frac{\rho_L 10^{0.0122(291-T)}}{\lambda^2} (\text{m}^{-1}), \tag{13.115}$$

where ρ_L is the density of liquid water droplets in grams per cubic meter, λ is the wavelength in meters, and T is in kelvins. This formula is valid for λ greater than ~3 mm where the droplet sizes are small compared with $\lambda/(2\pi)$. For shorter wavelengths the absorption is less than predicted by Eq. (13.115) (Freeman 1987, Ray 1972). A very wet cumulus cloud with a water density of 1 g m^{-3} and a size of 1 km will have an absorption coefficient of 7×10^{-5} m^{-1} and will therefore have a brightness temperature of about 20 K at 22 GHz. The index of refraction of liquid water is about 5 at 22 GHz for $T = 280$ K (Goldstein 1951). The actual excess propagation path through the cloud due to liquid water would be about 4 mm, but the predicted excess path from Eq. (13.114) is 10 cm. Thus, the brightness temperature at a single frequency cannot be used reliably to estimate the excess path length when clouds are present. In order to eliminate the brightness temperature contribution of clouds, measurements must be made at two frequencies, ν_1 and ν_2, one near the water line and one well off the water line, respectively. The brightness temperature is

$$T_{Bi} = T_{BVi} + T_{BCi}, \tag{13.116}$$

where T_{BVi} and T_{BCi} are the brightness temperatures due to water vapor and clouds at frequency i. Here we neglect the effects of atmospheric O_2. Since, from Eq. (13.115), $T_{BC} \propto \nu^2$, we can form the observable

$$T_{B1} - T_{B2}\frac{\nu_1^2}{\nu_2^2} = T_{BV1} - T_{BV2}\frac{\nu_1^2}{\nu_2^2}, \tag{13.117}$$

which eliminates the effect of clouds. The correlation between $T_{BV1} - T_{BV2} \times \nu_1^2/\nu_2^2$ and $\mathcal{L}_V$ can be estimated from model calculations based on Eqs. (13.45) and (13.16). The off-resonance frequency ν_2 is generally chosen to be about 31 GHz. The problem of finding the two best frequencies and the appropriate correlation coefficients to use in predicting $\mathcal{L}_V$ has been widely discussed (Westwater 1978, Wu 1979, Westwater and Guiraud 1980). The liquid content of clouds can also be measured by dual-frequency techniques [see, e.g., Snider, Burdick, and Hogg (1980)].

The application of multi-frequency microwave radiometry to the calibration of wet path length has been described by Guiraud, Howard, and Hogg (1979), Elgered, Rönnäng, and Askne (1982), Resch (1984), Elgered et al. (1991), and Tahmoush and Rogers (2000). The results show that $\mathcal{L}_V$ can be estimated to an accuracy better than a few mm. This is useful for calibrating VLBI delay measurements and extending coherence times. Measurements of T_B at the antennas of short-baseline interferometers can be useful in correcting the interferometer phase (see Section 13.2). More accurate predictions of $\mathcal{L}_V$, or interferometer phase, can be obtained by including measurements in other bands. For example, measurements of the wings of the terrestrial oxygen line near 50 GHz can be used to probe the vertical temperature structure of the troposphere [see, e.g., Miner, Thornton, and Welch (1972), Snider (1972)]. The accuracy of these schemes has been analyzed by Solheim et al. (1998).

13.2 ATMOSPHERIC EFFECTS AT MILLIMETER WAVELENGTHS

Site Testing by Opacity Measurement

At millimeter and submillimeter wavelengths, absorption and path length fluctuations in the atmosphere limit performance in synthesis imaging. This section is concerned with monitoring of atmospheric parameters for optimum choice of sites, and with methods of calibrating the atmosphere to reduce phase errors. This subject has received much attention as a result of the development of several major instruments at millimeter and submillimeter wavelengths.

For given atmospheric parameters, the zenith opacity (optical depth) τ_0 can be calculated as a function of frequency using the propagation model of Liebe (1989). Figure 13.12 shows curves of transmission, $\exp(-\tau_0)$, for 4 mm of precipitable water at an elevation of 2124 m and 1 mm at 5000 m, corresponding to the VLA and ALMA sites, respectively. For the purpose of choosing a suitable observatory site, detailed monitoring of the atmosphere covering both diurnal and annual variation is necessary. We assume that the zenith opacity has the form

$$\tau_\nu = A_\nu + B_\nu w, \tag{13.118}$$

where A_ν and B_ν are empirical constants that depend on frequency, site elevation, and meteorological conditions. Selected measurements of these constants are given in Table 13.4.

The opacity can be monitored by measuring the total noise power received in a small antenna as a function of zenith angle (i.e., the tipping scan method

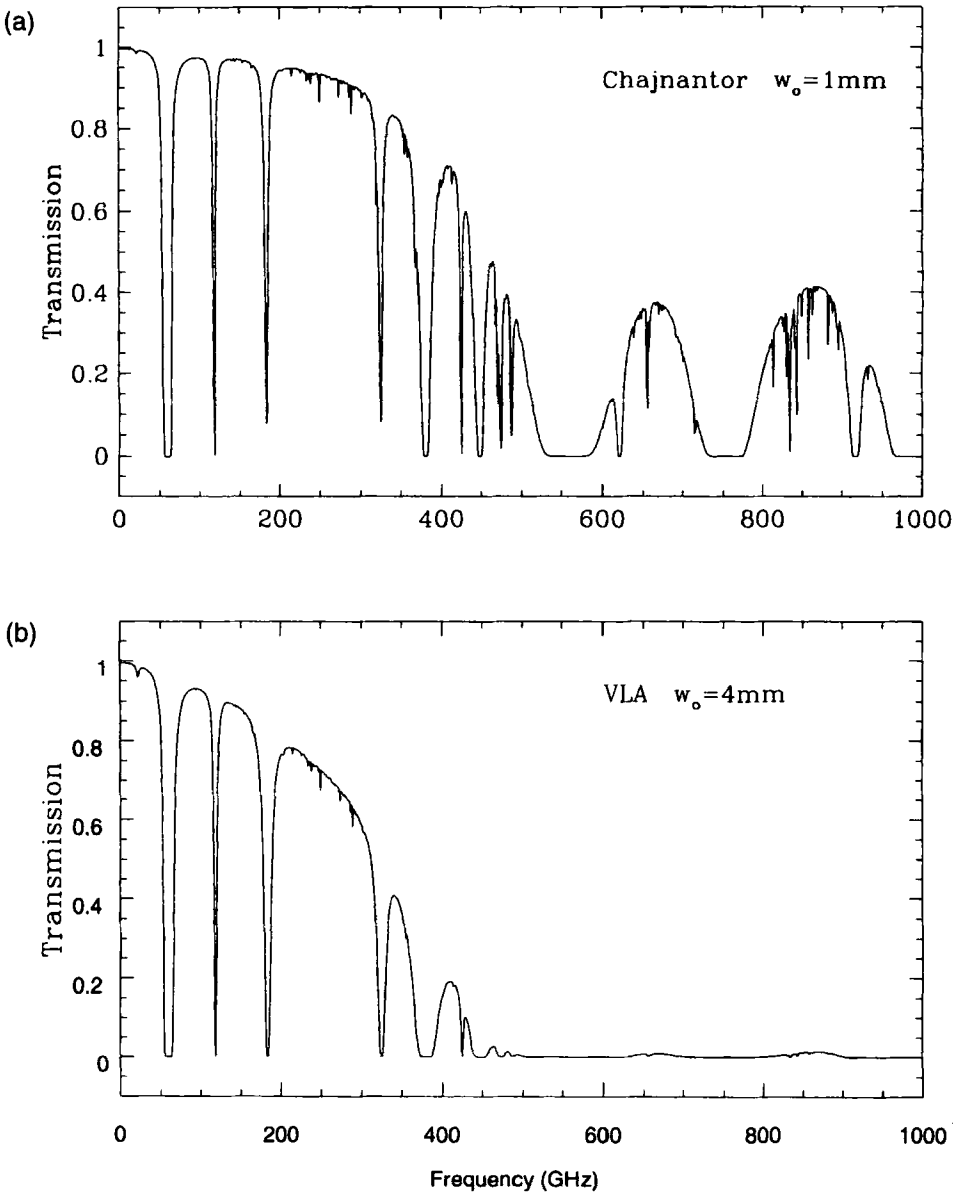


Figure 13.12 (a) The zenith atmospheric transmission at a 5000-m site with 1 mm of precipitable water vapor, calculated from the model of Liebe (1989) over the frequency range of 0–1000 GHz. There are additional windows with transmissions of about 0.10 near 1100, 1300, and 1500 GHz. (b) Transmission for a site at 2124 m with 4 mm of precipitable water. Note that the atmospheric transmission depends on the altitude because of the pressure broadening of the absorption lines. In general, the transmission at any frequency in an atmospheric window will be worse at lower sites with the same amount of precipitable water vapor. From Carilli and Holdaway (1999), ©1999 by the American Geophys. Union.

TABLE 13.4 Zenith Opacity as a Function of Column Height of Water Vapor

ν (GHz)	Location[a]	Altitude (m)	A_ν (nepers)	B_ν (nepers mm^{-1})	Method[b]	Ref.[c]
15	Sea level	0	0.013	0.002	1	1
22.2	Sea level	0	0.026	0.02	1	1
35	Sea level	0	0.039	0.006	1	1
90	Sea level	0	0.039	0.018	1	1
225	South Pole	2835	0.030	0.069	2	2
225	Mauna Kea	4070	0.01	0.04	2	3
225	Chajnantor	5000	0.006	0.033	2	4
225	Chajnantor	5000	0.007	0.041	2	5
493	South Pole	2835	0.33	1.49	2	6

[a] *Locations:* South Pole = Amundsen–Scott Station; Mauna Kea = site of submillimeter telescopes on Mauna Kea; Chajnantor = Llano de Chajnantor, Atacama Desert, Chile.

[b] *Methods:* (1) opacity derived from radiosonde data, water vapor estimated from surface humidity and scale height of 2 km; (2) opacity derived from tipping radiometer, water vapor column height derived from radiosonde data.

[c] *References:* (1) Waters (1976); (2) Chamberlin and Bally (1995); (3) Masson (1994a); (4) Holdaway et al. (1996); (5) Delgado et al. (1998); (6) Chamberlin, Lane, and Stark (1997).

described in Section 13.1 under *Absorption*). A commonly used frequency for opacity monitoring is 225 GHz, which lies within the 200–310 GHz atmospheric window (see Figs. 13.6 and 13.12) in the vicinity of an important CO line.

A typical site-test radiometer designed for opacity measurements uses a small parabolic primary reflector with a beamwidth of ~3° at 225 MHz. A wheel with blades that act as plane reflectors is inserted at the beam waist between the primary and secondary reflectors, and sequentially directs the input of the receiver to the output of the antenna, a reference load at 45°C, and a calibration load at 65°C. The amplified signals go to a power-linear detector, and then to a synchronous detector that produces voltages proportional to the difference between the antenna and the 45°C load, which is the required output, and the difference between the 45 and 65°C loads, which provides a calibration. Measurements of the antenna temperature are made at a range of different zenith angles. When connected to the antenna, the measured noise temperature of such a system, T_{meas}, consists of three components:

$$T_{\text{meas}} = T_{\text{const}} + T_{\text{at}}(1 - e^{-\tau_0 \sec z}) + T_{\text{cb}} e^{-\tau_0 \sec z}. \tag{13.119}$$

Here T_{const} represents the sum of noise components that remain constant as the antenna elevation is varied, that is, the receiver noise, thermal noise resulting from losses between the antenna and the receiver input, any offset in the radiometer detector, and so on. The second term in Eq. (13.119) represents the component of noise from the atmosphere: T_{at} is the temperature of the atmosphere, and z is the zenith angle. $T_{\text{cb}} \simeq 2.7$ K represents the cosmic background radiation. It will be assumed that T_{at} and T_{cb} represent brightness temperatures that are re-

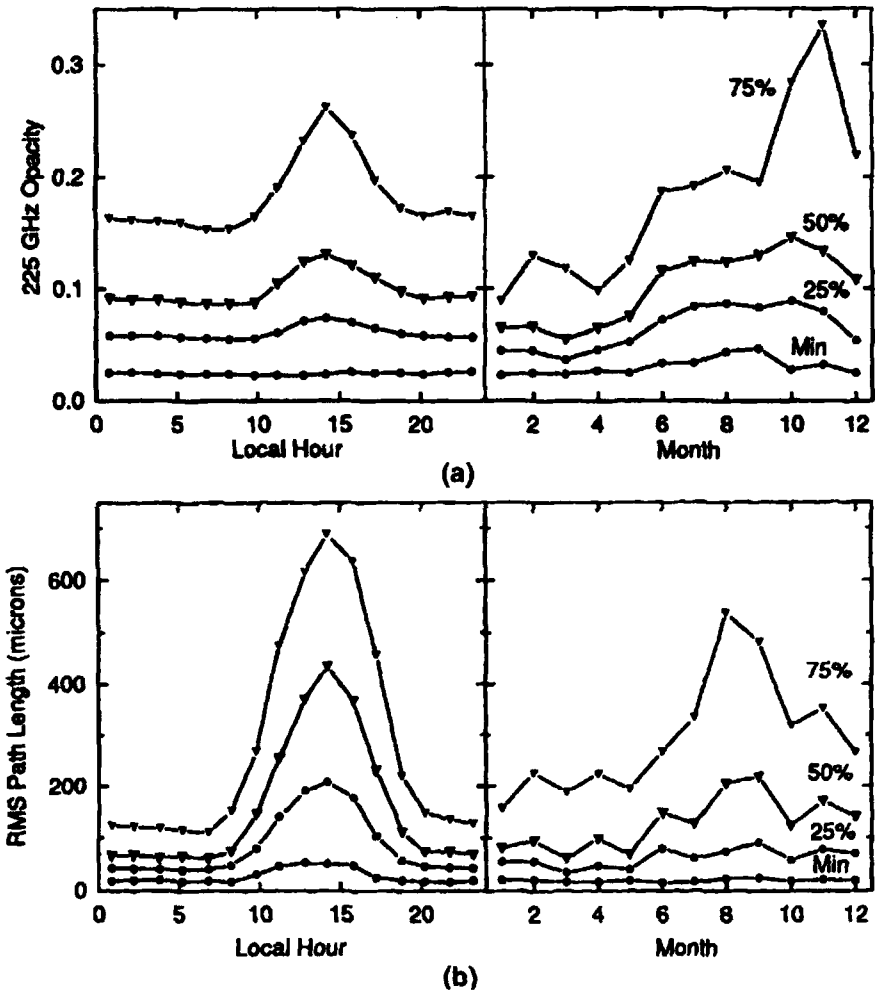


Figure 13.13 (a) Diurnal and seasonal zenith opacity at 225 GHz measured at the CSO site on Mauna Kea (4070 m elevation) for a three-year period (August 1989–July 1992) computed from 14,900 measurements. The minimum value and the 25th, 50th, and 75th percentiles are shown. The increase in opacity during the day is caused by an inversion layer that rises above the mountain in the afternoons. (b) Diurnal and seasonal variation of the rms path length on Mauna Kea on a 100-m baseline, determined from observations of a geostationary satellite at 11 GHz. From Masson (1994a), courtesy of the Astron. Soc. Pacific Conf. Ser.

lated to the physical temperatures by the Planck or Callen and Welton formulas (see Section 7.1 under *Noise Temperature Measurement*). If T_{at} is known, it is straightforward to determine τ_0 from T_{meas} as a function of z. The temperature of the atmosphere is assumed to fall off from the ambient temperature at the earth's surface T_{amb}, with a lapse rate l considered to be constant with height. Thus, at height h the temperature is $T_{amb} - lh$. We require the mean temperature weighted in proportion to the density of water vapor, which is exponentially distributed with scale height h_0:

$$T_{at} = T_{amb} - \frac{l \int_0^\infty h e^{h/h_0}\, dh}{\int_0^\infty e^{h/h_0}\, dh} = T_{amb} - l h_0. \tag{13.120}$$

The lapse rate resulting from adiabatic expansion of rising air, 9.8 K km^{-1}, can be used as an approximate value, but as indicated earlier, a typical measured value is ~6.5 K km^{-1}. The scale height of water vapor is approximately 2 km. Thus T_{at} is typically less than T_{amb} by ~13–20 K.

Figure 13.13 displays examples of data taken on Mauna Kea, which show the diurnal and seasonal effects at this site. The cumulative distribution of zenith opacity at 225 GHz as measured at Llano de Chajnantor in Chile, Mauna Kea, and the South Pole are shown in Fig. 13.14. Measurements of mean opacity provide a basis for calculating the loss in sensitivity due to absorption of the signal and the addition of noise from the atmosphere [see Eq. (13.50)]. The opacity varies both diurnally and annually, so measurements at hourly intervals over a year or more are required for reliable comparison of different sites. Long-term variability due to climatic effects (e.g., El Niño) can be significant. Table 13.4 shows the effect of site altitude on opacity. Comparison of the measurements of A_ν and B_ν show that both parameters decrease with altitude because of the effects of pressure broadening. Comparisons of opacities at various frequencies can be made with broadband Fourier transform spectrometers (Hills et al. 1978; Matsushita et al. 1999; Paine et al. 2000; Pardo, Serabyn, and Cernicharo 2001).

Site Testing by Direct Measurement of Phase Stability

Interferometer observations provide a direct method of determining atmospheric phase fluctuations. Signals from a geostationary satellite are usually used, since strong signals can be obtained using small, non-tracking antennas. This technique was developed by Ishiguro, Kanzawa, and Kasuga (1990); Masson (1994a); and Radford, Reiland, and Shillue (1996). It was used in site testing for the SAO Submillimeter Array on Mauna Kea, Hawaii and Atacama Large Millimeter Array at Llano de Chajnantor. Several suitable geostationary-orbit satellites operate in bands allocated to the fixed and broadcasting services near 11 GHz. Two commercial satellite TV antennas of diameter 1.8 m provide signal-to-noise ratios close to 60 dB. For measurements of atmospheric phase, baselines of 100–300 m have been used. The residual motion of the satellite, as well as any temperature variations, can cause unwanted phase drifts. These are generally slow compared

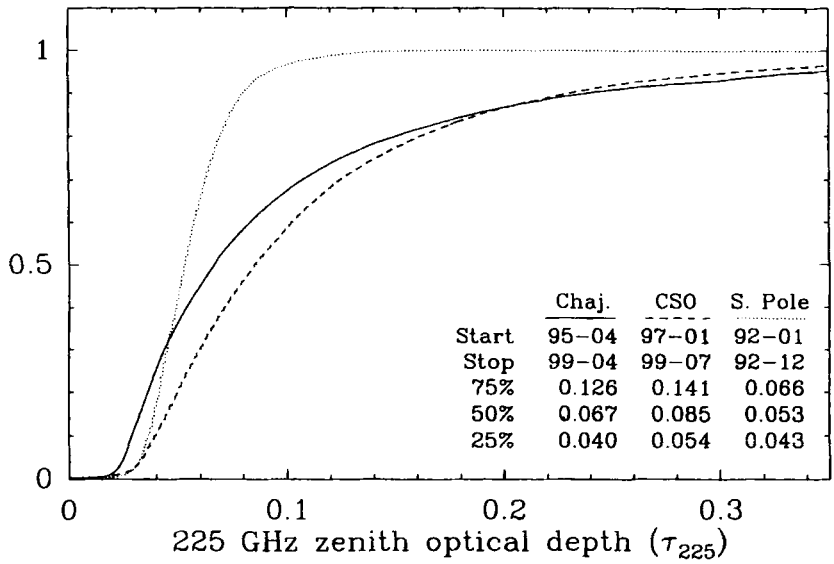


Figure 13.14 Cumulative distributions of the zenith optical depth at 225 GHz on Llano de Chajnantor, Chile (5000-m elevation; solid line), the CSO site on Mauna Kea, Hawaii (4070-m elevation; dashed line), and at the South Pole (2835-m elevation; dotted line) for the periods April 1995–April 1999, Jan. 1997–July 1999, and Jan. 1992–Dec. 1992, respectively. Note that the median opacity at the VLBA site on Mauna Kea (3720-m elevation) for the same time interval at the CSO site was 0.13. The median opacity for the VLA site (2124-m elevation) for the period 1990–1998 was 0.3 (Butler 1998). Conditions at lower elevation sites are correspondingly worse. For example, at a sea-level site in Cambridge, Mass. the 225 GHz opacity, inferred from measurements at 115 GHz, was 0.5 for the 6-month winter observing seasons spanning 1994–1997. See also Radford and Chamberlin (2000).

with the atmospheric effects and can be removed by subtracting a mean and slope from the output data. The variance of the fluctuations resulting from the system noise can also be determined and subtracted from the variance of the measured phases. The test interferometer provides a measure of the structure function of phase $\mathcal{D}_\phi(d)$ for one value of projected baseline d (see Fig. 13.13b).

With the frozen-screen approximation, the power-law exponent can be determined from the power spectrum of the fluctuations. An example is shown in Fig. 13.15. Thus, in extrapolating $\mathcal{D}_\phi(d)$ from a single-spacing measurement, one does not have to depend on the theoretical values of the exponent of d, but can use the measurements of $\mathcal{D}_\phi(\tau)$ to determine the range and variation [see Eq. (13.108) and Table 13.2]. For the example shown in Fig. 13.15, the power-law slope for frequencies above 0.01 Hz is 2.5, slightly below the value of 2.67 predicted for Kolmogorov turbulence. The spectrum flattens at frequencies below 0.01 Hz because of the filtering effect of the interferometer. Fluctuations larger than the baseline, 100 m in this case, cause little phase effect (see Fig. 13.9). Hence the corner frequency f_c is v_s/d. In this case the wind speed along the baseline direction can be inferred to be about 1 m s^{-1}.

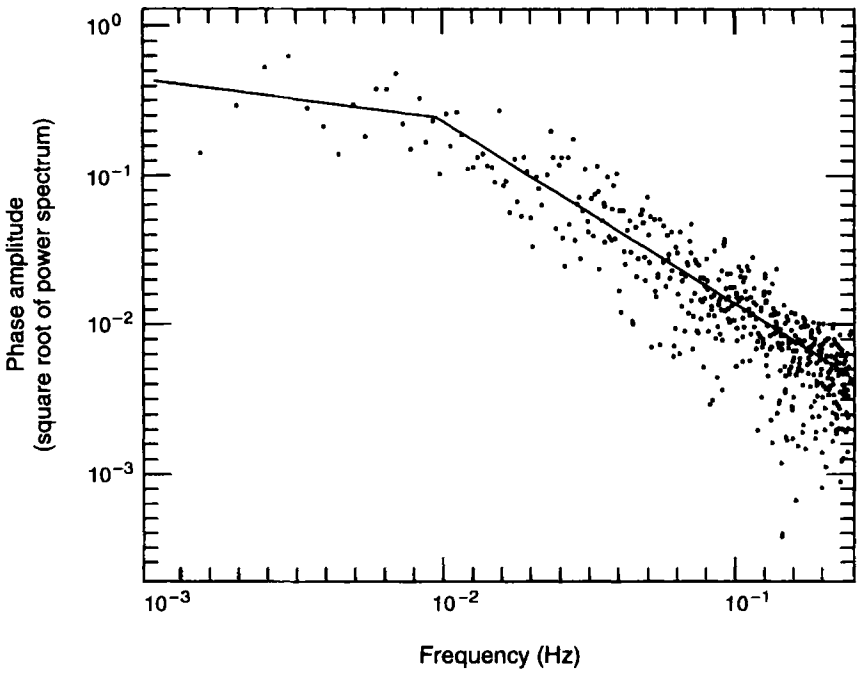


Figure 13.15 The square root of the temporal power spectrum [i.e., Eq. (13.107)] measured on a 100-m baseline on Mauna Kea (CSO site). The tropospheric wind speed along the baseline can be computed from the break in the spectrum. From Masson (1994a), courtesy of the Astron. Soc. Pacific Conf. Ser.

In calculating expected phase fluctuations it should be noted that the variation with zenith angle depends on the baseline length. For baselines short compared with the thickness of the water vapor layer, the (rms) phase variations are proportional to $\sqrt{\sec z}$, and for long baselines they are proportional to $\sec z$. This result can be visualized by noting that on short baselines the effects of large-scale irregularities cancel out between the two antennas, and the variations result from small irregularities, the number of which is roughly proportional to the path length. For long baselines the effects of the largest irregularities, of size comparable to the layer thickness, predominate.

Atmospheric phase errors can be treated like antenna-based phase errors in considering their effect on a map or an image. In Section 11.5 it is shown that the dynamic range of a snapshot image is approximately

$$\frac{\sqrt{n_a(n_a-1)}}{\phi_{\rm rms}}, \tag{13.121}$$

where $\phi_{\rm rms}$ is the rms of the phase error in radians measured with pairs of antennas, and n_a is the number of antennas. For example, if $\phi_{\rm rms}$ is 1 rad and

$n_a = 30$, the dynamic range is ~30. As a rough guide, the range of $\phi_{\rm rms}$ from 0.5 to 1 rad represents array performance from fair to marginal. The improvement in the image with longer integration depends on the spectrum of the phase fluctuations.

Reduction of Atmospheric Phase Errors by Calibration

For phase calibration at centimeter wavelengths, it is common to observe a phase calibrator at intervals of ~20–30 min. At millimeter wavelengths this is generally not satisfactory, because of the much greater phase fluctuations resulting from the atmosphere. Procedures that can be used at millimeter and submillimeter wavelengths to reduce the effect of atmospheric phase fluctuations are described below.

Self-Calibration. The simplest way to remove the effects of atmospheric phase fluctuations is to use self-calibration, as described in Sections 10.3 and 11.4. This method depends on phase closure relationships in groups of three or more antennas. In applying this method it is necessary to integrate the correlator output data for a long enough time that the source can be detected; that is, the measured visibility phase must result mainly from the source, not the instrumental noise. However, the integration time is limited by the fluctuation rate, so self-calibration is not useful for sources that require long integration times to detect.

Frequent Calibration (Fast Switching). Frequent phase calibration using an unresolved source close to the target source (the source under study) can greatly reduce atmospheric phase errors (Holdaway et al. 1995, Lay 1997b). To ensure that the atmospheric phase measured on the calibrator is close to that for the target source, the angular distance between the two sources must be no more than a few degrees. The time difference must be less than ~1 min, so fast position switching between the target source and the calibrator is required. In the layer in which most of the water vapor occurs, the lines of sight from the antennas to the target source and the calibrator pass within a distance d_{tc} of one another. For a nominal screen height of 1 km, $d_{tc} \simeq 17\theta$, where θ is the angular separation in degrees and d_{tc} is in meters. For one antenna, the rms phase difference between the two paths is $\sqrt{\mathcal{D}_\phi(d_{tc})}$ at any instant. If $t_{\rm cyc}$ is the time to complete one observing cycle of the target source and the calibrator, then the mean time difference between the measurements on these two sources is $t_{\rm cyc}/2$. In time $t_{\rm cyc}/2$ the atmosphere will have moved $v_s t_{\rm cyc}/2$. Thus, the phase difference between the measurements on the two paths is effectively $\mathcal{D}_\phi(d_{tc} + v_s t_{\rm cyc}/2)$. This is a worst-case estimate, since we have taken the scalar sum of vector quantities corresponding to d_{tc} and v_s. For the difference in the paths to the two antennas as measured by the interferometer, the rms value will be $\sqrt{2}$ times that for one antenna, so the residual atmospheric phase error in the measured visibility is

$$\phi_{\rm rms} = \sqrt{2\mathcal{D}_\phi(d_{tc} + v_s t_{\rm cyc}/2)}. \tag{13.122}$$

Note that $\phi_{\rm rms}$ is independent of the baseline, so the phase errors should not increase with baseline length. The total time for one cycle of observation of the two sources is the sum of the observing times on the target source and the calibrator, plus twice the antenna slew time between the sources and twice the setup time between ending the slew motion and starting to record data. The observing times required on each of the sources depend on the flux densities and the sensitivity of the instrument. For the calibrator there may be a choice between a weak source nearby and a stronger one that requires less observing time but more antenna slew time. In order to use calibration sources as a general solution to the atmospheric phase problem, suitable calibrators must be available within a few degrees of any point on the sky. Since calibrator flux densities generally decrease with increasing frequency, it may be necessary to observe the calibrator at a lower frequency than is used for the target source. The measured phase for the calibrator must then be multiplied by $\nu_{\rm source}/\nu_{\rm cal}$ (since the troposphere is essentially nondispersive) before subtraction from the target source phases, so the accuracy required for the calibrator phase is increased. Thus, the observing frequency for the calibrator should not be too low; a frequency near 90 GHz may be a practical choice with observations of the target source up to a few hundred gigahertz. The effectiveness of the fast-switching technique is demonstrated by the data in Fig. 13.16. Note that the break in the curve for the 300 s averaging time at antenna spacing 1500 m indicates that the wind speed was about $2 \times 1500/300 = 10$ m s^{-1} (Carilli and Holdaway 1999).

Paired or Clustered Antennas. Location of antennas in closely spaced pairs is an alternative to fast movement between the target source and the calibrator. One antenna of each pair continuously observes the target source and the other observes the calibrator. With this scheme $t_{\rm cyc}$ is zero in Eq. (13.122), but the spacing of the paired antennas, d_p, should be included. The rms residual atmospheric error in the visibility phase becomes

$$\phi_{\rm rms} = \sqrt{2\mathcal{D}_\phi(d_{tc} + d_p)}. \tag{13.123}$$

As in Eq. (13.122), $\phi_{\rm rms}$ is a worst-case estimate, since we have taken a scalar sum of vector quantities corresponding to d_{tc} and d_p. For a 2° position difference between the target source and the calibrator, and an effective height of 1 km for the water vapor, $d_{tc} = 35$ m. For antennas of diameter $\sim$10 m, which is typical for antennas operating up to 300 GHz, d_p should be about 15 m to avoid serious shadowing, and this is smaller than $v_s t_{\rm cyc}/2$ for the fast-switching scheme, since v_s is typically 6–12 m s^{-1} and $t_{\rm cyc}$ is 10 s or more. Thus, with paired antennas the residual phase errors are somewhat less than with fast switching. Also, observing time is not wasted during antenna slewing and setup. However, with fast switching about half of the *time* is devoted to the target source, whereas with paired antennas half of the *antennas* are devoted to the target source, so in the latter case the sensitivity is less by a factor $\sim\sqrt{2}$. If the antennas are grouped in clusters of three or four instead of in pairs, with one antenna in each cluster observing the calibrator, the loss in sensitivity is decreased.

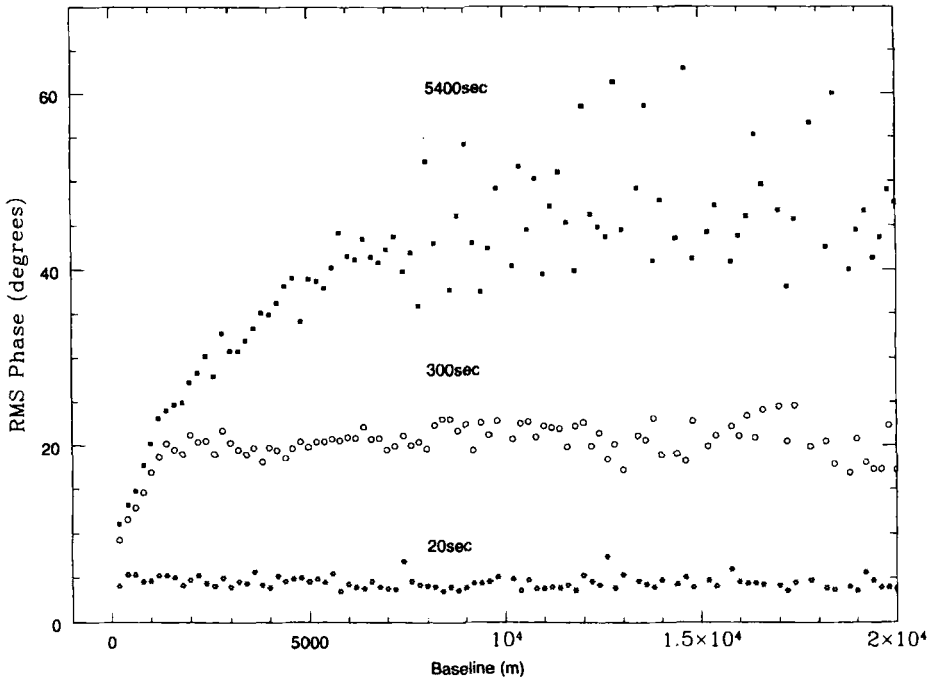


Figure 13.16 The square root of the phase structure, that is, the rms phase deviation versus baseline length, for data taken at the VLA at 22 GHz for various averaging times. These data show the effectiveness of fast phase switching. In these measurements the target source and calibrator source were the same, 0748 + 240. The solid squares (integration time 540 s) show the rms phase fluctuations with no switching (same data as in Fig. 13.11). The circles and the stars show the rms phase deviation for cycle times 300 s and 20 s, respectively. From Carilli and Holdaway (1999), ©1999 by the American Geophys. Union.

Direct Measurement of Water Vapor. A practical method of calibrating the phase fluctuations is to measure the integrated water vapor in the direction of each antenna beam. This usually requires an auxiliary water radiometer at each antenna to measure the sky brightness temperature, as described in Section 13.1 under *Water Vapor Radiometry*. Various techniques are discussed by Welch (1999). For correction of delay in VLBI systems it is usually sufficient to use an auxiliary antenna for the water vapor radiometer. For correction of phase in millimeter and submillimeter interferometers it is important to match the beam of the water radiometer system with that of the interferometer elements. Since the troposphere is in the near field of the beams, the two beams can be arranged to pass though nearly the same volume of the troposphere. Water vapor is the main cause of opacity at radio frequencies (except for the oxygen bands at 50–70 and 118 GHz), even at frequencies well away from the centers of water vapor lines, as can be seen in Fig. 13.6. Away from the centers of spectral lines, the opacity is due to the far line wings of infrared transitions. There is also an important *continuum* compo-

nent of the absorption caused by water vapor, which varies as ν^2 (Rosenkranz 1998). This component includes various quantum mechanical effects involving water molecules such as dimers (Chylek and Geldart 1997). It is usually necessary to model this component with an empirical coefficient. In addition, as described in Section 13.1 under *Water Vapor Radiometry*, the water droplets in the form of clouds and fog, as well as ice crystals, contribute absorption that varies as ν^2. Hence, there are two distinct methods of calibration: those based on measurement of sky brightness in the bands between the lines and those based on measurements near a spectral line. The sensitivities of the brightness temperature to the propagation delay are listed in Table 13.5 for selected frequencies and specific opacities.

The method of measuring the continuum sky brightness at, say, 90 or 230 GHz has several advantages. The same radiometers used for the astronomical measurements can be used for the sky brightness measurements. At 230 GHz, if phase calibration to an accuracy of a twentieth of a wavelength is required, then, from the sensitivity listed in Table 13.5, the brightness temperature accuracy required is 0.1 K. For a system temperature of 200 K, this accuracy requires a gain stability of 5×10^{-4}. Such stability usually requires special attention to the temperature stabilization of the receiver cryogenics. In addition, the gain scales must be accurately calibrated. Changes in ground pickup can be misinterpreted as sky brightness temperatures changes. The presence of clouds defeats this method, because of the contribution of liquid water to the opacity.

The observation of a spectral line provides a calibration technique that is not sensitive to gain variations and ground pickup. As described in Section 13.1 under *Water Vapor Radiometry*, multiple frequencies can be monitored to correct for clouds and the variable distribution of water vapor with height. For millimeter observations at moderately dry sites, the 22-GHz line may be the best choice. An example of phase correction based on this line is shown in Fig. 13.17. For submillimeter wavelength observations at very dry sites, the 183-GHz line may

TABLE 13.5 Change in Brightness Temperature[a] Caused by a Change in Path Length of 1 mm (Δw = 0.16 mm) for a Site at 5000-m Elevation and w = 1 mm

ν (GHz)	Origin of Opacity	ΔT_B (K)
22.2	Line center (6_{16}–5_{23})	0.5
90	Continuum	0.3
183.3	Line center (3_{13}–2_{20})	10.0[b]
185.5	Line wing	16.0
230	Continuum	2.0

[a]Calculated from data in Carilli and Holdaway (1999).
[b]Line is saturated for w = 1 mm.

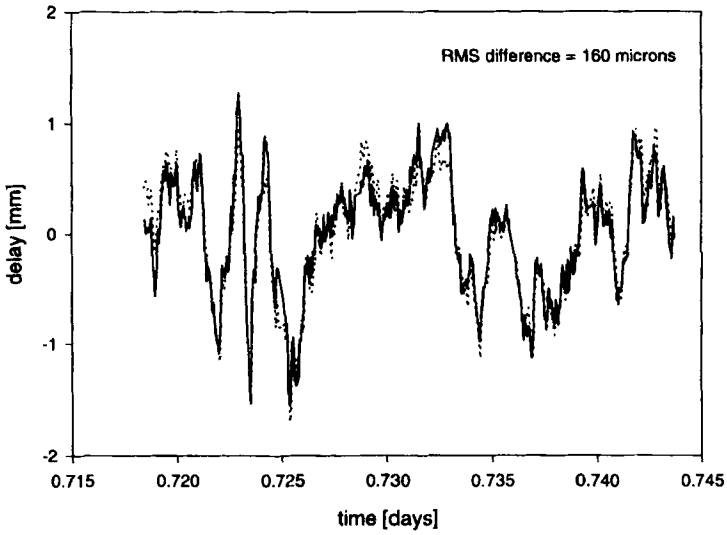


Figure 13.17 The interferometric phase (in units of delay) measured at 3-mm wavelength on one baseline of the interferometer at Owens Valley Radio Observatory (solid line), and the delay predicted by 22 GHz water vapor radiometer measurements (dotted line), versus time. The rms deviation of the difference is 160 μm. The source is 3C273. From Welch (1999); see also Woody, Carpenter, and Scoville (2000).

give better results (Lay 1998, Wiedner and Hills 2000). The 183-GHz line is intrinsically about 40 times more sensitive than the 22-GHz line. However, the 183-GHz line is much more easily saturated (i.e., its opacity exceeds unity) than is the 22-GHz line, which greatly reduces its usefulness. To avoid this problem, the 183-GHz line can be observed in its wings where the opacity is less than unity. Also, the absorption term that varies as ν^2 will be 70 times stronger at 183 GHz than at 22 GHz, which may prove to be a disadvantage for phase correction based on observations at the higher frequency because of the contributions of non-water-vapor components.

13.3 IONOSPHERE

The ionosphere has been studied extensively since the pioneering experiments of Appleton and Barnett (1925) and Breit and Tuve (1926). The literature on the subject is vast. Magneto-ionic propagation theory relevant to the ionosphere is treated in depth by Ratcliffe (1962) and Budden (1961); the morphology of the ionosphere is described by Rawer (1956); and an excellent general treatment of ionospheric propagation is given by Davies (1965). Reviews of particular relevance to radio astronomy can be found in Evans and Hagfors (1968) and Hagfors (1976). Beynon (1975) gives interesting historical anecdotes on the early development of ionospheric research. In this section, we treat only those aspects of the ionosphere that have a deleterious effect on interferometric observations. Table 13.6

TABLE 13.6 Maximum Likely Values of Ionospheric Effects at 100 MHz for a Zenith Angle of 60°[a]

Effect	Maximum[b] (Daytime)	Minimum[c] (Night)	Frequency Dependence
Faraday rotation	15 rotations	1.5 rotations	ν^{-2}
Group delay	12 μs	1.2 μs	ν^{-2}
Excess (phase) path length	3500 m	350 m	ν^{-2}
Phase change	7500 rad	750 rad	ν^{-1}
Phase stability (peak to peak)	±150 rad	±15 rad	ν^{-1}
Frequency stability (rms)	±0.04 Hz	±0.004 Hz	ν^{-1}
Absorption (in D and F regions)	0.1 dB[d]	0.01 dB	ν^{-2}
Refraction (ambient)	0.05°	0.005°	ν^{-2}
Isoplanatic patch	—	~5°	ν

Adapted from Evans and Hagfors (1968).
[a] For values of parameters at the zenith, divide numbers (except refraction) by $\sec z_i$, which is approximately 1.7 [see Eq. (13.140)]. For typical (rather than maximum) parameters, divide numbers by 2.
[b] Total electron content $= 5 \times 10^{17}$ m^{-2}.
[c] Total electron content $= 5 \times 10^{16}$ m^{-2}.
[d] 1 dB = 0.230 nepers.

gives the magnitude of various propagation effects for the daytime and nighttime ionosphere. Most of these effects scale as ν^{-2}, and they can be minimized by observing at higher frequencies. The magnitude of the ionospheric excess path typically equals that of the neutral atmosphere at approximately 2 GHz, but the frequency of this equality can vary from about 1 to 5 GHz. Thus, at 20 GHz the ionospheric excess path length is typically only 1% of the tropospheric excess path length.

Basic Physics

The ionization of the upper atmosphere is caused by the ultraviolet radiation from the sun. Typical daytime and nighttime vertical profiles of the electron density are shown in Fig. 13.18. The electron distribution and the total electron content vary also with geomagnetic latitude, time of year, and sunspot cycle. There are also substantial winds, traveling disturbances, and irregularities in the ionosphere. The ionosphere is permeated by the quasi-dipole magnetic field of the earth. Propagation is governed by the theory of waves in a magnetized plasma with collisions.

We derive some of the fundamental properties of the ionosphere related to the propagation of electromagnetic waves by considering elementary cases. First, consider a plane monochromatic wave of linear polarization that propagates through a uniform plasma of electron density n_e, where the magnetic field and collisions between particles can be neglected. The electrons oscillate with the

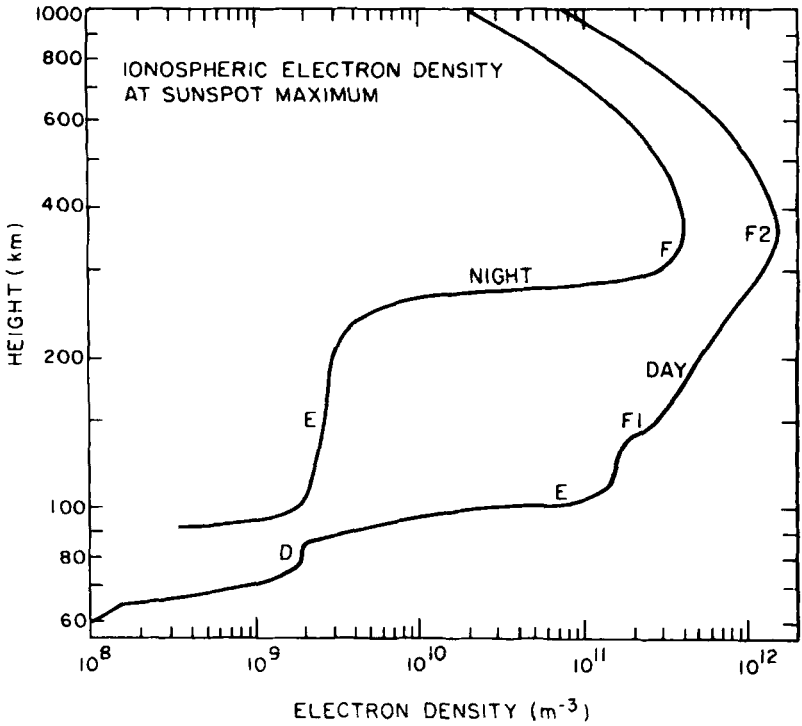


Figure 13.18 Idealized electron density distribution in the earth's ionosphere. The curves indicate the densities to be expected at sunspot maximum in temperate latitudes. Peak sunspot activity occurs at 11-year intervals, most recently in 1989 and 2000. From Evans and Hagfors (1968).

electric field, but the protons, because of their greater mass, remain relatively unperturbed. The index of refraction can be found by calculating either the induced current or the dipole moment. Either method yields the same result. We use the latter method, as we did when considering the index of refraction of water vapor using the bound oscillator model in Section 13.1 under *Origin of Refraction*. The equation of motion of a free electron in the plasma is

$$m\ddot{\mathbf{x}} = -e\mathbf{E}_0 e^{-j2\pi\nu t}, \tag{13.124}$$

where m, e, and $\mathbf{x}$ are the mass, charge *magnitude*, and displacement of the electron, and $\mathbf{E}_0$ and ν are the amplitude and frequency of the electric field $\mathbf{E}$ of the incident wave. The magnetic field of the plane wave has negligible influence on the electrons as long as the electron velocity is much less than c, and the electric field has negligible influence on the motion of the protons. The steady-state solution to Eq. (13.124) is

$$\mathbf{x} = \frac{e}{(2\pi\nu)^2 m}\mathbf{E}_0 e^{-j2\pi\nu t}. \tag{13.125}$$

Note that the induced current density is $\mathbf{i} = n_e e\dot{\mathbf{x}}$, where $\dot{\mathbf{x}}$, the velocity of the particle, is 90° out of phase with the driving electric field. Thus, the work done by the wave on the particles, which is $\langle \mathbf{i}\cdot\mathbf{E}\rangle$, is zero, and the wave propagates without loss, as expected, since Eq. (13.124) has no dissipative terms. The dipole moment per unit volume $\mathbf{P}$ is equal to $n_e e\mathbf{x}_0$, where $\mathbf{x}_0$ is the amplitude of oscillation. The dielectric constant ε is $1 + (\mathbf{P}/\mathbf{E}_0)/\epsilon_0$, where ϵ_0 is the permittivity of free space, so that

$$\varepsilon = 1 - \frac{n_e e^2}{4\pi^2\nu^2\epsilon_0 m}. \tag{13.126}$$

The dielectric constant is real and less than unity because the induced dipole is 180° out of phase with the driving field. The index of refraction n is equal to the square root of ε, and in this case is real, so

$$n = \sqrt{1 - \frac{\nu_p^2}{\nu^2}}, \tag{13.127}$$

where

$$\nu_p = \frac{e}{2\pi}\sqrt{\frac{n_e}{\epsilon_0 m}} \simeq 9\sqrt{n_e}\ (\text{Hz}), \tag{13.128}$$

and n_e is in meters^{-3}. ν_p is known as the *plasma frequency*, which is also the natural frequency of mechanical oscillations in the plasma [see, e.g., Holt and Haskell (1965)]. The plasma frequency of the ionosphere (see Fig. 13.18) is usually less than 12 MHz. Waves normally incident on a plasma with frequencies below ν_p are perfectly reflected. The phase velocity of a wave with $\nu > \nu_p$ in the plasma is c/n, which is greater than c, and the group velocity of a wave packet is cn, which is less than c.

Now consider a plasma with a static magnetic field $\mathbf{B}$ in the direction of propagation of the plane wave. The equation of motion of an electron is

$$m\dot{\mathbf{v}} = -e\left[\mathbf{E} + \mathbf{v}\times\mathbf{B}\right]. \tag{13.129}$$

Let the incident field be a circularly polarized wave. If $\mathbf{B}$ is zero, the particle will follow the tip of the electric field vector and move in a circular orbit. If $\mathbf{B}$ is nonzero, the sum of the $\mathbf{v}\times\mathbf{B}$ force term, which will be in the radial direction, and the electric force term must be balanced by centripetal acceleration. Thus, there is a basic anisotropy in the plasma depending on whether the wave is right or left circularly polarized, since the sign of the $\mathbf{v}\times\mathbf{B}$ term changes between the two cases. The radius R_e of the circular orbit of the electron is derived from the

balance-of-forces equation $eE_0 \pm evB = mv^2/R_e$, where $v = 2\pi\nu R_e$, B is the magnitude of the magnetic field, and the upper and lower signs refer to left and right circular polarization, respectively. Thus, we obtain

$$R_e = \frac{eE_0}{4\pi^2 m\nu^2 \mp 2\pi\nu eB}. \tag{13.130}$$

Following the same procedure as the one described below Eq. (13.125), we find that the index of refraction is given by the equation

$$n^2 = 1 - \frac{\nu_p^2}{\nu(\nu \mp \nu_B)}, \tag{13.131}$$

where ν_B is the gyrofrequency, or cyclotron frequency, given by

$$\nu_B = \frac{eB}{2\pi m}. \tag{13.132}$$

The gyrofrequency is the frequency at which an electron would spiral around a magnetic field line in the absence of any electromagnetic radiation. In the absence of damping, R_e would go to infinity if the applied electric field frequency were ν_B. The gyrofrequency of the earth's magnetic field in the ionosphere ($\sim 0.5 \times 10^{-4}$ tesla) is about 1.4 MHz.

Equation (13.131) gives the index of refraction for the case of a longitudinal magnetic field, that is, where the field is parallel to the direction of wave propagation. The solution for the transverse case is different. The solution for the quasi-longitudinal case is obtained by replacing B with $B\cos\theta$, where θ is the angle between the propagation vector and the direction of the magnetic field. The quasi-longitudinal solution is applicable when the angle θ is less than that specified by the inequality (Ratcliffe 1962)

$$\tfrac{1}{2}\sin\theta\tan\theta < \frac{\nu^2 - \nu_p^2}{\nu\nu_B}. \tag{13.133}$$

When $\nu > 100$ MHz, $\nu_p \simeq 10$ MHz, and $\nu_B \simeq 1.4$ MHz, the quasi-longitudinal solution is valid for $|\theta| < 89°$, or virtually all cases of interest. Therefore, to a high accuracy, when $\nu \gg (\nu_p$ and $\nu_B)$ we can expand Eq. (13.131) to obtain

$$n \simeq 1 - \tfrac{1}{2}\frac{\nu_p^2}{\nu^2} \mp \tfrac{1}{2}\frac{\nu_p^2\nu_B}{\nu^3}\cos\theta, \tag{13.134}$$

where we neglect terms in ν^4 and higher order. For propagation in the direction of **B**, the index of refraction is lower for a left circularly polarized wave than for a right circularly polarized wave.

The difference in the index of refraction for right and left circularly polarized waves leads to the important phenomenon of Faraday rotation, whereby a linearly polarized wave has its plane of polarization rotated as it propagates through the

plasma. A linearly polarized wave with position angle ψ can be decomposed into right and left circularly polarized waves of equal amplitude and phase difference 2ψ. The phase of the two circular waves as they propagate in the y direction through a plasma is $2\pi \nu n_r y/c$ and $2\pi \nu n_\ell y/c$, where n_r and n_ℓ are the indices of refraction for the right circular and left circular modes, respectively. The phase difference between the waves is $2\pi \nu(n_r - n_\ell)y/c$. From Eq. (13.134), $n_r - n_\ell = \nu_p^2 \nu_B \nu^{-3} \cos\theta$, so it is clear that the plane of polarization is rotated by the angle

$$\Delta\psi = \frac{\pi}{c\nu^2} \int \nu_p^2 \nu_B \cos\theta \, dy, \tag{13.135}$$

where ν_p, ν_B, and θ may be functions of y.

For constant magnetic field and electron density, Eq. (13.135) can be written

$$\Delta\psi = 2.6 \times 10^{-13} n_e B \lambda^2 L \cos\theta, \tag{13.136}$$

where $\Delta\psi$ is in radians, n_e is in meters^{-3}, B is in tesla, and is positive when the field is pointed toward the observer, λ is the wavelength in meters, and L is the path length in meters. A magnetic field pointed toward the observer causes the position angle to increase (i.e., a counterclockwise rotation of the plane of polarization of incident radiation as viewed from the surface of the earth).

Refraction and Propagation Delay

Refraction in the ionosphere decreases the zenith angle of signals arriving from outside the earth's atmosphere. This bending is caused by the curvature of the ionosphere. If the ionosphere were a plane-parallel layered structure, then from Eqs. (13.26) the bending angle would be zero. In a well-known approximation (Bailey 1948), the ionosphere is assumed to consist of a layer of thickness Δh, within which there is a parabolic distribution of electron density having a maximum at height h_i. The bending angle in this case is

$$\Delta z = \frac{2\,\Delta h \sin z}{3r_0} \left(\frac{\nu_p}{\nu}\right)^2 \left(1 + \frac{h_i}{r_0}\right) \left(\cos^2 z + \frac{2h_i}{r_0}\right)^{-3/2}, \tag{13.137}$$

where ν_p is the plasma frequency at h_i. If the constant $\Delta h \nu_p^2$ is chosen properly, Δz is accurate to better than 5% for all values of z.

The excess path length in the zenith direction can be calculated using Eqs. (13.5), (13.128), and (13.134) with the assumption that $\nu \gg (\nu_p$ and $\nu_B)$. The result is

$$\mathcal{L}_0 \simeq -\tfrac{1}{2} \int_0^\infty \left[\frac{\nu_p(h)}{\nu}\right]^2 dh \simeq -\frac{40.3}{\nu^2} \int_0^\infty n_e(h)\, dh, \tag{13.138}$$

where ν is in hertz and $n_e(h)$ and $\nu_p(h)$ are the electron distribution (m^{-3}) and plasma frequency as a function of height. The integral of electron density over

height in Eq. (13.138) is called the *total electron content* or *column density*. The excess path corresponds to a phase delay and is negative for the ionosphere. If we approximate the ionosphere by a thin layer at height h_i, then the excess path length will vary as the secant of the zenith angle of the ray as it passes through the layer. Thus

$$\mathcal{L} = \mathcal{L}_0 \sec z_i, \tag{13.139}$$

where z_i (see Fig. 13.4) is given by

$$z_i = \sin^{-1}\left[\left(\frac{r_0}{r_0 + h_i}\right)\sin z\right]. \tag{13.140}$$

When $z = 90°$, $\sec z_i$ is only ~ 3 if $h_i = 400$ km. The secant law provides a reasonable model for estimating the excess ionospheric path length. A more complex model can be found in Spoelstra (1983). Plots of Δz and $\mathcal{L}$ obtained from Eqs. (13.137) and (13.139) as well as from actual ray-tracing calculations are shown in Fig. 13.19.

In some applications, it is necessary to correct the measurements of fringe frequency for the effects of ionospheric delay. The ionospherically induced frequency shift at an antenna is $(\nu/c)d\mathcal{L}/dt$. The time rate of change in excess path length $d\mathcal{L}/dt$ has two components: one caused by the time rate of change of zenith angle dz/dt, and the other caused by the time rate of change of $\mathcal{L}_0$, $d\mathcal{L}_0/dt$. At many times, especially near sunrise and sunset, the latter term may dominate (Mathur, Grossi, and Pearlman 1970).

Calibration of Ionospheric Delay

The excess ionospheric path length must be calibrated as accurately as possible in experiments involving precise determination of source positions or baselines. Three approaches are possible. Models of the ionosphere can be constructed that depend on parameters such as geomagnetic latitude, solar time, season, and solar activity. Two such models are the International Reference Ionosphere (IRI) (Bilitza 1997) and the Parameterized Ionospheric Model (PIM) (Daniell et al. 1995). Alternatively, estimates of the total electron content can be obtained from measurements of the dual-frequency transmissions from the Global Positioning System (GPS) (Ho et al. 1997, Mannucci et al. 1998). GPS is rapidly replacing the more traditional methods such as ionosondes, Faraday rotation of satellite signals, and incoherent backscatter radar (Evans 1969). Finally, the differential path length effects can be virtually eliminated for unresolved sources by making astronomical observations simultaneously at two widely separated frequencies, ν_1 and ν_2. If the interferometer phases are ϕ_1 and ϕ_2 at the two frequencies, then the quantity

$$\phi_c = \phi_2 - \left(\frac{\nu_1}{\nu_2}\right)\phi_1 \tag{13.141}$$

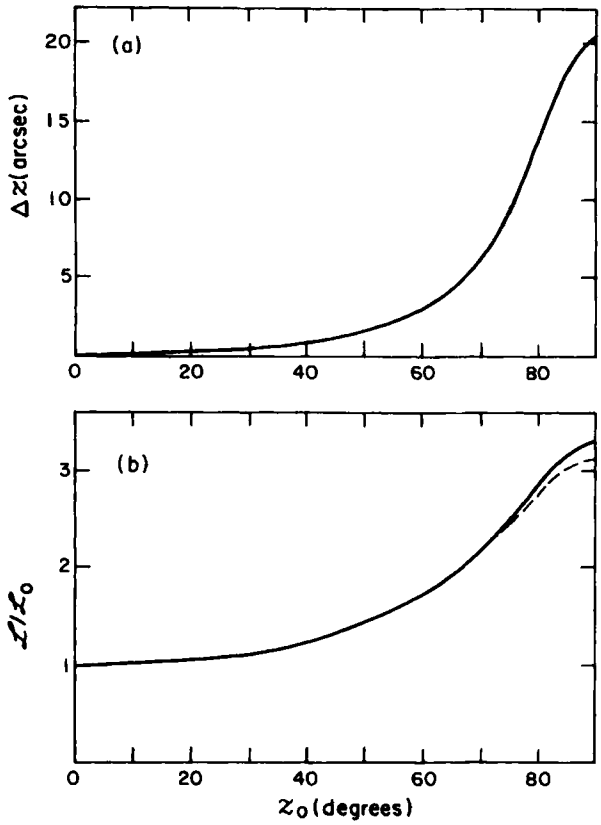


Figure 13.19 (a) Ionospheric bending angle versus zenith angle at 1000 MHz from a ray-tracing calculation for the daytime electron density profile in Fig. 13.12. The bending predicted by Eq. (13.137), with parameters $\nu_p = 12$ MHz, $h_i = 350$ km, $\Delta h = 225$ km, and $r_0 = 6370$ km, differs from the curve shown by no more than 5%. (b) Normalized ionospheric excess path length versus zenith angle for the same electron density profile from a ray-tracing calculation (solid curve) and from Eq. (13.139) (dashed curve). The total electron content is 6.03×10^{17} m^{-2}, and the excess path length at the zenith is 24.3 m. The bending and excess path length scale as ν^{-2}.

will preserve source position information and be substantially free of ionospheric delay effects. A small residual error remains because of higher-order frequency terms in the index of refraction and because the rays at the two frequencies traverse slightly different paths through the ionosphere. Dual-frequency observations are widely used in astrometric radio interferometry where source structure can be neglected [see, e.g., Fomalont and Sramek (1975), Kaplan et al. (1982), Shapiro (1976)]. Note that the difference in total electron content along the ray paths to the interferometer elements can be estimated from measurement of $\phi_2 - (\nu_2/\nu_1)\phi_1$. Similar dual-frequency systems can be employed for the trans-

fer of a local oscillator reference to a space-based VLBI station, see, for example, Moran (1989) and Section 9.10.

Absorption

Absorption in the ionosphere is caused by collisions of electrons with ions and neutral particles. At frequencies much greater than ν_p, the power absorption coefficient is

$$\alpha = 2.68 \times 10^{-7} \frac{n_e \nu_c}{\nu^2} (\mathrm{m}^{-1}), \tag{13.142}$$

where ν_c is the collision frequency and n_e is in meters^{-3}. The collision frequency in hertz is approximately

$$\nu_c \simeq 6.1 \times 10^{-9} \left(\frac{T}{300}\right)^{-3/2} n_i + 1.8 \times 10^{-14} \left(\frac{T}{300}\right)^{1/2} n_n, \tag{13.143}$$

where n_i is the ion density and n_n is the neutral particle density, both in meters^{-3} (Evans and Hagfors 1968). Numerical values of absorption are listed in Table 13.6.

Small- and Large-Scale Irregularities

The small-scale irregularities in the electron density distribution introduce random changes in the wavefront of a passing electromagnetic wave. As a consequence, fluctuations in fringe amplitude and phase can be readily observed with an interferometer at frequencies below a few hundred megahertz. In the early days of radio astronomy, signals from Cygnus A and other compact sources were observed to fluctuate on timescales of 0.1–1 min. At first these fluctuations were thought to be intrinsic to the sources (Hey, Parsons, and Phillips 1946), but later observations with spaced receivers showed that the fluctuations were uncorrelated for receiver separations of more than a few kilometers (Smith, Little, and Lovell 1950). This result led to the conclusion that irregularities in the ionosphere were perturbing the cosmic signals. The predominant scale sizes in the ionization irregularities were found to be a few kilometers or less. The timescale of the fluctuations indicates that ionospheric wind speeds are in the range of 50–300 m s^{-1}. The effects of these fluctuations have been studied extensively at frequencies between about 20 and 200 MHz, and have been observed at frequencies as high as 7 GHz (Aarons et al. 1983). An early example of the fluctuations seen in interferometer measurements is given in Fig. 13.20. Hewish (1952), Booker (1958), and Lawrence, Little, and Chivers (1964) reviewed the early results and techniques. A comprehensive review of theory and observations of ionospheric fluctuations can be found in Crane (1977), Fejer and Kelley (1980), and Yeh and Liu (1982), and summaries of global morphology can be found in Aarons (1982) and Aarons et al. (1999). Measurements with the GPS can be very useful in monitoring ionospheric fluctuations [e.g., Ho et al. (1996), Pi et al. (1997)]. The effects of iono-

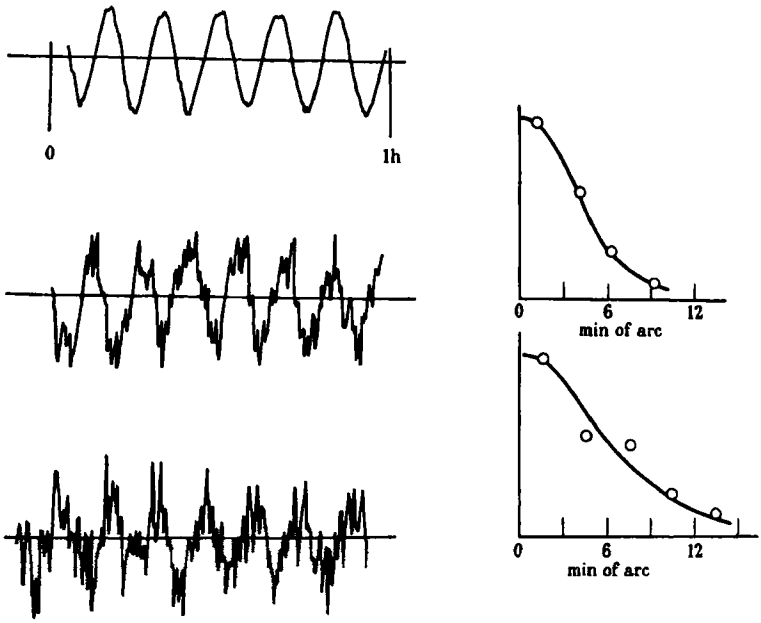


Figure 13.20 (Left) Typical records of the correlator output on three occasions from a phase-switching interferometer at Cambridge, England, having a 1-km baseline and operating at a wavelength of 8 m. The irregular responses are caused by disturbances in the ionosphere. (Right) Probability distributions of the angle of arrival deduced from the zero crossings of the correlator response. From Hewish (1952).

spheric scintillation on a synthesis telescope have been described by Spoelstra and Kelder (1984). In Section 13.4 we discuss a theory of scintillation, which can be applied to the ionosphere as well as to the interplanetary and interstellar media.

Large-scale variations in the electron density integrated along the line of sight are caused by traveling ionospheric disturbances (TIDs). TIDs, which are manifestations of acoustic-gravity waves in the upper atmosphere, are quasi-periodic, large-scale perturbations in electron density. The atmosphere has a natural buoyancy, so that a parcel of gas displaced vertically and released will oscillate at a frequency known as the Brunt–Väisälä, or buoyancy, frequency. This frequency is about 0.5–2 mHz (periods of 10–20 min) at ionospheric heights. For waves with frequencies above the buoyancy frequency, the restoring force is pressure (acoustic wave), and for waves with frequencies below the buoyancy frequency, the restoring force is gravity (gravity wave). Hunsucker (1982) and Hocke and Schlegel (1996) have reviewed the literature on acoustic-gravity waves. There are many potential sources of TIDs, including auroral heating, severe weather fronts, earthquakes, and volcanic eruptions. Medium-scale TIDs have scale lengths of 100–200 km, timescales of 10–20 min, and cause a variation in total electron content of 0.5–5%. Such TIDs are present for a substantial fraction of the time.

Large-scale TIDs, which are relatively uncommon, have scale lengths of 1000 km, timescales of hours, and can cause variations in total electron content of up to 8%. One such disturbance, excited by a volcano, was observed by VLBI (Roberts et al. 1982). The effects of TIDs on radio interferometry, which are primarily slow phase variations, are described by Hinder and Ryle (1971). The effects on satellite tracking are described by Evans, Holt, and Wand (1983).

13.4 SCATTERING CAUSED BY PLASMA IRREGULARITIES

Understanding the propagation of radiation in a random medium is an important problem in many fields. The signals from cosmic radio sources propagate through several random media, including the ionized interstellar gas of our Galaxy, the solar wind, and the ionosphere, as well as the neutral gas of the troposphere. In the observer's plane there are two effects. First, the amplitude varies with the position of the observer, which leads to temporal amplitude variations if there are relative motions among the source, scattering medium, and observer. Second, the image of the source is also distorted. Much of the research in this area has been motivated by the attempt to understand the observational characteristics of pulsars [see, e.g., Gupta (2000)].

Gaussian Screen Model

We begin the discussion by considering a simple model that serves to illustrate many features of the problem. This model was first developed by Booker, Ratcliffe, and Shinn (1950) to explain ionospheric scintillation and was refined by Ratcliffe (1956). Scheuer (1968) applied it to pulsar observations. The model assumes that the irregular medium is confined to a thin screen and that the irregularities (blobs) have one characteristic scale size a. Diffraction effects are neglected within the irregular medium; only the phase change imposed by the medium is considered. Diffraction is taken into account in the free-space region between the irregular medium and the receivers.

The geometric situation is shown in Fig. 13.21. The thin-screen assumption is not particularly restrictive. However, the assumption that the screen is filled with plasma blobs having one characteristic size is restrictive, and distinguishes this model from the power-law model where a range of scale sizes is present. From Eqs. (13.128) and (13.134), the index of refraction of the plasma can be written

$$n \simeq 1 - \frac{r_e n_e \lambda^2}{2\pi}, \tag{13.144}$$

where r_e is the classical electron radius, equal to $e^2/4\pi\epsilon_0 mc^2$ or 2.82×10^{-15} m, and the term in ν_B is neglected. Thus, the excess phase shift across one blob is

$$\Delta\phi_1 = r_e \lambda a \Delta n_e, \tag{13.145}$$

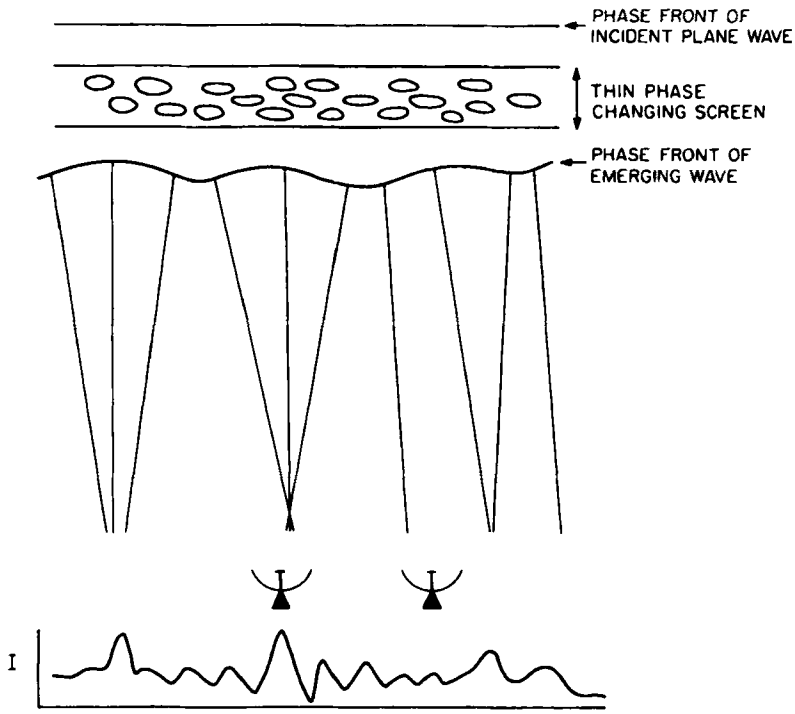


Figure 13.21 Geometry of a thin-screen scintillation model. An initially plane wave is incident on a thin phase-changing screen. The emerging wavefront is irregular. As the wave propagates to the observer, amplitude fluctuations develop, as suggested by the crossing rays. Below the antenna is a plot of intensity versus position along the wavefront. If there is motion between the screen and the observer, the spatial fluctuations will be observed as temporal fluctuations in the power received or the fringe visibility.

where Δn_e is the excess electron density in the blob over the ambient level. If the thickness of the screen is L, then the wave will encounter about L/a blobs, and the rms phase deviation $\Delta\phi$ will be $\Delta\phi_1\sqrt{L/a}$, or

$$\Delta\phi = r_e\lambda\,\Delta n_e\sqrt{La}. \tag{13.146}$$

The wave emerging from the screen is crinkled; that is, the amplitude is unchanged, but the phase is no longer constant and has random fluctuations with rms deviation $\Delta\phi$. The wave can therefore be decomposed into an angular spectrum of waves propagating with a variety of angles. The full width of the angular spectrum, θ_s, can be estimated by imagining that the random medium consists of refracting wedges that tilt the wavefront by the amount $\pm\Delta\phi\lambda/2\pi$ over a distance a. Thus

$$\theta_s = \frac{1}{\pi}r_e\lambda^2\,\Delta n_e\sqrt{\frac{L}{a}}. \tag{13.147}$$

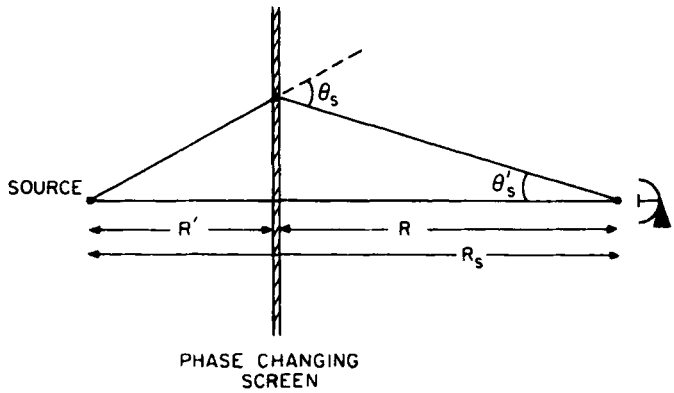


Figure 13.22 Path of a refracted ray in the thin phase screen model. The rms scattering angle θ_s is given by Eq. (13.147).

If the source is not infinitely distant, then the incident wave will not be plane. In that case the observed scattering angle θ'_s depends on the location of the screen with respect to the source and the observer. Since θ_s and θ'_s are small angles, it follows from the geometry in Fig. 13.22 that

$$\theta'_s = \frac{R'}{R + R'}\theta_s, \tag{13.148}$$

where R and R' are defined in Fig. 13.6. Therefore, the effectiveness of the scattering screen is diminished if the screen is moved toward the source. This lever effect is very important in astrophysical situations. It can be used to distinguish galactic and extragalactic sources whose radiation passes through the same scattering screen (Lazio and Cordes 1998).

Amplitude fluctuations build up as the wave propagates away from the screen. If the phase fluctuations are large, $\Delta\phi > 1$, then significant amplitude fluctuations occur when rays cross (see Fig. 13.21). The critical distance beyond which large-amplitude fluctuations are observed is

$$R_f \simeq \frac{a}{\theta'_s}. \tag{13.149}$$

Note that if $\Delta\phi = 2\pi$, then R_f is the distance for which the size of a blob is equal to the size of the first Fresnel zone. The random electric field distribution at the earth, in the plane perpendicular to the propagation direction, is called the diffraction pattern. It has a characteristic correlation length d_c given by

$$d_c \simeq \frac{\lambda}{\theta'_s}. \tag{13.150}$$

If the screen moves with relative velocity v_s in the direction perpendicular to the propagation direction, so that the diffraction pattern sweeps across the observer, then the timescale of variability is

$$\tau_d \simeq \frac{d_c}{v_s}\frac{R'}{R+R'} \simeq \frac{\lambda}{\theta_s v_s}. \tag{13.151}$$

The signal reaching the observer by traveling along the scattered ray path is delayed by an amount

$$\tau_c \simeq \frac{RR'\theta_s^2}{2c(R+R')} \tag{13.152}$$

with respect to the unscattered signal. The phase of the scattered wave is $2\pi\nu\tau_c$ with respect to the direct (unscattered) wave, and interference between these two waves causes scintillation. The bandwidth over which the relative phase changes by 2π is called the correlation bandwidth, $\Delta\nu_c$. The correlation bandwidth is the reciprocal of τ_c, and for the case $R = R'$ is

$$\Delta\nu_c \simeq \frac{8c}{R_s\theta_s^2}, \tag{13.153}$$

where R_s is the distance between the source and the observer. If the observations are made with a receiver of bandwidth greater than $\Delta\nu_c$, the amplitude fluctuations will be greatly reduced. Note from Eqs. (13.153) and (13.147) that $\Delta\nu_c$ varies as λ^{-4}.

Finally, if the source has two equal components separated by distance ℓ, then each component will produce the same diffraction pattern, but these patterns will be displaced at the earth by distance $\ell R/R'$. If this distance is greater than d_c, then the diffraction pattern will be smoothed and the amplitude fluctuations reduced. Thus, if the source size is greater than a critical size θ_c, amplitude fluctuations will be sharply reduced because the diffraction patterns from the component parts overlap and are smoothed out. From Eqs. (13.148) and (13.150), θ_c can be written as

$$\theta_c = \frac{\lambda}{R\theta_s}. \tag{13.154}$$

Hence, only sources of small angular diameter scintillate. In the optical regime, the analogous phenomenon is that stars twinkle, but usually planets do not. An elegant application of Eq. (13.154) was made by Frail et al. (1997) to determine the angular size of the expanding radio source associated with a γ-ray burst. They determined that the amplitude fluctuations in the radio emission, assumed to be caused by interstellar scattering, ceased during the first weeks after the burst, indicating that the source diameter had increased beyond the critical size of 3 μarcsec at that time.

A useful quantity is the ensemble average fringe visibility, $\mathcal{V}_m$, measured by an interferometer in the presence of scintillation. Assume that the phases ϕ_1 and ϕ_2 at two points along the phase screen, separated by distance d, are random

variables with a joint Gaussian distribution with variance $\Delta\phi^2$ and normalized correlation $\rho(d)$. $\rho(d)$ is the correlation function of the phase, or of the variable component of the index of refraction. The joint probability density function of the phase along the wavefront is

$$p(\phi_1, \phi_2) = \frac{1}{2\pi\, \Delta\phi^2 \sqrt{1-\rho(d)^2}} \exp\left[-\frac{\phi_1^2 + \phi_2^2 - 2\rho(d)\phi_1\phi_2}{2\,\Delta\phi^2[1-\rho(d)^2]}\right], \tag{13.155}$$

where $\rho(d) = \langle\phi_1\phi_2\rangle/\Delta\phi^2$. The expectation of $e^{j(\phi_1-\phi_2)}$ is

$$\langle e^{j(\phi_1-\phi_2)}\rangle = \iint e^{j(\phi_1-\phi_2)}\, p(\phi_1, \phi_2)\, d\phi_1\, d\phi_2\,, \tag{13.156}$$

which can be evaluated directly from Eq. (13.155) with the result

$$\langle e^{j(\phi_1-\phi_2)}\rangle = e^{-\Delta\phi^2[1-\rho(d)]}. \tag{13.157}$$

For a point source of flux density S, the ensemble average of the fringe visibility is

$$\langle \mathcal{V}_m\rangle = S\langle e^{j\phi_1} e^{-j\phi_2}\rangle\,, \tag{13.158}$$

or

$$\langle \mathcal{V}_m\rangle = S e^{-\Delta\phi^2[1-\rho(d)]}\,. \tag{13.159}$$

If the source has an intrinsic visibility $\mathcal{V}_0$, the ensemble average is

$$\langle \mathcal{V}_m\rangle = \mathcal{V}_0 e^{-\Delta\phi^2[1-\rho(d)]}\,. \tag{13.160}$$

This result was first derived by Ratcliffe (1956) and Mercier (1962). In much of the early radio astronomical literature, $\rho(d)$ is assumed to be a Gaussian function

$$\rho(d) = e^{-d^2/2a^2}, \tag{13.161}$$

where the characteristic scale length a corresponds to the blob size in the discussion above. This model, called the Gaussian screen model, is probably unrealistically restrictive because there are undoubtedly many scale sizes present. In the case where $\Delta\phi \gg 1$, $\mathcal{V}_m$ decreases rapidly as d increases, and we need consider only the case of $d \ll a$. In this case substitution of Eq. (13.161) into Eq. (13.160) yields

$$\langle \mathcal{V}_m\rangle \simeq \mathcal{V}_0 e^{-\Delta\phi^2 d^2/2a^2}. \tag{13.162}$$

Thus, the intensity distribution of a point source observed through a Gaussian screen is a Gaussian distribution with a diameter (full width at half maximum) of

$$\theta_s \simeq \sqrt{2\ln 2}\,\frac{\Delta\phi\lambda}{\pi a} = \frac{\sqrt{2\ln 2}}{\pi} r_e \lambda^2\,\Delta n_e \sqrt{\frac{L}{a}}. \tag{13.163}$$

This formula for θ_s is essentially equivalent to the one given in Eq. (13.147). In the case where $\Delta\phi \ll 1$, the normalized visibility function drops from unity to $e^{-\Delta\phi^2}$ when $d \gg a$. Therefore, the resulting intensity distribution for a point source is an unresolved core surrounded by a halo. The ratio of the flux density in the halo to the flux density in the core is $e^{\Delta\phi^2} - 1$.

Power-Law Model

The spectrum of fluctuations in the electron density in ionized astrophysical plasmas is normally modeled as a power law,

$$P_{ne} = C_{ne}^2 q^{-\alpha}, \tag{13.164}$$

where q is the three-dimensional spatial frequency, $q^2 = q_x^2 + q_y^2 + q_z^2$, and C_{ne}^2 characterizes the strength of the turbulence. The definition of C_{ne}^2 varies in the literature, depending on whether it is used as a constant in the spectrum or in the structure function. The two-dimensional power spectrum of phase [see Eq. (13.145) for the relation between ϕ and n_e] is

$$P_\phi(q) = 2\pi r_e^2 \lambda^2 L P_{ne}. \tag{13.165}$$

Hence, from Eq. (13.104), the structure function of phase is

$$\mathcal{D}_\phi(d) = 8\pi^2 r_e^2 \lambda^2 L \int_0^\infty [1 - J_0(qd)] P_{ne}(q) q\, dq. \tag{13.166}$$

For a power-law spectrum of the form of Eq. (13.164), the structure function is

$$\mathcal{D}_\phi(d) = 8\pi^2 r_e^2 \lambda^2 C_{ne}^2 L f(\alpha)\, d^{\alpha-2}, \tag{13.167}$$

where $f(\alpha)$ is of order unity. The index α is often taken to be $11/3$, which is its value for Kolmogorov turbulence, for which $f(\alpha) = 1.45$ [see Cordes, Pidwerbetsky, and Lovelace (1986) for other values of $f(\alpha)$]. The ensemble average of the interferometric visibility [see Eq. (13.81)] is

$$\langle \mathcal{V} \rangle = \mathcal{V}_0 e^{-\mathcal{D}_\phi/2}, \tag{13.168}$$

or

$$\langle \mathcal{V} \rangle = \mathcal{V}_0 e^{-4\pi^2 r_e^2 \lambda^2 C_{ne}^2 L f(\alpha)\, d^{\alpha-2}}. \tag{13.169}$$

The observed intensity distribution, the Fourier transform of Eq. (13.169), differs slightly from a Gaussian distribution, as can be seen in Fig. 13.10b. The scattering angle (full width at half maximum) obtained from the width of the intensity distribution is

$$\theta_s \simeq 4.1 \times 10^{-13} (C_{ne}^2 L)^{3/5} \lambda^{11/5} \text{ (arcsec)}, \tag{13.170}$$

where λ is in units of meters and $C_{ne}^2 L$ is in meters$^{-17/3}$. Thus, a difference between the power-law model and the Gaussian screen model is that θ_s, measured by Fourier transformation of visibility data over a range of baselines, is proportional to $\lambda^{2.2}$ in the former model and to λ^2 in the latter. Note that if $\langle \mathcal{V} \rangle$ were measured on a single baseline, that is, with d fixed, and if θ_s were estimated from comparison of the measured visibility with the visibility expected for a Gaussian intensity distribution, then θ_s would appear to vary as λ^2 in both models.

Measurements of visibility must be made over sufficiently long integration times to achieve an ensemble average if Eqs. (13.168), (13.169), and (13.170) are to be valid (Cohen and Cronyn 1974). A detailed discussion of the averaging time necessary to achieve an ensemble average is given by Narayan (1992).

For plasmas we can expect that the power law will hold from an inner scale q_0 to an outer scale q_1; that is, there are no fluctuations on length scales smaller than $1/q_1$ or larger than $1/q_0$. For the case where $qd \ll 1$, that is, where the baseline is shorter than the inner length scale, the Bessel function in Eq. (13.166) becomes $1 - q^2 r^2/4$ and the integration is straightforward, yielding

$$\mathcal{D}_\phi(d) = \frac{2\pi^2 r_e^2 \lambda^2 L C_{ne}^2}{4-\alpha} (q_1^{4-\alpha} - q_0^{4-\alpha})\, d^2. \tag{13.171}$$

This result has two interesting consequences. First, the structure function varies as d^2 regardless of α. Second, for $\alpha < 4$, the structure function is dominated by the effect of the smallest irregularities, whereas for $\alpha > 4$, it is dominated by the effect of the largest-scale irregularities. This result also suggests an important demarcation in phenomena between plasmas with $\alpha < 4$ and those with $\alpha > 4$. The case where $\alpha < 4$ is called Type A (shallow spectrum), and the case where $\alpha > 4$ is called Type B (steep spectrum) (Narayan 1988).

Consider the situation where the spectrum has three regimes:

$$\begin{aligned} P_{ne} &= C_{ne}^2 q_0^{-\alpha}, & q &< q_0 \\ &= C_{ne}^2 q^{-\alpha}, & q_0 &< q < q_1 \\ &= 0, & q &> q_1. \end{aligned} \tag{13.172}$$

Substitution of Eq. (13.172) into Eq. (13.166) gives

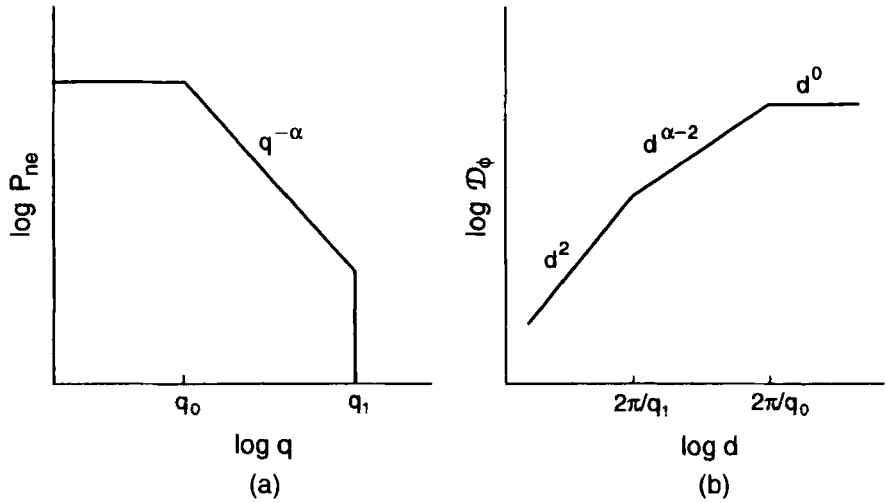


Figure 13.23 (a) A model spectrum of the electron density fluctuations with inner and outer scales of spatial frequency q_0 and q_1. (b) The corresponding structure function of phase: see Eqs. (13.172) and (13.173). From Moran (1989), ©1989 by Kluwer Academic Publishers, reproduced with permission.

$$\begin{aligned}
\mathcal{D}_\phi(d) &= \left(\frac{2\pi}{q_1 d_0}\right)^\alpha d^2, & d < \frac{2\pi}{q_1} \\
&= \left(\frac{d}{d_0}\right)^{\alpha-2}, & \frac{2\pi}{q_1} < d < \frac{2\pi}{q_0} \\
&= \left(\frac{2\pi}{q_0 d_0}\right)^{\alpha-2}, & d > \frac{2\pi}{q_0},
\end{aligned} \tag{13.173}$$

where we have introduced the normalization factor d_0, such that $\mathcal{D}_\phi(d_0) = 1$, as in the discussion of the troposphere in Section 13.2 under *Kolmogorov Turbulence*. We have also assumed that $2\pi/q_1 < d_0 < 2\pi/q_0$. This spectrum and structure function for the model are shown in Fig. 13.23.

13.5 INTERPLANETARY MEDIUM

Refraction

Radio waves passing near the sun are bent by the ionization of the solar corona and the solar wind. The general characteristics of the solar corona and the solar wind can be found in Winterhalter et al. (1996). Calculation of the refraction in the extended solar atmosphere is important for the understanding of solar radio emission at low frequencies, where the bending angles are large (Kundu 1965), and for tests of the general relativistic bending of electromagnetic radiation passing near the sun (Shapiro 1967, Fomalont and Sramek 1977). The accuracy of the

measurement of relativistic bending in the radio region is limited by the accuracy with which the bending by the ionized media can be accounted for. We now discuss the refraction in the case relevant to relativistic bending experiments, that is, the microwave region where the bending is small.

The electron density as a function of distance from the sun can be measured in a variety of ways. Optical observations of Thomson scattering during solar eclipses have been analyzed to give an electron density model

$$n_e = \left(1.55r^{-6} + 2.99r^{-16}\right) \times 10^{14}(\mathrm{m}^{-3}), \tag{13.174}$$

where r, the radial distance from the sun in units of the solar radius, is less than ~4. Equation (13.174) is the well-known Allen–Baumbach formula (Allen 1947). Eclipse data have also been interpreted by a model of the form

$$n_e = a_1 r^{-6} + a_2 r^{-2.33}, \tag{13.175}$$

for $r > 3$ (see references in Muhleman, Ekers, and Fomalont 1970). The coefficients a_1 and a_2 depend on solar activity and can vary by a factor of 5 during an 11-year cycle. Scintillation measurements of the occultation of the Crab Nebula at 26 MHz can be represented reasonably well by the model

$$n_e = 5 \times 10^{11} r^{-2}(\mathrm{m}^{-3}), \tag{13.176}$$

for $4 < r < 20$ (Erickson 1964, Evans and Hagfors 1968). Dispersion measurements of pulsars during solar occultation give about the same result as Eq. (13.176) (Counselman and Rankin 1972, Counselman et al. 1974). The angle of refraction of a ray passing near the sun can be calculated readily for the case where this angle is small. A ray obeys Snell's law in spherical coordinates, $nr \sin z$ = constant (Smart 1977), where n is the index of refraction and z is the angle between the ray and a line from the center of the sun, as shown in Fig. 13.24. From this relation, the bending angle is found to be

$$\theta_b = \pi - 2\int_{r_m}^{\infty} \frac{dr}{r\sqrt{(nr/p)^2 - 1}}, \tag{13.177}$$

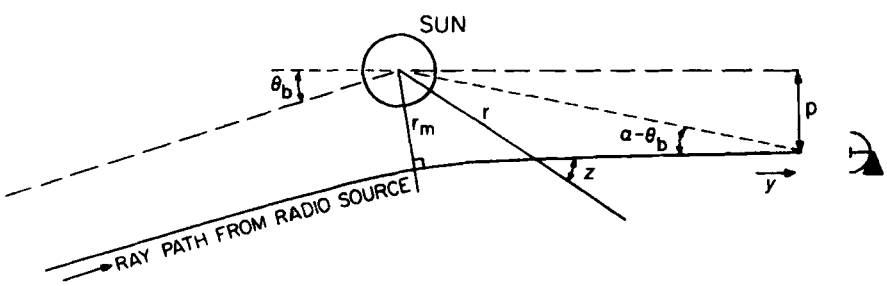


Figure 13.24 Path of a ray passing through the ionized gas surrounding the sun. p is the impact parameter and α is the solar elongation angle, that is, the angle between the sun and the source in the absence of solar bending.

where r_m is the distance of closest approach of the ray to the sun and p is the impact parameter (see Fig. 13.24). Assume that the electron density has a power-law distribution given by

$$n_e = n_{e0} r^{-\beta}, \tag{13.178}$$

where n_{e0} is the electron density in meters^{-3} at one solar radius and β is a constant. For a fully ionized solar wind, characterized by a constant mass loss rate and velocity, β is equal to 2. This case is applicable for $r \gtrsim 10$.

The index of refraction is obtained by substituting Eqs. (13.178) and (13.128) into Eq. (13.134) and neglecting the term in ν_B. Graphical solutions of Eq. (13.177) for large bending angles are given by Jaeger and Westfold (1950). For small bending angles, an approximate solution to Eq. (13.177) can be obtained by the use of the substitution $nr/p = \sec\theta$ [see also the discussion below Eq. (13.185)],

$$\theta_b \simeq 80.6\sqrt{\pi}\,\frac{n_{e0}}{\nu^2}\,\frac{\Gamma\left(\frac{\beta+1}{2}\right)}{\Gamma\left(\frac{\beta}{2}\right)}\,p^{-\beta}, \tag{13.179}$$

where p is in units of the solar radius and Γ is the gamma function. Note that the rays are bent away from the sun. The bending angle associated with the model in Eq. (13.176) is

$$\theta_b \simeq 2.4\lambda^2 p^{-2} \text{ (arcmin)}, \tag{13.180}$$

where λ is the wavelength in meters. For a multiple power-law model of electron density such as given in Eqs. (13.174) and (13.175), the bending angles for each component can be summed when the bending angles are small.

Relativistic Bending. The *general relativistic bending* can be described classically by an effective index of refraction given by $1 + 2GM_\odot/rc^2$, where G is the gravitational constant, and $M_\odot$ is the mass of the sun. The bending angle, for small values of p, is (Weinberg 1972)

$$\theta_{\text{GR}} \simeq -1.75 p^{-1} \text{ (arcsec)}. \tag{13.181}$$

The negative sign indicates that the bending is toward the sun. The general relativistic bending effect has been verified by VLBI observations to an accuracy of better than one part in a thousand (Lebach et al. 1995). In experiments to measure the relativistic bending, a model of the interferometer phase is formulated using Eq. (13.179) and a solar model with power-law components given by Eq. (13.178), with a separate density coefficient n_{e0i} for each component. The relativistic bending and the coefficients can be estimated simultaneously from the interferometer data. If, however, the electron density distribution has a component with $\beta = 1$, the relativistic effect would be masked if measurements were made

at only one frequency. The solar wind is highly variable, and attempts to characterize it by a power-law model with constant coefficients may not be adequate for high precision experiments.

At large angular distances from the sun, a more appropriate approximation for the general relativistic bending is (Misner, Thorne, and Wheeler 1973)

$$\theta_{\rm GR} \simeq -0.00407\sqrt{\frac{1+\cos\alpha}{1-\cos\alpha}}\ \text{(arcsec)}, \tag{13.182}$$

where α is the solar elongation, that is, the angle between the sun and the source. This bending, which is about 4.1 mas at $\alpha = 90°$, can be detected at almost all values of solar elongation with VLBI measurements (Robertson, Carter, and Dillinger 1991). Correction for the effect of relativistic bending must be made for many interferometric observations. Equation (13.182) is appropriate for a source at infinity. This equation must be modified if the source under investigation is within the solar system (Shapiro 1967).

The excess phase path for a ray passing through the corona, for the case where the effect of ray bending can be neglected, is, from Eq. (13.138),

$$\mathcal{L} \simeq -\frac{40.3}{\nu^2}\int_{-\infty}^{\infty} n_e\, dy, \tag{13.183}$$

where y is measured along the ray path as shown in Fig. 13.24. For a power-law model given by Eq. (13.178), the excess path is

$$\mathcal{L} \simeq -\frac{40.3 n_{e0}}{\nu^2}\int_{-\infty}^{\infty} \frac{dy}{(p^2+y^2)^{\beta/2}}, \tag{13.184}$$

which can be integrated to give

$$\mathcal{L} \simeq -\frac{40.3\sqrt{\pi}}{\nu^2}\frac{\Gamma\left(\frac{\beta-1}{2}\right)}{\Gamma\left(\frac{\beta}{2}\right)} n_{e0} p^{1-\beta}. \tag{13.185}$$

Note that the change in $\mathcal{L}$ with p describes the tilting of the wavefront and is the bending angle; hence $\theta_b \simeq d\mathcal{L}/dp$ (Bracewell, Eshleman, and Hollweg 1969). Differentiation of Eq. (13.185) with respect to p gives Eq. (13.179).

Interplanetary Scintillation

Detection of scintillation of extragalactic radio sources due to irregularities in the solar wind was reported by Clarke (1964) and Hewish, Scott, and Wills (1964). Interplanetary scintillation is readily distinguishable from ionospheric scintillation, since the timescale [Eq. (13.151)] and critical source size [Eq. (13.154)] are approximately 1 s and 0.5 arcsec for interplanetary scintillation and 30 s and 10 arcmin for ionospheric scintillation. Further observations of interplanetary scin-

tillation by Cohen et al. (1967) showed that the angular size of the radio source 3C273B is smaller than 0.02 arcsec, based on the application of Eq. (13.154). This result and the long-baseline interferometric results stimulated the development of VLBI. A comprehensive discussion of the interpretation of interplanetary scintillation can be found in Salpeter (1967), Young (1971), and Scott, Coles, and Bourgois (1983). For rough calculations, the scattering angle due to the interplanetary medium may be approximated by (Erickson 1964)

$$\theta_s \simeq 50\left(\frac{\lambda}{p}\right)^2 \text{ (arcmin)}, \tag{13.186}$$

where λ is in meters and p, the impact parameter, is in solar radii. This relationship is based on measurements taken in 1960–61 at 11 meters wavelength for impact parameters between 5 and 50 solar radii. Analysis of VLBI observations at 3.6 and 6 cm obtained in 1991 for a range of impact parameters of 10–50 solar radii led to a model for C_{ne}^2 of the form $C_{ne}^2 = 1.5 \times 10^{14}(r/R_{\text{sun}})^{-3.7}$ (Spangler and Sakurai 1995). Note that the power-law exponent is expected to be about -4 from the elementary consideration that C_{ne}^2 is proportional to the variance of the electron density, which is proportional to the square of the density. For a constant wind speed the density is proportional to r^{-2} and hence C_{ne}^2 is proportional to r^{-4}. Deviations from 4 are caused by the radial dependence of the magnetic field strength, which plays a role in driving the turbulence. Integrating C_{ne}^2 along the line of sight, and using Eq. (13.170), we derive an estimate for the scattering angle of $\theta_s = 3100(p/\lambda)^{-2.2}$ arcsec, which is comparable to the result in Eq. (13.186).

The concept that extended sources do not scintillate as much as point sources [see Eq. (13.154)] can be generalized to obtain more information about source structure. We assume that the scintillation is caused by a screen at a distance R from the earth as shown in Fig. 13.22, where $R \ll R_s$, and that the intensity at the earth is $I(x, y)$, where x and y are coordinates in a plane parallel to the screen in Fig. 13.21. The function $\Delta I(x, y)$ is equal to $I(x, y) - \langle I(x, y)\rangle$, where $\langle I(x, y)\rangle$ is the mean intensity. It has a power spectrum $\mathcal{S}_{I0}(q_x, q_y)$ for a point source and $\mathcal{S}_I(q_x, q_y)$ for an extended source, where q_x and q_y are the spatial frequencies (cycles per meter). If the visibility of the source is $\mathcal{V}(q_x R, q_y R)$, then it can be shown (Cohen 1969) that

$$\mathcal{S}_I(q_x, q_y) = \mathcal{S}_{I0}(q_x, q_y)\left|\mathcal{V}(q_x R, q_y R)\right|^2, \tag{13.187}$$

where $q_x R$ and $q_y R$ correspond to the projected baseline coordinates u and v. The scintillation index of the source m_s is defined by

$$m_s^2 = \frac{\langle \Delta I(x, y)^2\rangle}{\langle I(x, y)\rangle^2} = \frac{1}{\langle I(x, y)\rangle^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{S}_I(q_x, q_y)\, dq_x\, dq_y. \tag{13.188}$$

In principle, $\mathcal{S}_I(q_x, q_y)$ could be computed from the simultaneous measurements of $\Delta I(x, y)$ with a large number of spaced receivers. In practice, the motion of

the solar wind sweeps the diffraction pattern across a single telescope so that, from measurements of $\Delta I(t)$, the temporal power spectrum $\mathcal{S}(f)$ can be calculated. If the diffraction pattern moves with velocity v_s in the x direction, then $\mathcal{S}(f)$ can be related to the spatial spectrum since $q_x = f/v_s$:

$$\mathcal{S}(f) = \frac{1}{v_s}\int_{-\infty}^{\infty} \mathcal{S}_I\left(q_x = \frac{f}{v_s}, q_y\right) dq_y. \tag{13.189}$$

In principle, $|\mathcal{V}|^2$ can be recovered from Eq. (13.187) by observing a source over a range of different orientations with respect to the solar wind vector. The situation is entirely analogous to that of lunar occultation observations (Section 16.2) except that with lunar occultation observations the visibility phase can also be obtained. An estimate of the source diameter can be deduced from the width of the temporal power spectrum (Cohen, Gundermann, and Harris 1967) or from the scintillation index [Eq. (13.188)] (Little and Hewish 1966).

Interplanetary scattering is generally weak, except in directions close to the sun. The effects of refractive scattering (discussed in the next section and in Section 14.3), which can be important in the strong scattering regime, have been studied by Narayan, Anantharamaiah, and Cornwell (1989).

13.6 INTERSTELLAR MEDIUM

Table 13.7 lists the typical magnitudes and scale sizes of various effects caused by the interstellar medium. These are discussed individually in the following sections.

Dispersion and Faraday Rotation

The smooth, ionized component of the interstellar medium of our Galaxy affects propagation by introducing delay and Faraday rotation. The time of arrival of a pulse of radiation, such as that from a pulsar, is

$$t_p = \int_0^L \frac{dy}{v_g}, \tag{13.190}$$

where L is the propagation path, $v_g = cn$ is the group velocity, and n is given by Eq. (13.127). Differentiation of Eq. (13.190) gives

$$\frac{dt_p}{d\nu} \simeq -\frac{e^2}{4\pi\epsilon_0 mc\nu^3}\int_0^L n_e\, dy. \tag{13.191}$$

The integral of n_e over the path length is called the *dispersion measure*,

$$D_m = \int_0^L n_e\, dy, \tag{13.192}$$

TABLE 13.7 Typical Values[a] of the Effects of the Interstellar Medium on Radiation at 100 MHz

Effect	Equation Number	Magnitude	Frequency Dependence
Angular broadening[b]	13.163	0.3 arcsec	ν^{-2}
Pulse broadening[b]	13.152	10^{-4} s	ν^{-4}
Scintillation bandwidth[b]	13.153	10^4 Hz	ν^4
Spectral broadening[b]	—	1 Hz	ν^{-1}
Scintillation timescale[b]	13.151	10 s	ν^1
Scintillation timescale[c]	—	10^6 s	ν^{-1}
Free-free optical depth	13.142	0.01	ν^{-2}
Faraday rotation	13.193	10 rad	ν^{-2}

Adapted from Cordes (2000).
[a]For a source in the Galactic plane at a distance of 1 kpc. Actual values can differ by an order of magnitude.
[b]Diffractive scattering.
[c]Refractive scattering.

which is the same quantity as the total electron content. $dt_p/d\nu$ can be estimated by measuring the time of arrival of pulsar pulses at different frequencies, and the dispersion measure can then be found from Eq. (13.191). If the distance to the pulsar is known, then the average electron density can be calculated. A typical value of $\langle n_e \rangle$ in the plane of our Galaxy is 0.03 cm^{-3} (Weisberg, Rankin, and Boriakoff 1980). Alternatively, if a pulsar's distance is unknown, it can be estimated from Eq. (13.191) using an estimated average value of n_e

The magnetic field of the Galaxy causes Faraday rotation of the polarization plane of radiation from extragalactic radio sources. Equation (13.135) can be rewritten

$$\Delta\psi = \lambda^2 R_m, \tag{13.193}$$

where R_m is the *rotation measure* given by

$$R_m = 8.1 \times 10^5 \int n_e B_\parallel \, dy. \tag{13.194}$$

Here R_m is in radians per square meter, λ is in meters, $B_\parallel$ is the longitudinal component of magnetic field in gauss (1 gauss = 10^{-4} tesla), n_e is cm^{-3}, and dy is in parsecs (pc) (1 pc = 3.1×10^{16} m). The interstellar magnetic field can be estimated by dividing the rotation measure by the dispersion measure. Typical values of the magnetic field obtained in this way are 2 μG (Heiles 1976). This procedure underestimates the magnetic field if the field reverses direction along the line of sight. A formula for roughly estimating the rotation measure due to the galactic magnetic field is (Spitzer 1978)

$$R_m = -18|\cot b| \cos(\ell - 94°), \tag{13.195}$$

where ℓ and b are the galactic longitude and latitude. Extensive measurements of rotation measure as a function of direction can be found in Simard–Normandin and Kronberg (1980).

Faraday rotation that occurs within a radio source depolarizes the emergent radiation. This depolarization happens because radiation emitted from different depths in the source suffers different amounts of Faraday rotation. Such a source might be a relativistic gas emitting polarized synchrotron radiation immersed in a thermal plasma that causes the Faraday rotation. The degree of polarization of the observed radiation can be succinctly described in a Fourier transform relationship when self-absorption is negligible. We first introduce the function M, the complex degree of linear polarization, defined by

$$M = m_\ell \, e^{j2\psi} = \frac{Q + jU}{I}, \tag{13.196}$$

where m_ℓ is the degree of linear polarization, ψ is the position angle of the electric field, and Q, U, and I are the Stokes parameters as defined in Section 4.8 under *Parameters Defining Polarization*. If y is the linear distance into the source, $\psi(y)$ is the intrinsic position angle of the radiation at depth y, $j_\nu(y)$ is the volume emissivity of the source, and $\lambda^2\beta(y)$ is the Faraday rotation suffered by radiation emitted at depth y, then the degree of polarization of the observed radiation can be written

$$M(\lambda^2) = \frac{\int_0^\infty m_\ell(y)\, j_\nu(y)\, e^{j2[\psi(y)+\lambda^2\beta(y)]}\, dy}{\int_0^\infty j_\nu(y)\, dy}. \tag{13.197}$$

The denominator in Eq. (13.197) is the total intensity. $\beta(y)$ is the Faraday depth, which increases monotonically into the source as long as the sign of the longitudinal magnetic field direction does not change. In any case, we can superpose all the radiation from the same Faraday depth and write the integrals in Eq. (13.197) as a function of β instead of y, yielding

$$M(\lambda^2) = \int_{-\infty}^{\infty} F(\beta) e^{j2\lambda^2\beta}\, d\beta\,, \tag{13.198}$$

where

$$F(\beta) = \frac{m_\ell(y) j_\nu(y) e^{j2\psi(y)}}{\int_0^\infty j_\nu(y)\, dy}. \tag{13.199}$$

Thus $M(\lambda^2)$ and $F(\beta)$ form a Fourier transform pair. $F(\beta)$ is sometimes called the Faraday dispersion function. Unfortunately, $F(\beta)$, in general, cannot be found since M cannot be measured for negative values of λ^2. Because of this difficulty with the Fourier transform, $F(\beta)$ is usually estimated by model fitting. How-

ever, if $\psi(y)$ is constant, then $M(-\lambda^2) = M^*(\lambda^2)$, and $F(\beta)$ can be obtained by Fourier transformation.

Consider the result for a simple source model for which m_ℓ, ψ, and j_ν are constant. From Eq. (13.198), we have

$$M(\lambda^2) = M(0)\left[\frac{\sin \lambda^2 R_m}{\lambda^2 R_m}\right] e^{j\lambda^2 R_m}, \tag{13.200}$$

where R_m is the Faraday rotation measure through the whole source. If the Faraday rotation originates in front of the radiation source, the complex degree of polarization is

$$M(\lambda^2) = M(0) e^{j2\lambda^2 R_m}. \tag{13.201}$$

In this case there is no depolarization, and the Faraday rotation is twice that of Eq. (13.200), in which the source is uniformly distributed throughout the rotation medium. For detailed treatment of intrinsic Faraday rotation, see Burn (1966) and Gardner and Whiteoak (1966).

Diffractive Scattering

Diffractive interstellar scattering has been extensively investigated by observation of pulsars and compact extragalactic radio sources. For pulsars, the temporal broadening of the pulses [Eq. (13.152)], the decorrelation bandwidth [Eq. (13.153)], and the angular broadening [Eq. (13.147)] can be measured. Interpretation of the measurements in terms of a thin-screen model suggests that $\Delta n_e/n_e \simeq 10^{-3}$, and that the scale size responsible for the scintillation is on the order of 10^{11} cm. The temporal variations or scintillation of the signal from a pulsar are caused by the motions of the observer and the pulsar relative to the quasi-stationary interstellar medium. A measurement of the decorrelation bandwidth can be used to estimate the scattering angle [Eq. (13.153)]. This estimate of the scattering angle and the measurement of the timescale of fading (10^2–10^3 s at 408 MHz) can be used to estimate the relative velocity of the scattering screen by Eq. (13.151). From the relative velocity of the screen, the transverse velocity of the pulsar can be found. Velocities, and thus proper motions, of pulsars estimated in this way (Lyne and Smith 1982) agree with those measured directly with interferometers [see, e.g., Campbell et al. (1996)]. The transverse component of the orbital velocity of a binary pulsar has also been measured (Lyne 1984).

Observations show that the fluctuations in electron density can be described by a power-law spectrum with a power-law exponent of about 3.7 ± 0.3, which is similar to the value of $11/3$ for Kolmogorov turbulence (Rickett 1990; Cordes, Pidwerbetsky, and Lovelace 1986). The power-law spectrum appears to extend over a range of scale sizes from less than 10^{10} cm to more than 10^{15} cm. The inner scale may be set by the proton gyrofrequency ($\sim 10^7$ cm) and the outer scale by the scale height of the Galaxy ($\sim 10^{20}$ cm). Observational evidence for the inner scale is given by Spangler and Gwinn (1990).

Extensive measurements of the angular sizes of extragalactic radio sources have been used to derive an approximate formula for θ_s based on the Gaussian screen model, by Harris, Zeissig, and Lovelace (1970), Readhead and Hewish (1972), Cohen and Cronyn (1974), Duffett-Smith and Readhead (1976), and others. This formula is

$$\theta_s \simeq \frac{15}{\sqrt{|\sin b|}} \lambda^2 \text{ (mas)}, \qquad |b| > 15^\circ \tag{13.202}$$

where b is the Galactic latitude and λ is the wavelength in meters. The pulsar data have been interpreted by Cordes (1984) in terms of the power-law model to arrive at approximate formulas for θ_s:

$$\begin{aligned} \theta_s &\simeq 7.5\lambda^{11/5} \text{ (arcsec)}, && |b| \leq 0.6^\circ \\ &\simeq 0.5|\sin b|^{-3/5}\lambda^{11/5} \text{ (arcsec)}, && 0.6^\circ < |b| < 3^\circ\text{–}5^\circ \\ &\simeq 13|\sin b|^{-3/5}\lambda^{11/5} \text{ (mas)}, && |b| \geq 3^\circ\text{–}5^\circ. \end{aligned} \tag{13.203}$$

The accuracy of the representations in Eqs. (13.203) decreases with decreasing $|b|$. In particular, the scattering angle at low latitudes, $|b| < 1^\circ$, can take on a wide range of values (Cordes, Ananthakrishnan, and Dennison 1984). A much more detailed model with 23 parameters characterizing the electron distribution in the Galaxy has been constructed by Taylor and Cordes (1993). From this model more accurate estimates of θ_s can be computed.

An example of a compact radio source that suffers a high degree of interstellar scattering is Sagittarius A* at the dynamical center of our Galaxy. This source has an angular size of about 15 mas at a wavelength of 3.6 cm [compared with 7.7 mas predicted by Eq. (13.203)]. The angular size varies approximately as the wavelength squared over the entire measuring range ~0.3–30 cm, as shown in Fig. 13.25. The data at $\lambda = 0.7$ and 0.3 cm suggest that the structure of the source may be visible at $\lambda < 0.7$ mm (Lo et al. 1998).

Interstellar scattering probably places the ultimate limit on the angular size measurements that can be made with an interferometer. The apparent sizes of interstellar masers, which are mostly found in the Galaxy at low galactic latitudes, are sometimes set by interstellar scattering (Gwinn et al. 1988).

Refractive Scattering

The realization by Sieber (1982) that the characteristic period of amplitude scintillations of pulsars, on timescales of days to months, were correlated with their dispersion measures led Rickett, Coles, and Bourgois (1984) to the identification of another important scale length in the turbulent interstellar medium, the refractive scale d_{ref}. The refractive scale is the size of the diffractive scattering disk, which is the projection of the cone of scattered radiation on the scattering screen, located a distance R from the observer. The diameter of the diffractive scattering

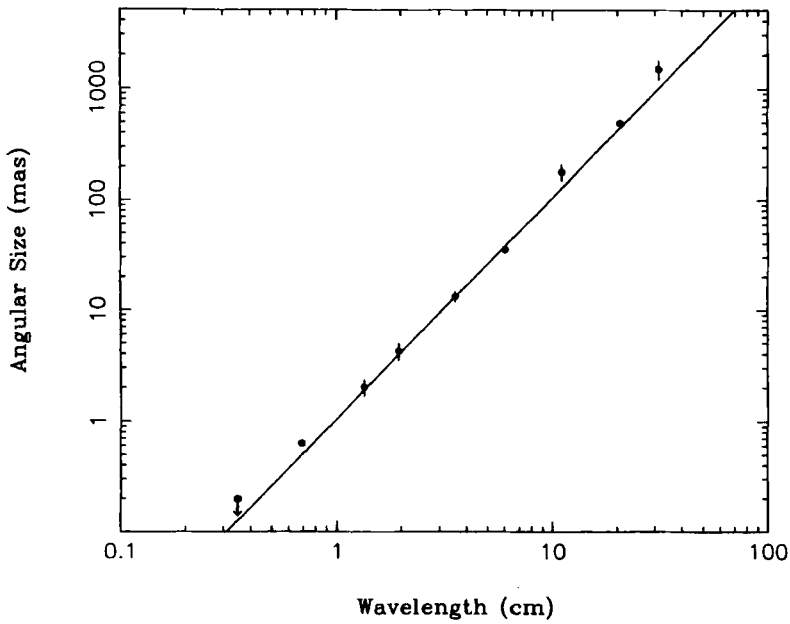


Figure 13.25 The angular size of the compact source in the galactic center (Sgr A*) versus wavelength. The scattering is slightly anisotropic [e.g., Lo et al. (1998)], and each data point plotted is the geometric mean of the angular diameters in right ascension and declination. The line through the data has the form $\theta_s = 1.04\lambda^{2.0}$, where θ_s is in milliarcseconds and λ is in centimeters. The $\lambda^{2.0}$ dependence suggests that the interferometer baselines (10^6 m or less) are all shorter than the inner scale of the turbulence [see Eqs. (13.168) and (13.171)].

disk is $R\theta_s$. The scattering disk represents the maximum extent on the screen from which radiation can reach the observer. With a power-law distribution of irregularities, it is the irregularities at the maximum allowed scale that have the largest amplitude and are the most influential. Thus, the refractive scale is $d_{\rm ref} \simeq R\theta_s$. Since $\theta_s = \lambda/d_0$, where d_0 is the diffractive scale size defined by $\mathcal{D}_\phi(d_0) = 1$, we can write

$$d_{\rm ref} = \frac{R\lambda}{d_0}, \tag{13.204}$$

or

$$d_{\rm ref} = \frac{d_{\rm Fresnel}^2}{d_0}, \tag{13.205}$$

where $d_{\rm Fresnel} = \sqrt{R\lambda}$ is the Fresnel scale. The scale lengths $d_{\rm ref}$ and d_0 are widely separated. Hence, the timescale associated with scintillation scattering for a screen velocity of v_s, $t_{\rm ref} = d_{\rm ref}/v_s$, is much longer than that associated with

diffractive scattering, $t_{\rm dif} = d_{\rm dif}/v_s$. Suppose that a source is observed through a scattering screen located at a distance of 1 kpc, at $b \simeq 20°$, and a wavelength of 0.5 m. For this case the diffractive scale length is 2×10^9 cm, the Fresnel scale is 4×10^{11} cm, and the refractive scale is 8×10^{13} cm. The typical velocity associated with the ISM is 100 km s^{-1} (Rickett, Coles, and Bourgois 1984). For this velocity the diffractive and refractive timescales for amplitude scintillation are 3 min and 3 months, respectively.

Refractive scattering is thought to be responsible for the slow amplitude variations observed in some pulsars and quasars at meter and decimeter wavelengths. This realization solved the long-standing problem of understanding the behavior of "long-wavelength variables," which could not be explained by intrinsic variability models based on synchrotron emission. The identification of two scales in the interstellar scattering medium provides strong support for the power-law model. The two scales provide a way of estimating the power-law index, because the relative importance of refractive scattering increases as the power spectrum steepens. It is interesting to note that these two scales arise from a power-law phenomenon, which has no intrinsic scale. The scales are related to the propagation and depend on the wavelength and distance of the screen.

In addition to amplitude scintillation, refractive scattering causes the apparent position of the source to wander with time. The amplitude and timescale are about equal to θ_s and $t_{\rm ref}$, respectively. The character of this wander depends on the power-law index of the fluctuations. Limits on the power-law index have been established from the limits on the amplitude of image wander in the relative positions among clusters of masers (Gwinn et al. 1988).

Rare sudden changes in the intensity of several extragalactic sources, called *Fiedler events*, or *extreme scattering events* (Fiedler et al. 1987), are probably caused by refractive scattering in the interstellar medium. In the archetypal example, the flux density of the extragalactic source 0954+658 increased by 30% and then dropped by 50% over a period of a month, after which it recovered in symmetric fashion. A large-scale plasma cloud presumably drifted between the source and the earth, creating flux density changes due to focusing and refraction.

Because there are two timescales associated with strong scattering in the interstellar medium, three distinct data averaging regimes are important for constructing images from interferometry data obtained on a timescale $t_{\rm int}$. These are: $t_{\rm int} > t_{\rm ref}$ (ensemble average image), $t_{\rm ref} > t_{\rm int} > t_{\rm dif}$ (average image), and $t_{\rm int} < t_{\rm dif}$ (snapshot image). The characteristics of these image regimes are described by Narayan (1992), Narayan and Goodman (1989), and Goodman and Narayan (1989). For ensemble averaging [see Eqs. (13.168) through (13.170)], the image is essentially convolved with the appropriate "seeing" function. The snapshot regime offers intriguing possibilities for image restoration. In this regime it should be possible to image the source with a resolution of $\lambda/d_{\rm ref}$, which can be very much smaller than that achievable with terrestrial interferometry. In this case the scattering screen functions as the aperture of the interferometer. Because of the multipath propagation provided by refractive scattering, which brings radiation from widely separated parts of the scattering screen to the observer, the effective

baselines can be very large. See Section 14.3 for further discussion, including an observation by Wolszczan and Cordes (1987).

BIBLIOGRAPHY

Baldwin, J. E. and Wang Shougun, Eds., *Radio Astronomical Seeing,* International Academic Publishers, Pergamon Press, Oxford, 1990.

Cordes, J. M., B. J. Rickett, and D. C. Backer, *Radio Wave Scattering in the Interstellar Medium,* American Institute of Physics Conference Proceedings 174, New York, 1988.

Janssen, M. A., *Atmospheric Remote Sensing by Microwave Radiometry,* Wiley, New York, 1993.

Narayan, R., The Physics of Pulsar Scintillation, *Phil. Tran. R. Soc. Lond. A,* **341**, 151–165, 1992.

Tatarski, V. I., *Wave Propagation in a Turbulent Medium*, Dover, New York, 1961.

Westwater, R., Ed., *Specialist Meeting on Microwave Radiometry and Remote Sensing Applications,* National Oceanic and Atmospheric Administration, U.S. Dept. Commerce, 1992.

REFERENCES

Aarons, J., Global Morphology of Ionospheric Scintillations, *Proc. IEEE*, **70**, 360–378, 1982.

Aarons, J., J. A. Klobuchar, H. E. Whitney, J. Austen, A. L., Johnson, and C. L. Rino, Gigahertz Scintillations Associated with Equatorial Patches, *Radio Sci.,* **18**, 421–434, 1983.

Aarons, J., M. Mendillo, B. Lin, M. Colerico, T. Beach, P. Kintner, J. Scali, B. Reinisch, G. Sales, and E. Kudeki, Equatorial F-Region Irregularity Morphology during an Equinoctial Month at Solar Minimum, *Space Science Reviews*, **87**, 357–386, 1999.

Allen, C. W., Interpretation of Electron Densities from Corona Brightness, *Mon. Not. R. Astron. Soc.,* **107**, 426–432, 1947.

Altenhoff, W. J., J. W. M. Baars, D. Downes, and J. E. Wink, Observations of Anomalous Refraction at Radio Wavelengths, *Astron. Astrophys.,* **184**, 381–385, 1987.

Appleton, E. V. and M. A. F. Barnett, On Some Direct Evidence for Downward Atmospheric Reflection of Electric Rays, *Proc. R. Soc. Lond. A,* **109**, 621–641, 1925.

Armstrong, J. W. and R. A. Sramek, Observations of Tropospheric Phase Scintillations at 5 GHz on Vertical Paths, *Radio Sci.,* **17**, 1579–1586, 1982.

Baars, J. W. M., Meteorological Influences on Radio Interferometer Phase Fluctuations, *IEEE Trans. Antennas Propag.,* **AP-15**, 582–584, 1967.

Bailey, D. K., On a New Method of Exploring the Upper Atmosphere, *J. Terr. Mag. Atmos. Elec.,* **53**, 41–50, 1948.

Bean, B. R., B. A. Cahoon, C. A. Samson, and G. D. Thayer, *A World Atlas of Atmospheric Refractivity,* U.S. Government Printing Office, Washington, DC, 1966.

Bean, B. R. and E. J. Dutton, *Radio Meteorology,* National Bureau of Standards Monograph 92, U.S. Government Printing Office, Washington, DC, 1966.

Beynon, W. J. G., Marconi, Radio Waves, and the Ionosphere, *Radio Sci.,* **10**, 657–664, 1975.

Bieging, J. H., J. Morgan, J. H., J. Morgan, W. J. Welch, S. N. Vogel, and M. C. H. Wright, Interferometer Measurements of Atmospheric Phase Noise at 86 GHz, *Radio Sci.*, **19**, 1505–1509, 1984.

Bilitza, D., International Reference Ionosphere—Status 1995/96, *Adv. Space Res.*, **20**, 1751–1754, 1997.

Bohlander, R. A., R. W. McMillan, and J. J. Gallagher, Atmospheric Effects on Near-Millimeter-Wave Propagation, *Proc. IEEE*, **73**, 49–60, 1985.

Booker, H. G., The Use of Radio Stars to Study Irregular Refraction of Radio Waves in the Ionosphere, *Proc. IRE*, **46**, 298–314, 1958.

Booker, H. G., J. A. Ratcliffe, and D. H. Shinn, Diffraction from an Irregular Screen with Applications to Ionospheric Problems, *Philos. Tran. R. Soc. Lond. A*, **242**, 579–607, 1950.

Bracewell, R.N., *The Fourier Transform and Its Applications*, 3rd ed., McGraw-Hill, New York, 2000.

Bracewell, R. N., V. R. Eshleman, and J. V. Hollweg, The Occulting Disk of the Sun at Radio Wavelengths, *Astrophys. J.*, **155**, 367–368, 1969.

Breit, G. and M. A. Tuve, A Test of the Existence of the Conducting Layer, *Phys. Rev.*, **28**, 554-575, 1926.

Budden, K. G., *Radio Waves in the Ionosphere*, Cambridge Univ. Press, Cambridge, UK, 1961.

Burn, B. J., On the Depolarization of Discrete Radio Sources by Faraday Dispersion, *Mon. Not. R. Astron. Soc.*, **133**, 67–83, 1966.

Butler, B., Precipitable Water at the VLA—1990–1998, MMA Memo. 237, Nat. Radio Astron. Obs., Socorro, NM, 1998.

Campbell, R. M., N. Bartel, I. I. Shapiro, M. I. Ratner, R. J. Cappallo, A. R. Whitney, and N. Putnam, VLBI-Derived Trigonometric Parallax and Proper Motion of PSR B2021+51, *Astrophys. J.*, **461**, L95–L98, 1996.

Carilli, C. L. and M. A. Holdaway, Tropospheric Phase Calibration in Millimeter Interferometry, *Radio Science*, **34**, 817–840, 1999.

Chamberlin, R. A. and J. Bally, The Observed Relationship Between the South Pole 225-GHz Atomosphere Opacity and the Water Vapor Column Density, *Int. J. Infrared and Millimeter Waves*, **16**, 907–920, 1995.

Chamberlin, R. A., A. P. Lane, and A. A. Stark, The 492 GHz Atmospheric Opacity at the Geographic South Pole, *Astrophys. J.*, **476**, 428–433, 1997.

Chylek, P., and D. J. W. Geldart, Water Vapor Dimers and Atmospheric Absorption of Electromagnetic Radiation, *Geophys. Res. Lett.*, **24**, 2015–2018, 1997.

Clarke, M., *Two Topics in Radiophysics*, Ph.D. thesis, Cambridge Univ., 1964 (see App. II).

COESA, *U.S. Standard Astmosphere, 1976*, NOAA-S/T 76-1562, U.S. Government Printing Office, Washington, DC, 1976.

Cohen, M. H., High-Resolution Observations of Radio Sources, *Ann. Rev. Astron, Astrophys.*, **7**, 619–664, 1969.

Cohen, M. H. and W. M. Cronyn, Scintillation and Apparent Angular Diameter, *Astrophys. J.*, **192**, 193–197, 1974.

Cohen, M. H., E. J. Gundermann, H. E. Hardebeck, and L. E. Sharp, Interplanetary Scintillations. II Observations, *Astrophys. J.*, **147**, 449–466, 1967.

Cohen, M. H., E. J. Gundermann, and D. E. Harris, New Limits on the Diameters of Radio Sources, *Astrophys. J.*, **150**, 767–782, 1967.

Cordes, J. M., Interstellar Scattering, in *VLBI and Compact Radio Sources*, R. Fanti, K. Kellermann, and G. Setti, Eds., IAU Symp. 110, Reidel, Dordrecht, Netherlands, 1984, pp. 303–307.

Cordes, J. M., Interstellar Scattering: Radio Sensing of Deep Space through the Turbulent Interstellar Medium, in *Radio Astronomy at Long Wavelengths*, R. G. Stone, K. W. Weiler, M. L. Goldstein, and J.-L. Bougeret, Eds., Geophysical Monograph 119, Amer. Geophys. Union, pp. 105–114, 2000.

Cordes, J. M., S. Ananthakrishnan, and B. Dennison, Radio Wave Scattering in the Galactic Disk, *Nature*, **309**, 689–691, 1984.

Cordes, J. M., A. Pidwerbetsky, and R. V. E. Lovelace, Refractive and Diffractive Scattering in the Interstellar Medium, *Astrophys. J.*, **310**, 737–767, 1986.

Coulman, C. E., Fundamental and Applied Aspects of Astronomical "Seeing," *Ann. Rev. Astron. Astrophys.*, **23**, 19–57, 1985.

Coulman, C. E., Atmospheric Structure, Turbulence and Radioastronomical "Seeing," in *Radio Astronomical Seeing*, J. E. Baldwin and Wang Shouguan, Eds., International Academic Publishers, Pergamon Press, Oxford, 1990, pp. 11–20.

Counselman, C. C., III., S. M. Kent, C. A. Knight, I. I. Shapiro, T. A. Clark, H. F. Hinteregger, A. E. E. Rogers, and A. R. Whitney, Solar Gravitational Deflection of Radio Waves Measured by Very-Long-Baseline Interferometry, *Phys. Rev. Lett.*, **33**, 1621–1623, 1974.

Counselman, C. C., III and J. M. Rankin, Density of the Solar Corona from Occultations of NP0532, *Astrophys. J.*, **175**, 843–856, 1972.

Cox, A. N., Ed., *Allen's Astrophysical Quantities*, 4th ed., AIP Press, New York, Springer, 2000, Sec. 11.20, p. 262.

Crane, R. K., Refraction Effects in the Neutral Atmosphere, in *Methods of Experimental Physics*, Vol. 12B, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 186–200.

Crane, R. K., Ionospheric Scintillation, *Proc. IEEE*, **65**, 180–199, 1977.

Crane, R. K., Fundamental Limitations Caused by RF Propagation, *Proc. IEEE*, **69**, 196–209, 1981.

Daniell, R. E., L. D. Brown, D. N. Anderson, M. W. Fox, P. H. Doherty, D. T. Decker, J. J. Sojka, and R. W. Schunk, Parameterized Ionospheric Model: A Global Ionospheric Parameterization Based on First Principles Models, *Radio Sci.*, **30**, 1499–1510, 1995.

Davies, K., *Ionospheric Radio Propagation*, National Bureau of Standards Monograph 80, U.S. Government Printing Office, Washington, DC, 1965.

Davis, J. L., T. A. Herring, I. I. Shapiro, A. E. E. Rogers, and G. Elgered, Geodesy by Radio Interferometry: Effects of Atmospheric Modeling Errors on Estimates of Baseline Length, *Radio Sci.*, **20**, 1593–1607, 1985.

Debye, P., *Polar Molecules*, Dover, New York, 1929.

Delgado, G., A. Otárola, V. Belitsky, and D. Urbain, The Determination of Precipitable Water Vapour at Llano de Chajnantor from Observations of the 183 GHz Water Line, ALMA Memo 271, National Radio Astronomy Observatory, Socorro, NM, 1998

Downes, D. and W. J. Altenhoff, Anomalous Refraction at Radio Wavelengths, in *Radio Astronomical Seeing*, J. E. Baldwin and Wang Shouguan, Eds., International Academic Publishers, Pergamon Press, Oxford, 1990, pp. 31–40.

Duffett-Smith, P. J. and A. C. S. Readhead, The Angular Broadening of Radio Sources by Scattering in the Interstellar Medium, *Mon. Not. R. Astron. Soc.*, **174**, 7–17, 1976.

Elgered, G., J. L. Davis, T. A. Herring, and I. I. Shapiro, Geodesy by Radio Interferometry: Water Vapor Radiometry for Estimation of the Wet Delay, *J. Geophys. Res.*, **96**, 6541–6555, 1991.

Elgered, G., B. O. Rönnäng, and J. I. H. Askne, Measurements of Atmospheric Water Vapor with Microwave Radiometry, *Radio Sci.*, **17**, 1258–1264, 1982.

Erickson, W. C., The Radio-Wave Scattering Properties of the Solar Corona, *Astrophys. J.*, **139**, 1290–1311, 1964.

Evans, J. V., Theory and Practice of Ionospheric Study by Thomson Scatter Radar, *Proc. IEEE*, **57**, 496–530, 1969.

Evans, J. V. and T. Hagfors, *Radar Astronomy*, McGraw-Hill, New York, 1968.

Evans, J. V., J. M. Holt, and R. H. Wand, A Differential-Doppler Study of Traveling Ionospheric Disturbances from Millstone Hill, *Radio Sci.*, **18**, 435–451, 1983.

Fejer, B. G. and M. C. Kelley, Ionospheric Irregularities, *Rev. Geophys. Space Sci.*, **18**, 401–454, 1980.

Fiedler, R. L., B. Dennison, K. J. Johnston, and A. Hewish, Extreme Scattering Events Caused by Compact Structures in the Interstellar Medium, *Nature*, **326**, 675–678, 1987.

Fomalont, E. B. and R. A. Sramek, A Confirmation of Einstein's General Theory of Relativity by Measuring the Bending of Microwave Radiation in the Gravitational Field of the Sun, *Astrophys. J.*, **199**, 749–755, 1975.

Fomalont, E. B. and R. A. Sramek, The Deflection of Radio Waves by the Sun, *Comments Astrophys.*, **7**, 19–33, 1977.

Frail, D. A., S. R. Kulkarni, L. Nicastro, M. Ferocl, and G. B. Taylor, the Radio Afterglow from the γ-ray Burst of 8 May 1997, *Nature*, **389**, 261–263, 1997.

Freeman, R. L., *Radio System Design for Telecommunications* (1–100 GHz), Wiley, New York, 1987.

Fried, D. L., Statistics of a Geometric Representation of Wavefront Distortion, *J. Opt. Soc. Am.*, **55**, 1427–1435, 1965.

Fried, D. L., Optical Resolution Through a Randomly Inhomogeneous Medium for Very Lang and Very short Exposures, *J. Opt. Soc. Am.*, **56**, 1372–1379, 1966.

Fried, D. L., Optical Heterodyne Detection of an Atmospherically Distorted Signal Wave Front, *Proc. IEEE*, **55**, 57–67, 1967.

Gardner, F. F. and J. B. Whiteoak, The Polarization of Cosmic Radio Waves, *Ann. Rev. Astron. Astrophys.*, **4**, 245–292, 1966.

Goldstein, H., Attenuation by Condensed Water, in *Propagation of Short Radio Waves*, MIT Radiation Laboratory Series, Vol. 13, D. E. Kerr, Ed., McGraw-Hill, New York, 1951, p. 671.

Goodman, J. and R. Narayan, The Shape of a Scatter-Broadened Image: II. Interferometric Visibilities, *Mon. Not. R. Astron. Soc.*, **238**, 995–1028, 1989.

Guiraud, F. O., J. Howard, and D. C. Hogg, A Dual-Channel Microwave Radiometer for Measurement of Preciptable Water Vapor and Liquid, *IEEE Trans. Geosci, Electron.*, **GE-17**, 129–136, 1979.

Gupta, Y., Pulsars and Interstellar Scintillations, in *Pulsar Astrometry—2000 and Beyond*, M. Kramer, N. Wex, and R. Wielebinski, Eds., Astron. Soc. Pacific Conf. Ser., **202**, 539–544, 2000.

Gwinn, C. R., J. M. Moran, M. J. Reid, and M. H. Schneps, Limits on Refractive Interstellar Scattering Toward Saggitarius B2, *Astrophys. J.*, **330**, 817–827, 1988.

Hagfors, T., The Ionosphere, in *Methods of Experimental Physics*, Vol. 12B, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 119–135.

Hamaker, J. P., Atmospheric Delay Fluctuations with Scale Sizes Greater than One Kilometer, Observed with a Radio Interferometer Array, *Radio Sci.*, **13**, 873–891, 1978.

Harris, D. E., G. A. Zeissig, and R. V. Lovelace, The Minimum Observable Diameter of Radio Sources, *Astron. Astrophys.*, **8**, 98–104, 1970.

Heiles, C., The Interstellar Magnetic Field, *Ann. Rev. Astron. Astrophys.*, **14**, 1–22, 1976.

Hess, S. L., *Introduction to Theoretical Meteorology*, Holt, Rinehart, Winston, New York, 1959.

Hewish, A., The Diffraction of Galactic Radio Waves as a Method of Investigating the Irregular Structure of the Ionosphere, *Proc. R. Soc. Lond. A*, **214**, 494–514, 1952.

Hewish, A., P. F. Scott, and D. Wills, Interplanetary Scintillation of Small Diameter Radio Sources, *Nature*, **203**, 1214–1217, 1964.

Hey, J. S., S. J. Parsons, and J. W. Phillips, Fluctuations in Cosmic Radiation at Radio Frequencies, *Nature*, **158**, 234, 1946.

Hill, R. J., Water Vapor-Absorption Line Shape Comparison Using the 22-GHz Line: the Van Vleck–Weisskopf Shape Affirmed, *Radio Sci.*, **21**, 447–451, 1986.

Hill, R. J. and S. F. Clifford, Contribution of Water Vapor Monomer Resonances to Fluctuations of Refraction and Absorption for Submillimeter through Centimeter Wavelengths, *Radio Sci.*, **16**, 77–82, 1981.

Hill, R. J., R. S. Lawrence, and J. T. Priestly, Theoretical and Calculational Aspects of the Radio Refractive Index of Water Vapor, *Radio Sci.*, **17**, 1251–1257, 1982.

Hills, R. E., A. S. Webster, D. A. Alston, P. L. R. Morse, C. C. Zammit, D. H. Martin, D. P. Rice, and E. I., Robson, Absolute Measurements of Atmospheric Emission and Absorption in the Range 100–1000 GHz, *Infrared Phys.*, **18**, 819–825, 1978

Hinder, R. A., Observations of Atmospheric Turbulence with a Radio Telescope at 5 GHz, *Nature*, **225**, 614–617, 1970.

Hinder, R. A., Fluctuations of Water Vapour Content in the Troposphere as Derived from Interferometric Observations of Celestial Radio Sources, *J. Atmos. Terr. Phys.*, **34**, 1171–1186, 1972.

Hinder, R. A. and M. Ryle, Atmospheric Limitations to the Angular Resolution of Aperture Synthesis Radio Telescopes, *Mon. Not. R. Astron. Soc.*, **154**, 229–253, 1971.

Ho, C. M., A. J. Mannucci, U. J. Lindqwister, X. Pi, and B. T. Tsurutani, Global Ionospheric Perturbations Monitored by the Worldwide GPS Network, *Geophys. Res. Lett.*, **23**, 3219–3222, 1996.

Ho, C. M., B. D. Wilson, A. J. Mannucci, U. J. Lindqwister, and D. N. Yuan, A Comparative Study of Ionospheric Total Electron Content Measurements Using Global Ionospheric Maps of GPS, TOPEX Radar, and the Bent Model, *Radio Sci.*, **32**, 1499–1512, 1997.

Hocke, K. and K. Schlegel, A Review of Atmospheric Gravity Waves and Travelling Ionospheric Disturbances: 1982–1995, *Ann. Geophysicae*, **14**, 917–940, 1996.

Hogg, D. C., F. O. Guiraud, and W. B. Sweezy, The Short–Term Temporal Spectrum of Precipitable Water Vapor, *Science*, **213**, 1112–1113, 1981.

Holdaway, M. A., M. Ishiguro, S. M. Foster, R. Kawabe, K. Kohno, F. N. Owen, S. J. E. Radford, and M. Saito, *Comparison of Rio Frio and Chajnantor Site Testing Data*, MMA Memo 152, National Radio Astronomy Observatory, Socorro, NM, 1996.

Holdaway, M. A., S. J. E. Radford, F. N. Owen, and S. M. Foster, *Fast Switching Phase Calibration: Effectiveness at Mauna Kea and Chajnantor*, MMA Memo 139, National Radio Astronomy Observatory, Socorro, NM, 1995.

Holt, E. H. and R. E. Haskell, *Foundations of Plasma Dynamics*, Macmillan, New York, 1965, p. 254.

Humphreys, W. J., *Physics of the Air*, 3rd ed., McGraw-Hill, New York, 1940.

Hunsucker, R. D., Atmospheric Gravity Waves Generated in the High-Latitude Ionosphere: A Review, *Rev. Geophys. Space Phys.*, **20**, 293–315, 1982.

Ishiguro, M., T. Kanzawa, and T. Kasuga, Monitoring of Atmospheric Phase Fluctuations Using Geostationary Satellite Signals, in *Radio Astronomical Seeing*, J. E. Baldwin and Wang Shouguan, Eds., International Academic Publishers, Pergamon Press, Oxford, 1990, pp. 60–63.

Jackson, J. D., *Classical Electrodynamics*, 3rd ed., Wiley, New York, 1999, pp. 775–784.

Jaeger, J. C. and K. C. Westfold, Equivalent Path and Absorption for Electromagnetic Radiation in the Solar Corona, *Aust. J. Phys.*, **3**, 376–386, 1950.

Kaplan, G. H., F. J. Josties, P. E. Angerhofer, K. J. Johnston, and J. H. Spencer, Precise Radio Source Positions from Interferometric Observations, *Astron. J.*, **87**, 570–576, 1982.

Kasuga, T., M. Ishiguro, and R. Kawabe, Interferometric Measurement of Tropospheric Phase Fluctuations at 22 GHz on Antenna Spacings of 27 to 540 m, *IEEE Trans. Antennas Propag.*, **AP-34**, 797–803, 1986.

Kundu, M. R., *Solar Radio Astronomy*, Wiley-Interscience, New York, 1965, p. 104.

Lawrence, R. S., C. G. Little, and H. J. A. Chivers, A Survey of Ionospheric Effects upon Earth-Space Radio Propagation, *Proc. IEEE*, **52**, 4–27, 1964.

Lay, O. P., The Temporal Power Spectrum of Atmospheric Fluctuations Due to Water Vapor, *Astron. Astrophys. Suppl.*, **122**, 535–545, 1997a.

Lay, O. P., Phase Calibration and Water Vapor Radiometry for Millimeter-Wave Arrays, *Astron. Astrophys. Suppl.*, **122**, 547–557, 1997b.

Lay, O. P., *183 GHz Radiometric Phase Correction for the Millimeter Array*, MMA Memo. 209, National Radio Astronomy Observatory, Socorro, NM, 1998.

Lazio, T. J. W. and J. M. Cordes, Hyperstrong Radio-Wave Scattering in the Galactic Center. I. A Survey for Extragalactic Sources Seen through the Galactic Center, *Astrophys. J. Suppl.*, **118**, 201–216, 1998.

Lebach, D. E., B. E. Corey, I. I. Shapiro, M. I. Ratner, J. C. Webber, A. E. E. Rogers, J. L. Davis, and T. A. Herring, Measurements of the Solar Deflection of Radio Waves Using Very Long Baseline Interferometry, *Phys. Rev. Lett.*, **75**, 1439–1442, 1995.

Liebe, H. J., Calculated Tropospheric Dispersion and Absorption Due to the 22-GHz Water Vapor Line, *IEEE Trans. Antennas Propag.*, **AP-17**, 621–627, 1969.

Liebe, H. J., Modeling Attenuation and Phase of Radio Waves in Air at Frequencies below 1000 GHz, *Radio Sci.*, **16**, 1183–1199, 1981.

Liebe, H. J., An Updated Model for Millimeter Wave Propagation in Moist Air, *Radio Sci.*, **20**, 1069–1089, 1985.

Liebe, H. J., MPM-An Atmospheric Millimeter-Wave Propagation Model, *Int. J. Infrared and MM Waves*, **10**, 631–650, 1989.

Little, L. T., and A. Hewish, Interplanetary Scintillation and Relation to the Angular Structure of Radio Souces, *Mon. Not. R. Astron. Soc.*, **134**, 221–237, 1966.

Lo, K. Y., Z.-Q. Shen, J. H. Zhao, and P. T. P. Ho, Instrinsic Size of Sagittarius A*: 72 Schwarzschild Radii, *Astrophys. J.*, **508**, L61–L64, 1998.

Loudon, R., *The Quantum Theory of Light*, 2nd ed., Oxford Univ. Press, London, 1983.

Lyne, A. G., Orbital Inclination and Mass of the Binary Pulsar PSR0655+64, *Nature*, **310**, 300–302, 1984.

Lyne, A. G. and F. G. Smith, Interstellar Scintillation and Pulsar Velocities, *Nature*, **298**, 825–827, 1982.

Mannucci, A. J., B. D. Wilson, D. N. Yuan, C. H. Ho, U. J. Lindqwister, and T. F. Runge, A Global Mapping Technique for GPS-Derived Ionospheric Total Electron Content Measurements, *Radio Sci.*, **33**, 565–582, 1998.

Marini, J. W., Correction of Satellite Tracking Data for an Arbitrary Tropospheric Profile, *Radio Sci.*, **7**, 223–231, 1972.

Masson, C. R., Atmospheric Effects and Calibrations, in *Astronomy with Millimeter and Submillimeter Wave Interferometry*, M. Ishiguro and W. J. Welch, Eds., Astron. Soc. Pacific Conf. Series **59**, 87–95, 1994a.

Masson, C. R., Seeing, in *Very High Angular Resolution Imaging*, J. G. Robertson and W. J. Tango, Eds., , IAU Symp. 158, Kluwer, Dordrecht, 1994b, pp. 1–10.

Mathur, N. C., M. D. Grossi, and M. R. Pearlman, Atmospheric Effects in Very Long Baseline Interferometry, *Radio Sci.*, **5**, 1253–1261, 1970.

Matsushita, S., H. Matsuo, J. R. Pardo, and S. J. E. Radford, FTS Measurements of Submillimeter–Wave Atmospheric Opacity at Pampa la Bola II: Supra-Terahertz Windows and Model Fitting, *Pub. Ast. Soc. Japan*, **51**, 603–610, 1999.

Mercier, R. P., Diffraction by a Screen Causing Large Random Phase Fluctuations, *Proc. R. Soc. Lond. A*, **58**, 382–400, 1962.

Miner, G. F., D. D. Thornton, and W. J. Welch, The Inference of Atmospheric Temperature Profiles from Ground-Based Measurements of Microwave Emission from Atmospheric Oxygen. *J. Geophys. Res.*, **77**, 975–991, 1972.

Misner, C. W., K. S. Thorne, and J. A. Wheeler, *Gravitation*, Freedman, San Francisco, 1973, Sec. 40.3.

Moran, J. M., The Effects of Propagation on VLBI Observations, in *Very Long Baseline Interferometry: Techniques and Applications*, M. Felli and R. E. Spencer, Eds., Kluwer, Dordrecht, 1989, pp. 47–59.

Moran, J. M. and B. R. Rosen, Estimation of the Propagation Delay through the Troposphere from Microwave Radiometer Data, *Radio Sci.*, **16**, 235–244, 1981.

Muhleman, D. O., R. D. Ekers, and E. B. Fomalont, Radio Interferometric Test of the General Relavistic Light Bending Near the Sun, *Phys. Rev. Lett.*, **24**, L1377–L1380, 1970.

Narayan, R., From Scintillation Observations to a Model of the ISM–The Inverse Problem, in *Radio Wave Scattering in the Interstellar Medium*, J. M. Cordes, B. J. Rickett, and D. C. Backer, Eds., American Institute of Physics Conf. Proc. 174, New York, 1988, pp. 17–31.

Narayan, R., The Physics of Pulsar Scintillation, *Phil. Tran. R. Soc. Lond. A*, **341**, 151–165, 1992.

Narayan, R., K. R. Anantharamaiah, and T. J. Cornwell, Refractive Radio Scintillation in the Solar Wind, *Mon. Not. R. Astron. Soc.*, **241**, 403–413, 1989.

Narayan, R. and J. Goodman, The Shape of a Scatter-Broadened Image: I. Numerical Simulations and Physical Principles, *Mon. Not. R. Astron. Soc.*, **238**, 963–994, 1989.

Niell, A. E., Global Mapping Functions for the Atmospheric Delay at Radio Wavelengths, *J. Geophys Res.*, **101**, 3227–3246, 1996.

NRAO, Recommended Site for the Millimeter Array, National Radio Astronomy Observatory, Charlottesville, VA, May 1998.

Olmi, L. and D. Downes, Interferometric Measurement of Tropospheric Phase Fluctuations at 86 GHz on Antenna Spacings of 24 m to 288 m, *Astron. Astrophys.*, **262**, 634–643, 1992.

Owens, J. C., Optical Refractive Index of Air: Dependence on Pressure, Temperature, and Composition, *Appl. Opt.*, **6**, 51–58, 1967.

Paine, S., R. Blundell, D. C. Papa, J. W. Barrett, and S. J. E. Radford, A Fourier Transform Spectrometer for Measurement of Atmospheric Transmission at Submillimeter Wavelengths, *Pub. Astron. Soc. Pacific*, **112**, 108–118, 2000.

Pardo, J. R., E. Serabyn, and J. Cernicharo, Submillimeter Atmospheric Transmission Measurements on Mauna Kea During Extremely Dry El Nino Conditions, *J. Quant. Spect and Rad Trans.*, **68**, 419–433, 2001.

Pi, X., A. J. Mannucci, U. J. Lindqwister, and C. M. Ho, Monitoring of Global Ionospheric Irregularities Using the Worldwide GPS Network, *Geophys. Res. Lett.*, **24**, 2283–2286, 1997.

Pol, S. L. C., C. S. Ruf, and S. J. Keihm, Improved 20- to 32-GHz Atmospheric Absorption Model, *Radio Sci.*, **33**, 1319–1333, 1998.

Radford, S. J. E. and R. A. Chamberlin, *Atmospheric Transparency at 225 GHz over Chajnantor, Mauna Kea, and the South Pole*, ALMA Memo. 334, National Radio Astronomy Observatory, Socorro, New Mexico, 2000.

Radford, S. J. E., G. Reiland, and B. Shillue, Site Test Interferometer, *Pub. Astron. Soc. Pacific*, **108**, 441–445, 1996.

Ratcliffe, J. A., Some Aspects of Diffraction Theory and Their Application to the Ionosphere, *Rep. Prog. Phys.*, **19**, 188–267, 1956.

Ratcliffe, J. A., *The Magneto-Ionic Theory and Its Application to the Ionosphere*, Cambridge Univ. Press, Cambridge, UK, 1962.

Rawer, K., *The Ionosphere*, Ungar, New York, 1956.

Ray, P. S., Broadband Complex Refractive Indices of Ice and Water, *Applied Optics*, **11**, 1836–1843, 1972.

Readhead, A. C. S. and A. Hewish, Galactic Structure and the Apparent Size of Radio Sources, *Nature*, **236**, 440–443, 1972.

Reber, E. E. and J. R. Swope, On the Correlation of Total Precipitable Water in a Vertical Column and Absolute Humidity, *J. Appl. Meteorol.*, **11**, 1322–1325, 1972.

Resch, G. M., Water Vapor Radiometry in Geodetic Applications, in *Geodetic Aspects of Electromagnetic Wave Propagation through the Atmosphere*, F. K. Brunner, Ed., Springer-Verlag, Berlin, 1984.

Rickett, B. J., W. A. Coles, and G. Bourgois, Slow Scintillation in the Interstellar Medium, *Astron. Astrophys.*, **134**, 390–395, 1984.

Rickett, B. J., Radio Propagation through the Turbulent Interstellar Medium, *Ann. Rev. Astron. Astrophys.*, **28**, 561–605, 1990.

Roberts, D. H., A. E. E. Rogers, B. R. Allen, C. L. Bennet, B. F. Burke, P. E. Greenfield, C. R. Lawerence, and T. A. Clark, Radio Interferometric Detection of a Traveling Ionospheric Disturbance Excited by the Explosion of Mt. St. Helens, *J. Geophys. Res.*, **87**, 6302–6306, 1982.

Robertson, D. S., W. E. Carter, and W. H. Dillinger, New Measurement of Solar Gravitational Deflection of Radio signals using VLBI, *Nature*, **349**, 768–770, 1991.

Roddier, F., The Effects of Atmospheric Turbulence in Optical Astronomy, in *Progress in Optics XIX*, E. Wolf, Ed., North-Holland, Amsterdam, 1981, pp. 281–376.

Rogers, A. E. E., A. T. Moffet, D. C. Backer, and J. M. Moran, Coherence Limits in VLBI Observations at 3-Millimeter Wavelength, *Radio Sci.*, **19**, 1552–1560, 1984.

Rogers, A. E. E. and J. M. Moran, Coherence Limits for Very Long Baseline Interferometry, *IEEE Trans. Instrum. Meas.*, **IM-30**, 283–286, 1981.

Rosenkranz, P. W., Water Vapor Microwave Continuum Absorption: A Comparison of Measurements and Models, *Radio Sci.*, **33**, 919–928, 1998.

Rybicki, G. B. and A. P. Lightman, *Radiative Processes in Astrophysics*, Wiley-Interscience, New York, 1979 (reprinted 1985).

Saastamoinen, J., Introduction to Practical Computation of Astronomical Refraction, *Bull. Geodesique*, **106**, 383–397, 1972a.

Saastamoinen, J., Atmospheric Correction for the Troposphere and Stratosphere in Radio Ranging of Satellites, in *The Use of Artificial Satellites for Geodesy*, Geophysical Monograph 15, American Geophysical Union, Washington, DC, 1972b, pp. 247–251.

Salpeter, E. E., Interplanetary Scintillations. I. Theory, *Astrophys. J.* **147**, 433–448, 1967.

Schaper, L. W., Jr., D. H. Staelin, and J. W. Waters, The Estimation of Tropospheric Electrical Path Length by Microwave Radiometry, *Proc. IEEE*, **58**, 272–273, 1970.

Scheuer, P. A. G., Amplitude Variations in Pulsed Radio Sources, *Nature*, **218**, 920–922, 1968.

Scott, S. L., W. A. Coles, and G. Bourgois, Solar Wind Observations Near the Sun Using Interplanetary Scintillation, *Astron. Astrophys.*, **123**, 207–215, 1983.

Shapiro, I. I., New Method for the Detection of Light Deflection by Solar Gravity, *Science*, **157**, 806–808, 1967.

Shapiro, I. I., Estimation of Astrometric and Geodetic Parameters, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 261–276.

Sieber, W., Causal Relationship Between Pulsar Long-Term Intensity Variations and the Interstellar Medium, *Astron. Astrophys.*, **113**, 311–313, 1982.

Simard–Normandin, M. and P. P. Kronberg, Rotation Measures and the Galactic Magnetic Field, *Astrophys. J.*, **242**, 74–94, 1980.

Smart, W. M., *Textbook on Spherical Astronomy*, 6th ed., rev. R. M. Green, Cambridge Univ. Press, Cambridge, UK, 1977.

Smith, E. K., Jr. and S. Weintraub, The Constants in the Equation for Atmospheric Refractive Index at Radio Frequencies, *Proc. IRE*, **41**, 1035–1037, 1953.

Smith, F. G., C. G. Little, and A. C. B. Lovell, Origin of the Fluctuations in the Intensity of Radio Waves from Galactic Sources, *Nature*, **165**, 422–424, 1950.

Snider, J. B., Ground-Based Sensing of Temperature Profiles from Angular and Multi-Spectral Microwave Emission Measurements, *J. Appl. Meteorol.*, **11**, 958–967, 1972.

Snider, J. B., H. M. Burdick, and D. C. Hogg, Cloud Liquid Measurement with a Ground-Based Microwave Instrument, *Radio Sci.*, **15**, 683–693, 1980.

Solheim, F., Godwin, J. R., Westwater, E. R., Han, Yong, Keihm, S. J., Marsh, K., and Ware, R., Radiometric Profiling of Temperature,, Water Vapor, and Cloud Liquid Water Using Various Inversion Methods, *Radio Sci.*, **33**, 393–404, 1998.

Spangler, S. R. and C. R. Gwinn, Evidence for an Inner Scale to the Density Turbulence in the Interstellar Medium, *Astrophys. J.*, **353**, L29–L32, 1990.

Spangler, S. R. and T. Sakurai, Radio Interferometry of Solar Wind Turbulence from the Orbit of Helios to the Solar Corona, *Astrophys. J.*, **445**, 999–1061, 1995.

Spitzer, L., *Physical Processes in the Interstellar Medium*, Wiley-Interscience, New York, 1978, p. 65.

Spoelstra, T. A. T., The Influence of Ionospheric Refraction on Radio Astronomy Interferometry, *Astron. Astrophys.*, **120**, 313–321, 1983.

Spoelstra, T. A. T. and H. Kelder, Effects Produced by the Ionosphere on Radio Interferometry, *Radio Sci.*, **19**, 779–788, 1984.

Sramek, R., VLA Phase Stability at 22 GHz on Baselines of 100 m to 3 km, VLA Test Memo. 143, National Radio Astronomy Observatory, Sororro, NM, 1983.

Sramek, R. A., Atmospheric Phase Stability at the VLA, in *Radio Astronomical Seeing*, J. E. Baldwin and Wang Shouguan, Eds., International Academic Publishers, Pergamon Press, Oxford, 1990, pp. 21–30.

Staelin, D. H., Measurements and Interpretation of the Microwave Spectrum of the Terrestrial Atmosphere near 1-Centimeter Wavelength, *J. Geophys. Res.*, **71**, 2875–2881, 1966.

Sutton, E. C., and R. M. Hueckstaedt, Radiometric Monitoring of Atmospheric Water Vapor as It Pertains to Phase Correction in Millimeter Interferometry, *Astron. Astrophysics Suppl.*, **119**, 559–567, 1996.

Sutton, E. C., S. Subramanian, and C. H. Townes, Interferometric Measurements of Stellar Positions in the Infrared, *Astron. Astrophys.* **110**, 324–331, 1982.

Tahmoush, D. A. and A. E. E. Rogers, Correcting Atmospheric Variations in Millimeter Wavelength Very Long Baseline Interferometry Using a Scanning Water Vapor Radiometer, *Radio Sciences*, **35**, 1241–1251, 2000.

Tatarski, V. I., *Wave Propagation in a Turbulent Medium*, Dover, New York, 1961.

Tatarski, V. I., *The Effects of the Turbulent Atmosphere on Wave Propagation*, National Technical Information Service, Springfield, VA, 1971.

Taylor, G. I., Spectrum of Turbulence, *Proc. Roy. Soc.*, **164A**, 476–490, 1938.

Taylor, J. H. and J. M. Cordes, Pulsar Distances and the Galactic Distribution of Free Electrons, *Astrophys. J.*, **411**, 674–684, 1993.

Thayer, G. D., An Improved Equation for the Radio Refractive Index of Air, *Radio Sci.*, **9**, 803–807, 1974.

Treuhaft, R. N. and G. E. Lanyi, The Effect of the Dynamic Wet Troposphere on Radio Interferometric Measurements, *Radio Sci.*, **22**, 251–265, 1987.

Van Vleck, J. H., E. M. Purcell, and H. Goldstein, Atmospheric Attenuation, in *Propagation of Short Radio Waves*, MIT Radiation Laboratory Series, Vol. 13, D. E. Kerr, Ed., McGraw-Hill, New York, 1951, pp. 641–692.

Waters, J. W., Absorption and Emission by Atmospheric Gases, in *Methods of Experimental Physics*, Vol. 12B, M. L. Meeks, Ed., Academic Press, New York, 1976, pp. 142–176.

Weinberg, S., *Gravitation and Cosmology: Principles and Applications of the General Theory of Relativity*, Wiley, New York, 1972, p. 188.

Weisberg, J. M., J. Rankin, and V. Boriakoff, HI Absorption Measurements of Seven Low Latitude Pulsars, *Astron. Astrophys.*, **88**, 84–93, 1980.

Welch, W. J., Correcting Atmospheric Phase Fluctuations by Means of Water-Vapor Radiometry, in *Review of Radio Science, 1996–1999*, W. R. Stone, Ed., Oxford Univ. Press, Oxford, 1999, pp. 787–808.

Westwater, E. R., *An Analysis of the Correction of Range Errors due to Atmospheric Refraction by Microwave Radiometric Techniques*, ESSA Technical Report, IER 30-ITSA 30, Institute for Telecommunication Sciences and Aeronomy, Boulder, CO, 1967.

Westwater, E. R., The Accuracy of Water Vapor and Cloud Liquid Determination by Dual-Frequency Ground-Based Microwave Radiometry, *Radio Sci.*, **13**, 677–685, 1978.

Westwater, E. R. and F. O. Guiraud, Ground-Based Microwave Radiometric Retrieval of Precipitable Water Vapor in the Presence of Clouds with High Liquid Content, *Radio Sci.*, **15**, 947–957, 1980.

Wiedner, M. C. and R. E. Hills, Phase Correction on Mauna Kea Using 183 GHz Water Vapor Monitors, in *Imaging at Radio through Submillimeter Wavelengths*, J. G. Mangum and S. J. E. Radford, Eds., Astron. Soc. Pacific Conf. Ser., **217**, 327–335, 2000.

Winterhalter, D., J. T. Gosling, S. R. Habbal, W. S. Kurth, and M. Neugebauer, Solar Wind Eight, Proc. 8th Int. Solar Wind Conf., *AIP Conference Proceedings*, Vol. 382, American Institute of Physics, New York, 1996.

Wolszczan, A. and J. M. Cordes, Interstellar Interferometry of the Pulsar PSR 1237+25, *Astrophys. J.*, **320**, L35–L39, 1987.

Woody, D., J. Carpenter, and N. Scoville, Phase Correction at OVRO Using 22 GHz Water Line, in *Imaging at Radio through Submillimeter Wavelengths*, J. G. Mangum and S. J. E. Radford, Eds., Astron. Soc. Pacific Conf. Ser., **217**, 317–326, 2000.

Woolf, N. J., High Resolution Imaging from the Ground, *Ann. Rev. Astron. Astrophys.* **20**, 367–398, 1982.

Wright, M. C. H., Atmospheric Phase Noise and Aperture-Synthesis Imaging at Millimeter Wavelengths, *Pub. Astron. Soc. Pacific.*, **108**, 520–534, 1996.

Wright, M. C. H. and Welch, W. J., Interferometer Measurements of Atmospheric Phase Noise at 3 mm, in *Radio Astronomical Seeing*, J. E. Baldwin and Wang Shouguan, Eds., International Academic Publishers, Pergamon Press, Oxford, 1990, pp. 71–74.

Wu, S. C., Optimum Frequencies of a Passive Microwave Radiometer for Tropospheric Path-Length Correction, *IEEE Trans. Antennas Propag.*, **AP-27**, 233–239, 1979.

Yeh, K. C. and C. H. Liu, Radio Wave Scintillations in the Ionosphere, *Proc. IEEE*, **70**, 324–360, 1982.

Young, A. T., Interpretation of Interplanetary Scintillations, *Astrophys. J.*, **168**, 543–562, 1971.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

14 Van Cittert–Zernike Theorem, Spatial Coherence, and Scattering

This chapter is concerned with the van Cittert–Zernike theorem, including an examination of the assumptions involved in its derivation, the requirement of spatial incoherence of a source, and the interferometer response to a coherent source. There is also a brief discussion of some aspects of scattering by irregularities in the propagation medium. Much of the development of the theory of coherence and similar concepts of electromagnetic radiation is to be found in the literature of optics. The terminology is sometimes different from that which has evolved in radio interferometry, but many of the physical situations are similar or identical. In some of the analyses we use optical terminology and introduce the concept of *mutual coherence*, which includes complex visibility.

14.1 VAN CITTERT–ZERNIKE THEOREM

We showed in Chapters 2 and 3 that the cross-correlation of the signals received in spaced antennas can be used to map the intensity distribution of a distant cosmic source through a Fourier transform relationship. This result is a form of the van Cittert–Zernike theorem, which originated in optics. The basis for the theorem is a study published by van Cittert in 1934, and followed a few years later by a simpler derivation by Zernike. A description of the result established by van Cittert and Zernike is given by Born and Wolf (1999, Ch. 10). The original form of the result does not specifically refer to the Fourier transform relationship between intensity and mutual coherence, but is essentially as follows.

Consider an extended, quasi-monochromatic, incoherent source, and let the mutual coherence of the radiation be measured at two points P_1 and P_2 in a plane normal to the direction of the source, as in Fig. 14.1. Then suppose that the source is replaced by an aperture of identical shape and size, and illuminated from behind by a spatially coherent wavefront. The distribution of the electric field amplitude over the aperture is proportional to the intensity distribution over the source. The Fraunhofer diffraction pattern of the aperture is observable in the plane containing P_1 and P_2. The relative positions of the points P_1 and P_2 are the same in the two

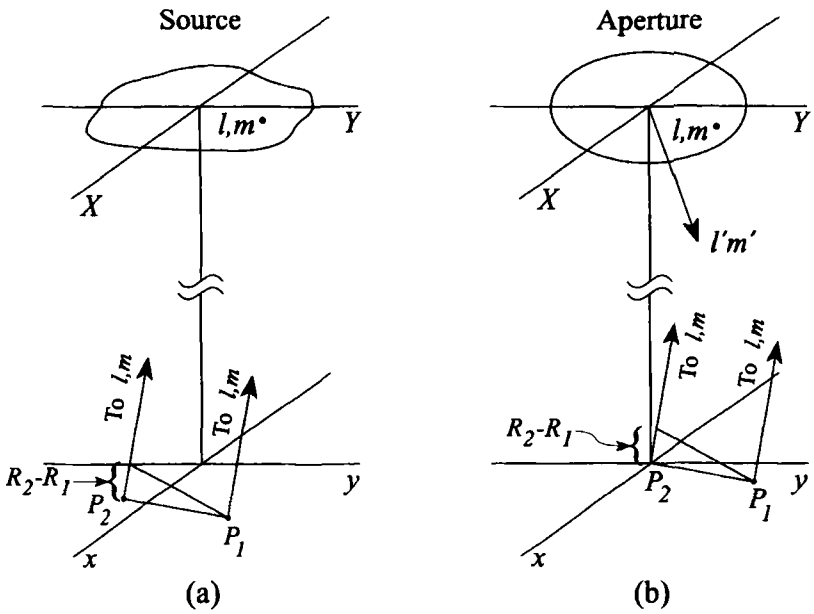


Figure 14.1 (a) Geometry of a distant spatially incoherent source and the points P_1 and P_2 at which the mutual coherence of the radiation is measured. The source plane (X, Y) is parallel to the measurement plane (x, y) but at a large distance from it. (b) Similar geometry for measurement of the radiation field from an aperture in the (X, Y) plane that is illuminated from above by a coherent wavefront. The radiated field has a maximum at the point P_2. Direction cosines (l, m) are defined with respect to the (x, y) axes in the measurement plane, and (l', m') with respect to the (X, Y) axes in the plane of the aperture.

cases, but for the aperture the geometric configuration is such that P_2 lies on the maximum of the diffraction pattern. Then the mutual coherence measured for the incoherent source, normalized to unity for zero spacing between P_1 and P_2, is equal to the complex amplitude of the field of the aperture diffraction pattern at the position P_1, normalized to the maximum value at P_2.

In this form the theorem results from the fact that the behavior of both the mutual coherence and the Fraunhofer diffraction can be represented by similar Fourier transform relationships. Derivation of the theorem provides an opportunity to examine the assumptions involved, and is given below. The analysis is similar to that given by Born and Wolf, but with some modifications to take advantage of the simplified geometry when the source is at an astronomical distance. First we note that in optics the *mutual coherence function* for a field $E(t)$, measured at points 1 and 2, is represented by

$$\Gamma_{12}(u, v, \tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} E_1(t) E_2^*(t-\tau)\, dt, \tag{14.1}$$

where u and v are the coordinates of the spacing between the two measurement points, expressed in units of wavelength. $\Gamma_{12}(u, v, 0)$, for zero time offset, is equivalent to the complex visibility $\mathcal{V}(u, v)$ used in the radio case.

Mutual Coherence of an Incoherent Source

The geometric situation for the incoherent source is shown in Fig. 14.1a. Consider the source located in a distant plane, indicated by (X, Y). The radiated field is measured at two points, P_1 and P_2, in the (x, y) plane that is parallel to the source plane. In the radio case these points are the locations of the interferometer antennas. It is convenient to specify the position of a point in the (X, Y) plane by the direction cosines (l, m) measured with respect to the (x, y) axes. The source is sufficiently distant that the direction of any point within it measured from P_1 is the same as that measured from P_2. The fields at P_1 and P_2 resulting from a single element of the source at the point (l, m) are given by

$$E_1(l, m, t) = \mathcal{E}\left(l, m, t - \frac{R_1}{c}\right) \frac{\exp[-j2\pi\nu(t - R_1/c)]}{R_1}, \tag{14.2}$$

and

$$E_2(l, m, t) = \mathcal{E}\left(l, m, t - \frac{R_2}{c}\right) \frac{\exp[-j2\pi\nu(t - R_2/c)]}{R_2}, \tag{14.3}$$

where $\mathcal{E}(l, m, t)$ is a phasor representation of the complex amplitude of the electric field at the source for an element at position (l, m). R_1 and R_2 are the distances from this element to P_1 and P_2, respectively, and c is the velocity of light. The exponential terms in Eqs. (14.2) and (14.3) represent the phase change in traversing the paths from the source to P_1 and P_2.

The complex cross-correlation of the field voltages at P_1 and P_2 due to the radiation from the element at (l, m) is, for zero time offset,

$$\begin{aligned}
&\langle E_1(l, m, t) E_2^*(l, m, t)\rangle \\
&\quad = \left\langle \mathcal{E}\left(l, m, t - \tfrac{R_1}{c}\right) \mathcal{E}^*\left(l, m, t - \tfrac{R_2}{c}\right)\right\rangle \\
&\qquad \times \frac{\exp[-j2\pi\nu(t - R_1/c)]\exp[j2\pi\nu(t - R_2/c)]}{R_1 R_2} \\
&\quad = \left\langle \mathcal{E}(l, m, t)\mathcal{E}^*\left(l, m, t - \frac{R_2 - R_1}{c}\right)\right\rangle \frac{\exp[j2\pi\nu(R_1 - R_2)/c]}{R_1 R_2},
\end{aligned} \tag{14.4}$$

where the asterisk denotes the complex conjugate, and the angle brackets $\langle\ \rangle$ represent a time average. Note that the source is assumed to be spatially incoherent, which means that terms of the form $\langle E_1(l_p, m_p, t) E_2^*(l_q, m_q, t)\rangle$, where p and q denote different elements of the source, are zero. If the quantity $(R_2 - R_1)/c$ is small compared with the reciprocal receiver bandwidth, we can neglect it within

the angle brackets of Eq. (14.4), where it occurs in the amplitude term for $\mathcal{E}$. Equation (14.4) then becomes

$$\left\langle E_1(l,m,t)E_2^*(l,m,t)\right\rangle = \frac{\langle \mathcal{E}(l,m,t)\mathcal{E}^*(l,m,t)\rangle \exp[j2\pi\nu(R_1-R_2)/c]}{R_1R_2}. \tag{14.5}$$

The quantity $\langle \mathcal{E}(l,m,t)\mathcal{E}^*(l,m,t)\rangle$ is a measure of the time-averaged intensity $I(l,m)$ of the source. To obtain the mutual coherence function of the fields at points P_1 and P_2, we integrate over the source, using ds to represent an element of area within the (X,Y) plane:

$$\Gamma_{12}(u,v,0) = \int_{\text{source}} \frac{I(l,m)\exp[j2\pi\nu(R_1-R_2)/c]}{R_1R_2}\,ds, \tag{14.6}$$

where u and v are the x and y components of the spacing between the points P_1 and P_2 measured in wavelengths. Note that (R_1-R_2) is the differential distance in the path lengths from (l,m) in the source to P_1 and P_2. The points P_1 and P_2 have coordinates (x_1,y_1) and (x_2,y_2) respectively, so $u=(x_1-x_2)\nu/c$ and $v=(y_1-y_2)\nu/c$, where c/ν is the wavelength. Thus we obtain $(R_2-R_1)=(ul+vm)c/\nu$. Because the distance of the source is very much greater than the distance between P_1 and P_2, for the remaining R terms we can put $R_1\simeq R_2\simeq R$, where R is the distance between the (X,Y) and (x,y) origins. Then $ds=R^2\,dl\,dm$, and from Eq. (14.6)

$$\Gamma_{12}(u,v,0) = \int\int_{\text{source}} I(l,m)e^{-j2\pi(ul+vm)}\,dl\,dm. \tag{14.7}$$

Since the integrand in Eq. (14.7) is zero outside the source boundary, the limits of the integral effectively extend to infinity and the mutual coherence $\Gamma_{12}(u,v,0)$, which is equivalent to the complex visibility $\mathcal{V}(u,v)$, is the Fourier transform of the intensity distribution $I(l,m)$ of the source. This result is generally referred to as the van Cittert–Zernike theorem. However, it is instructive to examine the definition of the theorem in terms of the diffraction pattern of an aperture given at the beginning of this section.

Diffraction at an Aperture and the Response of an Antenna

The Fraunhofer diffraction field of an aperture, as a function of angle, can be analyzed using the geometry shown in Fig. 14.1b. Here, an aperture is illuminated by an electromagnetic field of amplitude $\mathcal{E}(l,m,t)$, where again we use direction cosines with respect to the x and y axes to indicate points within the aperture as seen from P_1 and P_2. The (x,y) plane is in the far field of a wavefront from any point in the aperture, so such a wavefront can be considered plane over the distance P_1P_2. The aperture is centered on the point O and is normal to the direction OP_2. The phase over the aperture is assumed to be uniform, and components

of the field therefore combine in phase at P_2. Thus in the (x, y) plane the maximum field strength occurs at P_2. Now consider the field at the point P_1 which has coordinates (x, y). The component of the field at P_1 due to radiation from an element of the aperture at position (l, m) is given by Eq. (14.2). The path lengths from the point (l, m) at the source to P_1 and P_2 are R_1 and R_2, respectively, and $R_2 - R_1 = lx + my$. Thus from Eq. (14.2) we can write

$$E_1(l, m, t) = \frac{e^{-j2\pi\nu(t-R_2/c)}}{R_1}\mathcal{E}\left(l, m, t - \frac{R_1}{c}\right)e^{-j2\pi\nu(xl+ym)/c}. \tag{14.8}$$

Again, for the remaining R terms we put $R_1 \simeq R_2 \simeq R$. Integration over the aperture then gives the total field at P_1,

$$E(x, y) = \frac{e^{-j2\pi\nu(t-R/c)}}{R}\int_{\text{aperture}} \mathcal{E}\left(l, m, t - \frac{R}{c}\right)e^{-j2\pi[(x/\lambda)l+(y/\lambda)m]}\,ds, \tag{14.9}$$

where λ is the wavelength and the element of area ds is proportional to $dl\,dm$. The term on the right-hand side that is outside the integrals is a propagation factor that represents the variation in amplitude and phase over the path from the source to P_2 in Fig. 14.1b. In applying the result to the radiation pattern of an aperture, we replace the time-dependent functions E and $\mathcal{E}$ by the corresponding rms field amplitudes, which will be denoted by $\overline{E}$ and $\overline{\mathcal{E}}$, respectively:

$$\overline{E}(x, y) \propto \iint_{\text{aperture}} \overline{\mathcal{E}}(l, m)e^{-j2\pi[(x/\lambda)l+(y/\lambda)m]}\,dl\,dm, \tag{14.10}$$

where the propagation factor in Eq. (14.9) has been omitted. A comparison of Eqs. (14.7) and (14.10) explains the van Cittert–Zernike theorem as described at the beginning of this section. With the specified proportionality between the incoherent intensity and the coherent field amplitude, it will be found that

$$\frac{\Gamma_{12}(u, v, 0)}{\Gamma_{12}(0, 0, 0)} = \frac{\overline{E}(x, y)}{\overline{E}(0, 0)}. \tag{14.11}$$

In Eqs. (14.7) and (14.10) the integrand is zero outside the source or aperture. Thus, in each case, the limits of integration can be extended to $\pm\infty$, and the equations are seen to be Fourier transforms. The calculations of the mutual coherence of the source and the radiation pattern of the aperture yield similar results because the geometry and the mathematical approximations are the same in each case. It should be emphasized, however, that the physical situations are different. In the first case considered the source is spatially incoherent over its surface, whereas in the second case the field across the aperture is fully coherent.

The result in Eq. (14.10) also gives the angular radiation pattern for an antenna that has the form of an excited aperture. The application to an antenna is more

useful if the radiation pattern is specified in terms of an angular representation (l', m') of the direction of radiation from the antenna aperture instead of the position of the point P_1, and if the field distribution over the aperture is specified in terms of units of length rather than angle. (l', m') are direction cosines with respect to the (X, Y) axes. Since the angles concerned are small, we can substitute into Eq. (14.10) $x = Rl'$, $y = Rm'$, $l = X/R$, $m = Y/R$, $dl = dX/R$, and $dm = dY/R$, and obtain

$$\overline{E}'(l', m') \propto \iint_{\text{aperture}} \overline{\mathcal{E}}_{XY}(X, Y) e^{-j2\pi[(X/\lambda)l' + (Y/\lambda)m']} \, dX \, dY. \tag{14.12}$$

This is the expression for the field distribution resulting from Fraunhofer diffraction at an aperture [see, e.g., Silver (1949)]. It includes the case of a transmitting antenna in which the aperture of a parabolic reflector is illuminated by a radiator at the focus. If such an antenna is used in reception, the received voltage from a source in direction (l', m') is proportional to the right-hand side of Eq. (14.12). Thus the voltage reception pattern $V_A(l', m')$, introduced in Section 3.3 under *Antennas*, is proportional to the right-hand side of Eq. (14.12).

To obtain the power radiation pattern for an antenna, we need the response in terms of $|\overline{E}'(l', m')|^2$. From an autocorrelation theorem of Fourier transforms the squared amplitude of $\overline{E}'(l', m')$ is equal to the autocorrelation of the Fourier transform of $\overline{E}'(l', m')$ [see, e.g., Bracewell (2000), and note that this relationship is also a generalization of the Wiener–Khinchin relationship derived in Section 3.2]. Thus the power radiated as a function of angle is given by

$$\begin{aligned} &|\overline{E}'(l', m')|^2 \\ &\quad \propto \iint_{\text{aperture}} [\overline{\mathcal{E}}_{XY}(X, Y) \star\star \overline{\mathcal{E}}_{XY}(X, Y)] e^{-j2\pi[(X/\lambda)l' + (Y/\lambda)m']} \, dX \, dY, \end{aligned} \tag{14.13}$$

where $\overline{\mathcal{E}}(X, Y) \star \bullet \overline{\mathcal{E}}(X, Y)$ is the two-dimensional autocorrelation function of the field distribution over the aperture. To obtain absolute values of the radiated field, the required constant of proportionality can be determined by integrating Eq. (14.13) over 4π sr to obtain the total radiated power, and equating this to the power applied to the antenna terminals. In reception, the power collected by an antenna is proportional to the power radiated in transmission, so the form of the beam is identical in the two cases. To illustrate the physical interpretation of Eq. (14.13), consider the simple case of a rectangular aperture with uniform excitation of the electric field. The function $\overline{\mathcal{E}}_{XY}(X, Y)$ is then the product of two one-dimensional functions of X and Y. If d is the aperture width in the X direction, the autocorrelation function in X is triangular with a width $2d$, and Fourier transformation gives

$$|\overline{E}_X(l')|^2 \propto \left[\frac{\sin(\pi d l'/\lambda)}{\pi d l'/\lambda} \right]^2. \tag{14.14}$$

In the l' dimension the full width of this beam at the half-power level is $0.886\lambda/d$, for example, 1° for $d/\lambda = 50.8$ wavelengths. For a uniformly illuminated circular

aperture of diameter d, the response pattern is circularly symmetrical and is given by

$$|\overline{E}_r(l'_r)|^2 \propto \left[\frac{2J_1(\pi d l'_r/\lambda)}{\pi d l'_r/\lambda}\right]^2, \tag{14.15}$$

where the subscript r indicates a radial profile in which l'_r is measured from the center of the beam. The full width of the beam at the half-power level is $1.03\lambda/d$.

A more direct way of obtaining the Fraunhofer radiation pattern of an aperture antenna is to start by considering the field strength of the radiated wavefront as a function of direction, rather than the field strength at a single point P_1, as above. However, the method used was chosen to provide a more direct comparison with the interferometer response to a spatially incoherent source. For a more detailed analysis of the response of an antenna, see, for example, Booker and Clemmow (1950), Bracewell (1962), or the textbooks on antennas in the bibliography of Chapter 5.

Assumptions in the Derivation and Application of the Van Cittert–Zernike Theorem

At this point it is convenient to collect and review the assumptions and limitations that are involved in the theory of the interferometer response.

1. *Polarization of the electric field.* Although the electric fields are vector quantities with directions that depend on the polarization of the radiation, the components received by antennas from different elements of the source can be combined in the manner of scalar quantities. The fields are measured by antennas at P_1 and P_2, and each antenna responds to the component of the radiation for which the polarization matches that of the antenna. If the fields are randomly polarized and the antennas are identically polarized, then the signal product in Eq. (14.4) represents half the total power at each antenna. However, the antenna polarizations do not have to be identical since, in general, the interferometer system will respond to some combination of components of the source intensity determined by the antenna polarizations. The ways in which the antenna polarizations can be chosen to examine all polarizations of the incident radiation are described in Section 4.8 under *Stokes Visibilities*. Thus the scalar treatment of the field involves no loss of generality.

2. *Spatial incoherence of the source.* The radiation from any point on the source is statistically independent from that from any other point. This applies almost universally to astronomical sources, and permits the integration in Eq. (14.6) by allowing cross products representing different elements of the source to be omitted. The Fourier transform relationship provided by the van Cittert–Zernike theorem requires the source to be spatially incoherent. Spatial coherence and incoherence are discussed in Section 14.2. Note that an incoherent source gives rise to a coherent or partially coherent wavefront as its radiation propagates through space. If this were not the case the mutual coherence (or visibility) of an incoherent source, measured by spaced antennas, would always be zero.

3. *Bandwidth pattern.* The assumption required in going from Eqs. (14.4) to (14.5), that $(R_2 - R_1)/c$ is less than the reciprocal bandwidth $(\Delta\nu)^{-1}$, can be written

$$\frac{\Delta\nu}{\nu} < \frac{1}{l_d u}; \qquad \frac{\Delta\nu}{\nu} < \frac{1}{m_d v}, \tag{14.16}$$

where l_d and m_d are the maximum angular dimensions of the source. This is the requirement that the source be within the limits imposed by the bandwidth pattern of the interferometer, which is discussed in Section 2.2. Conversely, the required field of view limits the maximum bandwidth that can be used. The distortion caused by the bandwidth effect is discussed further in Section 6.3 and, if not severe, can often be corrected.

4. *Distance of the source.* For an array with maximum baseline D the departure of the wavefront from a plane for a source of distance R is $\sim D^2/R$. Thus the *far-field* distance R_{ff}, defined as that for which the divergence is small compared with the wavelength λ, is given by

$$R_{ff} \gg D^2/\lambda\,. \tag{14.17}$$

The far-field condition implies that the antenna spacing subtends a small angle as seen from the source and results in the approximation for Fraunhofer diffraction. If the source is at a known distance closer than the far-field distance, then the phase term can be compensated. This may sometimes be necessary in solar system studies. For example, for an antenna spacing of 35 km and a wavelength of 1 cm, the far-field distance is greater than 1.2×10^{11} m, or approximately the distance of the sun. Note that the requirement of the far-field distance means that no information concerning the structure of the object in the longitudinal direction is possible, only the intensity distribution as projected onto the celestial sphere.

5. *Use of direction cosines.* In going from Eqs. (14.6) to (14.7), the path difference $R_2 - R_1$ is specified in terms of the baseline coordinates (u, v) and angular coordinates (l, m). The expression for the path difference is precise if l and m are specified as direction cosines. In integration over the source, the element of area bounded by increments $dl\,dm$ is equal to $dl\,dm/n$, where n is the third direction cosine and is equal to $\sqrt{1 - l^2 - m^2}$. In optics, derivation of the van Cittert–Zernike theorem usually involves the assumption that the source subtends only small angles at the measurement plane. Then l and m can be approximated by the corresponding small angles and n can be approximated by unity. As a result, the relationship between $\mathcal{V}$ and I becomes a two-dimensional Fourier transform, as in the approximation for limited field size discussed in Section 3.1. In the radio case the less restrictive result in Eq. (3.7) is sometimes required; see Sections 3.1 and 11.8.

6. *Three-dimensional distribution of the visibility measurements.* As antennas track a source, the antenna-spacing vectors, designated above by (u, v) components, may not lie in a plane, and three coordinates, (u, v, w), are then required to specify them. The Fourier transform relationship is then more complicated, but a

simplifying approximation can be made if the field of view to be mapped is small. These effects are discussed in Sections 3.1 and 11.8.

7. *Refraction in space.* It has been implicitly assumed in the analysis above that the space between the source and the antennas is empty, or at least that any medium within it has a uniform refractive index, so that there is no distortion of the incoming wavefront from the source. In practice, the interstellar and interplanetary media, and the earth's atmosphere and ionosphere, can introduce effects including rotation of the position angle of a linearly polarized component as discussed in Section 13.3; see Eq. (13.135).

14.2 SPATIAL COHERENCE

In the derivation of the interferometer response in Chapters 2 and 3, and in Eq. (14.5), it is assumed that the source under discussion is spatially incoherent. This means that the waveforms received from different spatial elements of the source are not correlated, which enables us to add the correlator output from the different angular increments in the integration over the source. We now examine this requirement in more detail. To illustrate the principles involved, it is sufficient to work in one dimension on the sky, for which the position is given by the direction cosine l.

Incident Field

Consider the electric field $E(l, t)$ at the earth's surface resulting from a wavefront incident from the direction l at time t. Figure 14.2 shows the geometry of the situation, in which $l = 0$ in the direction OS of the center, or nominal position, of the source under observation. l is a direction cosine measured from OB, the normal

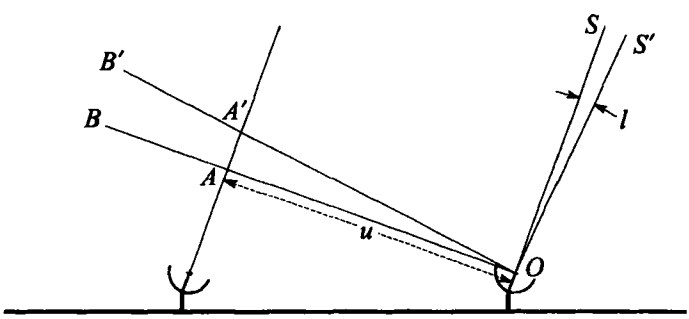


Figure 14.2 Diagram to illustrate the variation of phase along a line OB that is perpendicular to the direction of a source OS, where l is the direction cosine that specifies the direction OS', and is defined with respect to OB. The angle SOS' is small and is thus approximately equal to l, as indicated. The line OS' points toward another part of the same source, and OB' is perpendicular to it.

to OS. A path OS' is shown that indicates the direction of another part of the source. Radiation from the direction OS' produces a wavefront parallel to OB'. The wavefronts from points on the source are plane because we are considering a source in the far field of the interferometer. The line OA represents the projection of the baseline normal to the direction of the source, and the distance OA measured in wavelengths is equal to u. Now consider wavefronts from the directions S and S' that arrive at the same time at O. To reach the point A, the wavefront from S' has to travel a further distance AA'. With the usual small-angle approximation, we find that the distance AA' is equal to ulc/ν, that is, ul wavelengths. Thus the wave from direction S' is delayed at A by a time interval $\tau = ul/\nu$, relative to the wave from S. If we represent the wave from direction S' by $E(l, t)$ at O, at A it is $E(l, t - \tau)$. Now because the incident wavefronts are plane, the amplitude of the wave does not change over the distance AA'. However, the phase changes by $\nu\tau = ul$, so for the wave from S' at A we have

$$E(l, t - \tau) = E(l, t)e^{-j2\pi ul}. \tag{14.18}$$

If $e(u, t)$ is the field at A resulting from radiation from all parts of the source, then

$$e(u, t) = \int_{-\infty}^{\infty} E(l, t)e^{-j2\pi ul}\, dl. \tag{14.19}$$

It will be assumed that the angular dimensions of the source are not large, so also we have

$$E(l, t) = 0, \qquad |l| \geq 1. \tag{14.20}$$

The condition specified in Eq. (14.20) allows us to write the limits of the integral in Eq. (14.19) as $\pm\infty$. Note that Eq. (14.19) has the form of a Fourier transform, and the inverse transform gives $E(l, t)$ from $e(u, t)$. Equation (14.19) will be required in the following subsection.

Source Coherence

We now return to the spatial coherence of the source and follow part of a more extensive analysis by Swenson and Mathur (1968). As a measure of the spatial coherence we introduce the *source coherence function* γ. This is defined in terms of the cross-correlation of signals received from two different directions, l_1 and l_2, at two different times:

$$\begin{aligned}\gamma(l_1, l_2, \tau) &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} E(l_1, t)E^*(l_2, t - \tau)\, dt \\ &= \langle E(l_1, t)E^*(l_2, t - \tau)\rangle.\end{aligned} \tag{14.21}$$

Finite limits are used in the integral to ensure convergence. $\gamma(l_1, l_2, \tau)$ is similar to the coherence function of a source or object discussed by Drane and Parrent (1962) and Beran and Parrent (1964).

The *complex degree of coherence* of an extended source is the normalized source coherence function

$$\gamma_N(l_1, l_2, \tau) = \frac{\gamma(l_1, l_2, \tau)}{\sqrt{\gamma(l_1, 0)\gamma(l_2, 0)}}, \tag{14.22}$$

where $\gamma(l_1, \tau)$ is defined by putting $l_1 = l_2$ in Eq. (14.21), that is, $\gamma(l_1, \tau) = \gamma(l_1, l_1, \tau)$. It can be shown by using the Schwarz inequality that $0 \leq |\gamma_N(l_1, l_2, \tau)| \leq 1$. The extreme values of 0 and 1 correspond to the cases of complete incoherence and complete coherence, respectively. When dealing with extended sources of arbitrary spectral width, it is possible that, for a given pair of points, l_1 and l_2, $|\gamma_N(l_1, l_2, \tau)|$ is zero for one value of τ and nonzero for another value. Therefore, more stringent definitions of complete coherence and incoherence are necessary. The following definitions are adapted from Parrent (1959):

1. The emissions from the directions l_1 and l_2 are completely coherent (incoherent) if $|\gamma_N(l_1, l_2, \tau)| = 1(0)$ for all values of τ.
2. An extended source is coherent (incoherent) if the emissions from all pairs of directions l_1, l_2 within the source are coherent (incoherent).

In all other cases the extended source is described as partially coherent.

Consider now the coherence function of the field $e(x_\lambda, t)$ of a distant source measured, say, at the earth's surface, x_λ being a linear coordinate measured in wavelengths in a direction normal to $l = 0$:

$$\begin{aligned}\Gamma(x_{\lambda 1}, x_{\lambda 2}, \tau) &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e(x_{\lambda 1}, t) e^*(x_{\lambda 2}, t - \tau)\, dt \\ &= \langle e(x_{\lambda 1}, t) e^*(x_{\lambda 2}, t - \tau)\rangle.\end{aligned} \tag{14.23}$$

This is a variation of the mutual coherence function Γ_{12} in Eq. (14.1), in which the positions of the measurement points defined by $x_{\lambda 1}$ and $x_{\lambda 2}$ are retained, rather than just the relative positions given by the baseline components. By using the Fourier transform relationship between $E(l, t)$ and $e(u, t)$ derived in Eq. (14.19), and replacing u by x_λ, we obtain

$$\Gamma(x_{\lambda 1}, x_{\lambda 2}, \tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \gamma(l_1, l_2, \tau) e^{-j2\pi(x_{\lambda 1} l_1 - x_{\lambda 2} l_2)}\, dl_1\, dl_2\,, \tag{14.24}$$

and the inverse transform, which is

$$\gamma(l_1, l_2, \tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Gamma(x_{\lambda 1}, x_{\lambda 2}, \tau) e^{j2\pi(x_{\lambda 1} l_1 - x_{\lambda 2} l_2)}\, dx_{\lambda 1}\, dx_{\lambda 2}\,. \tag{14.25}$$

The relationships in Eqs. (14.24) and (14.25) do not provide a means of measuring the intensity distribution of a source, except in the case of complete incoherence. For complete incoherence, the coherence function can be expressed as

$$\gamma(l_1, l_2, \tau) = \gamma(l_1, \tau)\delta(l_1 - l_2), \tag{14.26}$$

where δ is the delta function. Using the relation in Eq. (14.26) in conjunction with Eqs. (14.24) and (14.25), we find that the self-coherence function of a completely incoherent source and its spatial frequency spectrum are Fourier transforms of each other:

$$\Gamma(u, \tau) = \int_{-\infty}^{\infty} \gamma(l, \tau) e^{-j2\pi ul}\, dl \tag{14.27}$$

$$\gamma(l, \tau) = \int_{-\infty}^{\infty} \Gamma(u, \tau) e^{j2\pi ul}\, du, \tag{14.28}$$

where $u = x_{\lambda 1} - x_{\lambda 2}$. It is clear that $\Gamma(u, \tau)$ is independent of $x_{\lambda 1}$ and $x_{\lambda 2}$ and depends only on their difference. As explained in Section 2.3 under *Convolution Theorem and Spatial Frequency*, u can be interpreted as the spacing of two sample points between which the coherence of the field is measured, and also as the spatial frequency of the visibility measured over the same baseline. For $\tau = 0$, from Eqs. (14.21) and (14.26),we obtain

$$\gamma(l, 0) = \langle |E(l)|^2 \rangle, \tag{14.29}$$

which is the one-dimensional intensity distribution of the source, I_1, introduced in Eq. (1.9). Then from Eqs. (14.27) and (14.29)

$$\Gamma(u, 0) = \int_{-\infty}^{\infty} \langle |E(l)|^2 \rangle e^{-j2\pi ul}\, dl. \tag{14.30}$$

$\Gamma(u, 0)$ is measured between points along a line normal to the direction $l = 0$. As measured with an interferometer, it is also the complex visibility $\mathcal{V}$. Equation (14.30) is the Fourier transform relationship between mutual coherence (visibility) and intensity.

When the incoherence condition in Eq. (14.26) is introduced into Eqs. (14.24) and (14.25), two results appear: the van Cittert–Zernike relation between mutual coherence and intensity, and the stationarity of the mutual coherence with respect to u. The physical reason underlying these results is seen in Fig. 14.2. When the wavefronts incident at different angles combine at any point, the relative phases of their (Fourier) frequency components vary linearly with the position of the point (e.g., the position of A along the line OB in Fig. 14.2), and for small l they also vary linearly with the angle on the sky. As a result, the phase differences of the Fourier components at two points depend only on the relative positions of the points, not their absolute positions. Interferometer measurements of mutual coherence incorporate the phase differences for a range of angles of incidence

governed by the angular dimensions of the source and the width of the antenna beams. The linear relationship between phase and position angle allows us to recover the angular distribution of the incident wave intensity from the variation of the mutual coherence as a function of u, by Fourier analysis. If the angular width of the source is small enough that the distance AA' in Fig. 14.2 is always much less than the wavelength, then the form of the electric field remains constant along the line OA, and the source is not resolved.

Completely Coherent Source

Parrent (1959) has shown that an extended source can be completely coherent only if it is monochromatic. As examples of such a source one may visualize the aperture of a distant, large antenna, or an ensemble of radiating elements all driven by the same monochromatic signal. The aperture considered in Section 14.1 under *Diffraction at an Aperture and the Response of an Antenna* is a conceptual example of a coherent source. The difference between the responses of an interferometer to a fully coherent source and to a fully incoherent one can be explained by the following physical picture. The source can be envisioned as an ensemble of radiators distributed over a solid angle on the sky. In the case of a coherent source the signals from the radiators are monochromatic and coherent. The radiation in any direction combines into a single monochromatic wavefront and produces a monochromatic signal in each antenna of an interferometer. The output of the correlator is directly proportional to the product of the two (complex) signal amplitudes from the antennas. Thus if a coherent source is observed with n_a antennas, the $n_a(n_a - 1)/2$ pairwise cross-correlations of the signals that are measured can be factored into n_a values of complex signal amplitude.

In contrast, for an incoherent source the outputs from radiating elements are uncorrelated and must be considered independently. Each one produces a component of the fringe pattern in the correlator output. But since the phases of these fringe components depend on the positions of the radiators within the source, the combined response is proportional not only to the signal amplitudes at the antennas but also to a factor that depends on the angular distribution of the radiators. This factor, of magnitude ≤ 1, is equal to the modulus of the visibility normalized to unity for an unresolved (point) source of flux density equal to that of the source under observation. Unless the source is unresolved, it is not possible to factor the measured cross-correlations into signal amplitude values at the antennas. Because the emissions of the radiating elements of a source are uncorrelated, the information on the source distribution is preserved in the ensemble of wavefronts they produce at the antennas.

As shown by the derivation of the angular dependence of the radiation from a coherently illuminated aperture [Eq. (14.12)], and suggested by the analogy with a large antenna, the radiation from a coherent source is highly directional. Thus the signal strengths observed depend on the absolute positions of the two antennas of an interferometer, as in Eqs. (14.24) and (14.25), not only on their relative positions as is the case for an incoherent source. The ability to factor the signal outputs from a series of baselines, and the nonstationarity of the correlator output

measurements with the absolute positions of the antennas, are two characteristics that could allow a coherent source to be recognized (MacPhie 1964). From the analysis in Section 14.1, it is clear that a similar range of antenna spacings is required to resolve an incoherent source or to explore the radiation pattern of a coherent source of the same angular size.

14.3 SCATTERING AND THE PROPAGATION OF COHERENCE

It is well known that optical telescope images of single stars made with exposure times short compared with the timescale of atmospheric scintillation exhibit multiple stellar images (see Section 16.4 under *Speckle Imaging*). These images result from the scattering of light from the star by irregularities in the earth's atmosphere. Something closely analogous to this occurs in the case of imaging of an unresolved radio source through a medium with strong irregular scattering, such as the interplanetary medium within a few degrees of the sun, as described in Section 13.5 under *Interplanetary Scintillation*. Since each scattered image results from the emission of the same source, one is led to expect that such a situation would simulate the effect of a distribution of coherent point sources. In this section we examine the effects of scattering by considering the propagation of coherence in space, following in part a discussion by Cornwell, Anantharamaiah, and Narayan (1989). This formalism suggests methods for the recovery of the unscattered image from the observed image.

Given a radiating surface, we wish to know the mutual coherence function on another (possibly virtual) surface in space. In the typical radio astronomy situation, a number of simplifying assumptions can be made about the geometry of the problem. Consider the situation illustrated in Fig. 14.3, in which narrowband radio waves propagate from surface S to surface Q. The mutual coherence of two points in space is the expectation of the product of the (copolarized) electric fields at the two points. For signals correlated with arbitrary time delay, the mutual coherence is

$$\Gamma(Q_1, Q_2, \tau) = \langle E(Q_1, t) E^*(Q_2, t - \tau) \rangle. \tag{14.31}$$

The mutual coherence function Γ is a function of the field at two points and the time difference τ. We consider the propagation of *mutual intensity*, that is, the mutual coherence evaluated for $\tau = 0$. Following common practice, we represent the mutual intensity by $J(Q_1, Q_2) \equiv \Gamma(Q_1, Q_2, 0)$. J will be subscripted by S, Q, or B to indicate the corresponding plane (Fig. 14.3) of the mutual intensity value. We assume that the emitting surface is completely incoherent, as is usually the case for astronomical objects, and that the observed radiation is restricted to a narrow band of frequencies as dictated by the characteristics of the receiving system. From Eq. (14.31) and the Huygens–Fresnel formulation of radiation, it can be shown (Born and Wolf 1999, Goodman 1985), by a calculation similar to the one used in deriving Eq. (14.6), that the mutual intensity for points Q_1 and Q_2 is

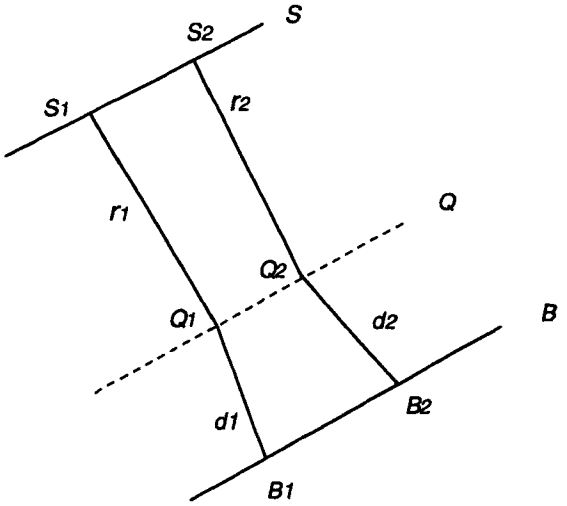


Figure 14.3 Simplified geometry for examining the propagation of coherence. S represents an extended source, Q is the location of a scattering screen, and B is the measurement plane. Surfaces S, Q, and B are plane and parallel, and r_1, r_2, d_1, and d_2 are much greater than the wavelength. All rays are nearly (but not necessarily exactly) perpendicular to the surfaces.

$$J_Q(Q_1, Q_2) = \lambda^{-2} \iint_S J_S(S_1, S_2) \frac{\exp[-j2\pi(r_1 - r_2)/\lambda]}{r_1 r_2} dS_1\, dS_2, \qquad (14.32)$$

where $dS_1\, dS_2$ is a surface element of S, and λ is the wavelength at the center of the observed frequency band.

The condition of incoherence can be represented by the use of a delta function (Beran and Parrent 1964), as in Eq. (14.26). Here the mutual intensity is represented by a delta function, and thus the intensity distribution on the surface Q is found by allowing points Q_1 and Q_2 to merge:

$$J_S(S_1, S_2) = \lambda^2 I(S_1)\delta(S_1 - S_2), \qquad (14.33)$$

where the factor λ^2 has been included to preserve the physical dimension of intensity. Equation (14.32) then becomes

$$J_Q(Q_1, Q_2) = \int_S I(S_1) \frac{\exp[-j2\pi(r_1 - r_2)/\lambda]}{r_1 r_2} dS. \qquad (14.34)$$

When the angular dimension of the source is infinitesimal, that is, when the source is unresolved, the integration over the source becomes trivial and the mutual intensity can be factored into terms depending, respectively, on r_1 and r_2:

$$J_Q(Q_1, Q_2) = I(S) \left(\frac{\exp(-j2\pi r_1/\lambda)}{r_1} \right) \left(\frac{\exp(j2\pi r_2/\lambda)}{r_2} \right), \qquad (14.35)$$

where r_1 and r_2 now originate at a single point S. In the more general case of a resolved source Eq. (14.34) cannot be factored. Equations (14.34) and (14.35) describe for their respective cases the propagation of mutual coherence in situations subject to the constraints of Fig. 14.3, and thus can be used to determine the mutual intensity on surface Q resulting from incoherent radiation from surface S. Examination of Eq. (14.31) reveals that, for the extended source S, the mutual intensity on Q depends on both r_1 and r_2 for all pairs of points on Q. Thus the field at Q is at least partially coherent for all sources, including those of finite extent. This is intuitively reasonable, as all points on Q are illuminated by all points on S. In fact, it can be demonstrated rigorously that an incoherent field cannot exist in free space (Parrent 1959).

Suppose now that we have a situation in which the surface Q is actually a screen of irregularities in the transmission medium, such as plasma or dust, which scatters the radiation from S. The mutual intensity incident on the screen is modified by a complex transmission factor $T(Q)$ to produce the transmitted mutual intensity

$$J_{Qt}(Q_1, Q_2) = T(Q_1)T^*(Q_2)J_{Qi}(Q_1, Q_2), \tag{14.36}$$

where subscripts i and t indicate the incident and transmitted mutual intensity, respectively. From Eq. (14.34) we now define a "propagator" (Cornwell, Anantharamaiah, and Narayan 1989) for mutual intensity:

$$W(S, B) = \int_S \frac{T(Q)\exp[-j2\pi(r+d)/\lambda]}{r\,d}\,dS, \tag{14.37}$$

where r and d are defined in Fig. 14.3. Then the mutual intensity on surface B is given, in terms of the mutual intensity of an extended source S, by

$$J_B(B_1, B_2) = \lambda^{-4}\iint_S J_S(S_1, S_2)W(S_1, B_1)W^*(S_2, B_2)\,dS_1\,dS_2. \tag{14.38}$$

For an incoherent extended source

$$J_B(B_1, B_2) = \lambda^{-2}\int_S I(S)W(S, B_1)W^*(S, B_2)\,dS, \tag{14.39}$$

and for a point source of flux density F, the mutual intensity on B becomes

$$J_B(B_1, B_2) = F\lambda^{-2}W(S, B_1)W^*(S, B_2). \tag{14.40}$$

Again, for the unresolved source the mutual intensity on B consists of two factors, each depending only on one position on B. For an extended incoherent source distribution on S, however, the mutual intensity depends on *differences* in position and therefore cannot be factored.

The existence of a scattering screen between a source and an observer with an instrument of limited aperture raises the possibility of greatly increased angular resolution resulting from the much larger extent of the scattering screen. The partial coherence of radiation from the screen requires that the intensity be measured at all points on the measurement plane B, spaced as dictated by the Nyquist criterion, rather than at all points in the spatial frequency spectrum as allowed by the van Cittert–Zernike theorem. The former observing mode results in very much more data than does the latter. In two spatial dimensions a large redundancy of data results, so that in principle not only can the scattering screen be characterized, but the source as well. In this respect the problem is similar to that of self-calibration (Section 11.4). Unfortunately, in the case of the scattering screen, the practical difficulties of such observations are enormous, and few significant attempts have been made to apply the principle. Cornwell and Narayan (1993) discuss the possibilities of statistical image synthesis using scattering to obtain ultrafine resolution in a manner somewhat analogous to speckle imaging (see Section 16.4).

Emission from a radio source that undergoes strong scattering during propagation through space has been investigated by Anantharamaiah, Cornwell, and Narayan (1989), and Cornwell, Anantharamaiah, and Narayan (1989). To demonstrate the response of a radio telescope to such a spatially coherent source distribution, they observed the strong and essentially pointlike source 3C279, which passes close to the sun each year. Under these conditions the scattering is strong enough to cause amplitude scintillation of the received signals. Anantharamaiah and colleagues used the VLA in its most extended configuration for which the longest baselines are approximately 35 km. The velocity of the solar wind, of order 100–400 km s^{-1}, causes irregularities to sweep across the array in ~100 ms, so it was necessary to make snapshot observations of duration 10–40 ms to avoid smearing of the image by the movement of the scattering screen. Observations were made at wavelengths of 20, 6, and 2 cm, with the source at angular distances of 0.9° to 5° from the sun. It was found that the correlator output values could be factored as expected for a coherent source. When correlated signals were averaged for about 6 s, an enlarged image of the source was obtained, and the enlargement increased as the distance from the sun decreased. It was also demonstrated that it would be possible to determine the characteristics of the scattering screen by measuring the mutual intensity function on the ground, provided that the latter is measured completely in the two-dimensional spatial frequency domain. It is not possible to distinguish between a spatially coherent extended source and a scattering screen illuminated by a point source.

A significant observation was made by Wolszczan and Cordes (1987), who were able to infer the dimensions of structure within pulsar PSR 1237 + 25 from an occurrence of interstellar scattering. The pulsar was observed with a single antenna, the 308-m spherical reflector at Arecibo, at a frequency of 430 MHz. Dynamic spectra of the received signal (i.e., the received power displayed as a function of both time and frequency) showed prominent band structure with maxima separated by ~300–700 kHz in frequency. This was interpreted in terms of a thin-screen model of the interstellar medium, in which refraction of rays

from the pulsar occurred at two separated points in the screen. The analysis of such a model is complicated by the occurrence of both diffractive and refractive scattering, resulting from structure smaller and larger than the Fresnel scale, respectively (Cordes, Pidwerbetsky, and Lovelace 1986). The refraction gave rise to two images of the source at the radio telescope, resulting in fringes in the intensity of the received signal. The distance of the pulsar (0.33 kpc) and its transverse velocity (178 km s^{-1}) were known from other observations, and the distance of the screen was taken to be half the distance of the pulsar. It was deduced that the angular separation of the images was ~3.3 mas, corresponding to a spacing of ~1 AU (astronomical unit) between the refracting structures. In effect the refracting structures constitute a two-element interferometer, with fringe spacing ~1 μarcsec. For comparison, the angular resolution of a baseline equal to the diameter of the earth at 430 MHz would be 44 mas. The particular conditions that resulted in this observation lasted for at least 19 days, and during that period observations of other pulsars did not show similar scattering. This strongly suggests that the observed phenomenon resulted from a fortuitous configuration of the interstellar medium in the direction of the pulsar.

Apart from cases of scattering such as that described, there are essentially no clear cases of spatially coherent astronomical sources, although coherent mechanisms may occur in pulsars and masers (Verschuur and Kellermann 1988). Fully coherent sources are not amenable to synthesis mapping using the van Cittert–Zernike principle, and thus do not fall within the area of principal concern of this book. Further material on coherence and partial coherence can be found, for example, in Beran and Parrent (1964), Born and Wolf (1999), Drane and Parrent (1962), Mandel and Wolf (1965, 1995), MacPhie (1964), and Goodman (1985).

REFERENCES

Anantharamaiah, K. R., T. J. Cornwell, and R. Narayan, Synthesis Imaging of Spatially Coherent Objects, in *Synthesis Imaging in Radio Astronomy*, R. A. Perley, F. R. Schwab, and A. H. Bridle, Eds., Astron. Soc. Pac. Conf. Ser., **6**, 415–430, 1989.

Beran, M. J. and G. B. Parrent, Jr., *Theory of Partial Coherence*, Prentice-Hall, Englewood Cliffs, NJ, 1964; repr. by Society of Photo-Optical Instrumentation Engineers, Bellingham, WA, 1974.

Booker, H. G. and P. C. Clemmow, The Concept of an Angular Spectrum of Plane Waves, and Its Relation to that of Polar Diagram and Aperture Distribution, *Proc. IEE*, **97**, 11–17, 1950.

Born, M. and E. Wolf, *Principles of Optics*, 7th ed., Cambridge Univ. Press, Cambridge, UK, 1999.

Bracewell, R. N., Radio Astronomy Techniques, in *Handbuch der Physik*, Vol. 54, S. Flugge, Ed., Springer-Verlag, Berlin, 1962, pp. 42–129, .

Bracewell, R. N., *The Fourier Transform and Its Applications*, McGraw-Hill, New York, 2000 (earlier eds. 1965, 1978).

Cordes, J. M., A. Pidwerbetsky, and R. V. E. Lovelace, Refractive and Diffractive Scattering in the Interstellar Medium, *Astrophys. J.*, **310**, 737–767, 1986.

Cornwell, T. J., K. R. Anantharamaiah, and R. Narayan, Propagation of Coherence in Scattering: An Experiment Using Interplanetary Scintillation, *J. Opt. Soc. Am.*, **6A**, 977–986, 1989.

Cornwell, T. J., and R. Narayan, Imaging with Ultra-Resolution in the Presence of Strong Scattering, *Astrophys. J.*, **408**, L69–L72, 1993.

Drane, C. J. and G. B. Parrent, Jr., On the Mapping of Extended Sources with Nonlinear Correlation Antennas, *IRE Trans. Antennas Propag.*, **AP-10**, 126–130, 1962.

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

MacPhie, R. H., On the Mapping by a Cross Correlation Antenna System of Partially Coherent Radio Sources, *IEEE Trans. Antennas Propag.*, **AP-12**, 118–124, 1964.

Mandel, L., and E. Wolf, Coherence Properties of Optical Fields, *Rev. Mod. Phys.*, **37**, 231–287, 1965.

Mandel, L., and E. Wolf, *Optical Coherence and Quantum Optics*, Cambridge Univ. Press, 1995.

Parrent, G. B., Jr., Studies in the Theory of Partial Coherence, *Opt. Acta.*, **6**, 285–296, 1959.

Silver, S., *Microwave Antenna Theory and Design*, Radiation Laboratory Series Vol. 12, McGraw-Hill, New York, 1949, p. 174.

Swenson, G. W., Jr. and N. C. Mathur, The Interferometer in Radio Astronomy, *Proc. IEEE*, **56**, 2114–2130, 1968.

Verschuur, G. L. and K. I. Kellermann, Eds., *Galactic and Extragalactic Astronomy*, Springer-Verlag, New York, 1988.

Wolszczan, A. and J. M. Cordes, Interstellar Interferometry of the Pulsar PSR 1237+25, *Astrophys. J.*, **320**, L35–39, 1987.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

15 Radio Interference

With the increasing use of the radio spectrum for communications, navigation, and other services, the avoidance of unwanted signals is an essential practical concern in radio astronomy. Interference poses particular problems to the radio astronomer because the signal levels from cosmic sources are much lower than the operating levels in active (transmitting) services, and wide bandwidths are required for adequate sensitivity. Although certain frequency bands are allocated solely to radio astronomy and passive (non-transmitting) sensing, some of those at meter and centimeter wavelengths are too narrow to allow the desired sensitivity to be obtained. Also, many cosmic spectral line frequencies fall outside the radio astronomy bands. Thus it is sometimes necessary for radio astronomers to observe within bands that are allocated to other services. Interference can then best be avoided by placing radio telescopes in locations remote from centers of industrial and similar activity and by taking advantage of shielding from transmitters by terrain features. A basic parameter in site selection and coordination with other spectrum users is the threshold of harmful interference, that is, the flux density above which an interfering signal falling within the passband of the radio telescope is detrimental to astronomical observations. The harmful threshold is a function of the type and operating parameters of the radio telescope, and this dependence is the principal concern of this chapter. The international system of regulation of the radio spectrum is briefly described in Appendix 15.1.

15.1 GENERAL CONSIDERATIONS

The ultimate limit on the sensitivity of a radio telescope is set by the system noise, and an interfering signal can generally be tolerated if its contribution to the output is small compared with the noise fluctuations. A response to interference of one-tenth the rms level of the noise in the measurements is a useful criterion in interference threshold calculations. The corresponding flux density of such a signal can be calculated if the effective collecting area of the antenna is known. Radio astronomy antennas usually have narrow beams, and the probability of the interfering signal being received in the main beam or nearby sidelobes is low, especially if the interfering transmitter is ground-based. Thus it can be assumed that interference usually enters the far sidelobes of the antenna. Figure 15.1 shows an empirical model curve for the maximum sidelobe gain as a function of angle from the main-beam axis. This curve is derived from the measured response patterns

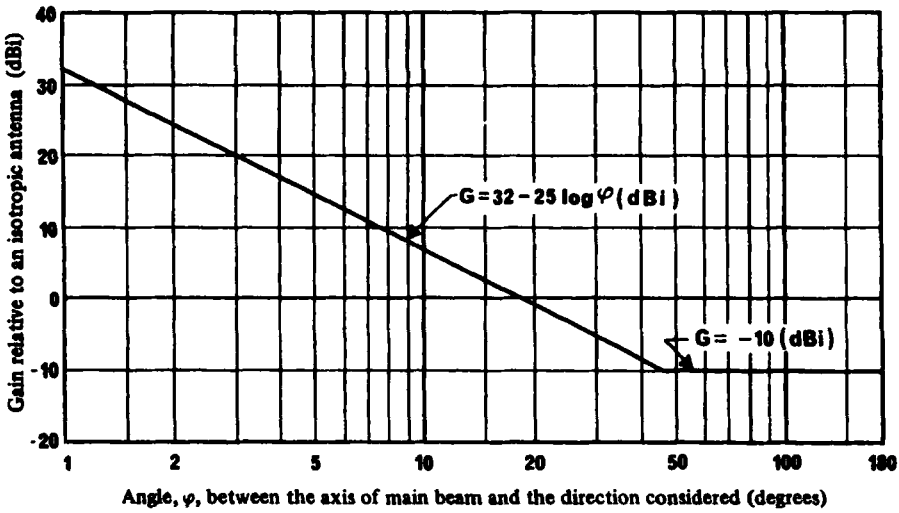


Figure 15.1 Empirical sidelobe-envelope model for reflector antennas of diameter greater than 100 wavelengths. Measurements on actual antennas show that 90% of sidelobe peaks lie below the curve. Sidelobe levels can be reduced by 3 dB or more in designs in which aperture blockage by feed structure is eliminated or minimized. The model shown is representative of large antennas with tripod or quadrupod feed supports of the type commonly used in radio astronomy. From ITU-R (1997a).

of a number of large reflector antennas. For the present calculations it is appropriate to use a gain of 0 dBi (i.e., 0 dB with respect to an isotropic radiator), which occurs at about 19° from the main beam. Zero dBi is also the mean gain of an antenna over 4π sr, and the effective collecting area for this gain is equal to $\lambda^2/4\pi$, where λ is the wavelength. If F_h (W m^{-2}) is the flux density of an interfering signal within the receiver passband, the interference-to-noise power ratio in the receiver is

$$\frac{F_h \lambda^2}{4\pi k T_S \, \Delta\nu}, \tag{15.1}$$

where k is Boltzmann's constant, T_S is the system noise temperature, and $\Delta\nu$ is the receiver bandwidth. In this expression it is assumed that the polarization of the interfering signal matches that of the antenna. Since radio astronomy antennas commonly receive two polarizations, crossed linear or opposite circular, choice of antenna polarization is of little help in avoiding interference. In practice the received level of the interfering signal varies with time because of propagation effects and the tracking motion of the radio telescope, which sweeps the sidelobe pattern across the direction of the transmitter.

For comparison with correlator systems, we first consider the simpler case of a receiver that measures the total power at the output of a single antenna. The interference-to-noise ratio of the output, after square-law detection and averaging for a time τ_a, is expression (15.1) multiplied by $\sqrt{\Delta\nu\tau_a}$. This result follows from

considerations similar to those discussed in Section 6.2 under *Signal and Noise Processing in the Correlator*. Then for an output interference-to-noise ratio of 0.1, which we use as the criterion for the threshold of harmful interference,

$$F_h = \frac{0.4\pi k T_S \nu^2 \sqrt{\Delta\nu}}{c^2 \sqrt{\tau_a}}. \tag{15.2}$$

Note that the harmful threshold increases with frequency as ν^2 as a result of the dependence of the sidelobe collecting area. With increasing frequency the system temperature and the usable bandwidth also generally increase. Expressed in spectral power flux density, the corresponding threshold level, S_h(W m^{-2} Hz^{-1}), is

$$S_h = \frac{F_h}{\Delta\nu} = \frac{0.4\pi k T_S \nu^2}{c^2 \sqrt{\tau_a \, \Delta\nu}}. \tag{15.3}$$

To determine the harmful interference level for continuum observations within a band allocated to radio astronomy, $\Delta\nu$ is usually taken to be the width of the allocated band. The total-power type of radio telescope is the most sensitive to interference. Thus, the results in Eqs. (15.2) and (15.3) provide a worst-case specification for the harmful thresholds of interference for radio astronomy. Values of F_h and S_h computed for total-power systems using typical parameters for the various radio astronomy bands are given in ITU-R (1995) and ITU-R (1997b). For S_h, the values are plotted as the bottom curve in Fig. 15.2. Since much of the interference to radio astronomy results from broadband spurious emissions, S_h is particularly useful.

Low level interference, of amplitude comparable to the noise in the receiver output, degrades the sensitivity and impedes the ability to detect weak sources. For stronger sources such interference degrades the accuracy of measurements, and thus reduces the possibility of detecting fine details or variations in structure or intensity, which are often key to new discoveries in astronomy. Thus in observations in which interference has occurred, it is necessary to delete any data that appear to be corrupted.

The analysis that follows considers the response to interference resulting from basic methods of observation and data reduction, and does not include procedures designed specifically for mitigation of interference. Such procedures include adaptive cancellation of interfering signals in the receiver, and adaptive nulling of the response of an array in the direction of incoming interference [see, e.g., Barnbaum and Bradley (1998)].

15.2 SHORT- AND INTERMEDIATE-BASELINE ARRAYS

We now consider the interference response of a correlator array with antenna spacings up to a few tens of kilometers, typical of connected-element arrays. Two effects reduce the response to interference compared to that of a total-power

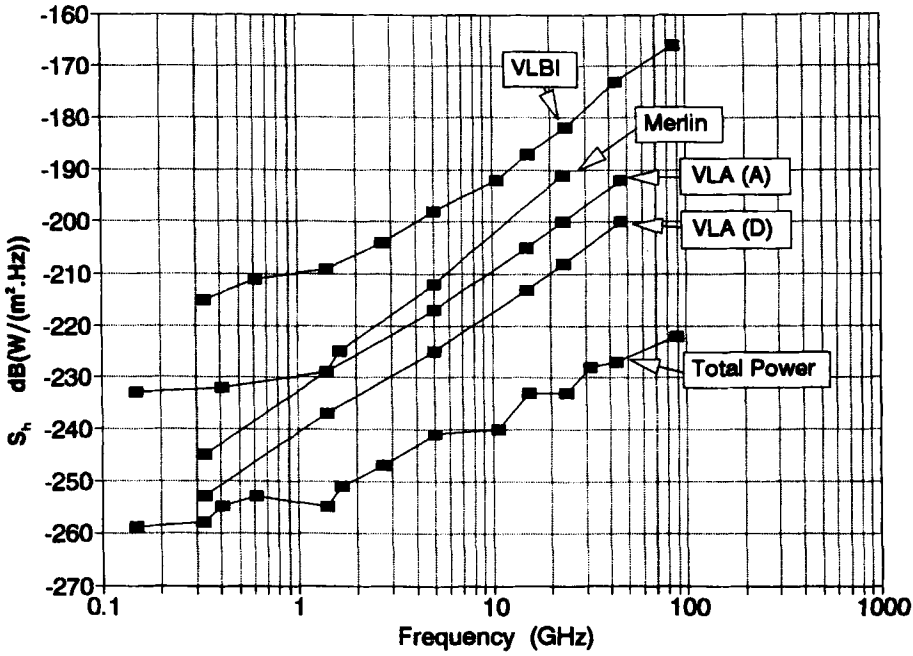


Figure 15.2 Curves of the harmful threshold of interference S_h, in decibel units of spectral power flux density dBW m^{-2} Hz^{-1}, for continuum observations. These are computed using typical instrumental characteristics for each frequency band and type of instrument. The curve for total-power radiometers is based on Eq. (15.3) with values from ITU-R (1997b). Connected-element arrays are represented by the VLA, with curves for the most compact and the most extended configurations, and by the MERLIN array. Curves for connected-element arrays are derived from Eq. (15.15). The curve for VLBI systems is based on Eq. (15.25). Note that for arrays, at any given frequency, S_h increases as the synthesized beamwidth is reduced.

system. First, the source of interference does not move across the sky with the sidereal motion of the object under observation, and thus it produces fringe oscillations of a different frequency from those of the wanted signal. Second, the instrumental delays are adjusted to equalize the signal paths for radiation incident from the direction under observation, and signals from another direction, if they are broadband, are to some extent decorrelated. The following analysis is based on Thompson (1982).

Fringe-Frequency Averaging

Consider first the fringe-frequency effect. Suppose that instrumental phase shifts are introduced, as described in Section 6.1 under *Delay Tracking and Fringe Rotation*, to slow the fringe oscillations of the wanted signal to zero frequency. The removal of the fringe-frequency phase shifts from the cosmic signals introduces corresponding shifts into the interfering signals. If the source of interference is

stationary with respect to the antennas, the interference at the correlator output has the form of oscillations at the natural fringe frequency for the source under observation, which from Eq. (4.9) (omitting the sign of dw/dt) is

$$\nu_f = \omega_e u \cos\delta. \tag{15.4}$$

Here ω_e is the angular rotation velocity of the earth, u is a component of antenna spacing, and δ is the declination of the source under observation. Averaging of such a fringe-frequency waveform for a period τ_a is equivalent to convolution with a rectangular function of width τ_a. The amplitude is thus decreased by a factor that follows from the Fourier transform of the convolving function. This factor is

$$f_1 = \frac{\sin(\pi \nu_f \tau_a)}{\pi \nu_f \tau_a}. \tag{15.5}$$

In order to derive a harmful threshold of interference, we compute the ratio of the rms level of interference to the rms level of noise in a radio map and, as before, equate the result to 0.1. The first step is to determine the mean squared value of the modulus of the interference component in the visibility data. Figure 6.7b, which depicts the spectral components at the correlator output, shows that the output from the correlated signal component, in this case the interference, is represented by a delta function. Assuming, as before, that the interference enters sidelobes of gain 0 dBi, and that the polarization is matched, we substitute in the magnitude of the delta function $kT_A\,\Delta\nu = F_h c^2/4\pi\nu^2$. Thus, the sum of the squared modulus of the interference over n_r grid points in the (u, v) plane is

$$\sum_{n_r} \langle |r_i|^2 \rangle = \left(\frac{H_0^2 F_h c^2}{4\pi\nu^2}\right) n_r \langle f_1^2 \rangle. \tag{15.6}$$

Here r_i is the correlator response to the interference, H_0 is a voltage gain factor, and $\langle f_1^2 \rangle$ is the mean squared value of f_1 as given in Eq. (15.5), which represents the effect of the visibility averaging on the fringe-frequency oscillations. To determine the mean squared value of f_1, a simple approach is to consider the variation of this factor in the (u', v') plane in which the spacing vector rotates with constant angular velocity ω_e, and sweeps out a circular locus as described in Section 4.2. Also, suppose that to interpolate the values of visibility at the rectangular grid points in the (u, v) plane, the measured values are averaged with uniform weight within rectangular cells centered on the grid points (see the description of cell averaging in Section 5.2 under *Discrete Two-Dimensional Fourier Transform*). Then the effective averaging time τ for the interference is equal to the time taken by the baseline vector to cross a cell, as shown in Fig. 15.3. Note from Eq. (15.4) that the fringe frequency goes through zero at the v' axis, and f_1 is then unity. For small values of ψ, as defined in Fig. 15.3, the path length through a cell is closely equal to Δu, and the cell crossing time is $\tau = \Delta u/\omega_e q'$, where $q' = \sqrt{u'^2 + v'^2}$. Also, $\nu_f \tau = \Delta u \sin\psi \cos\delta$. Now Δu is equal to the reciprocal of the width of

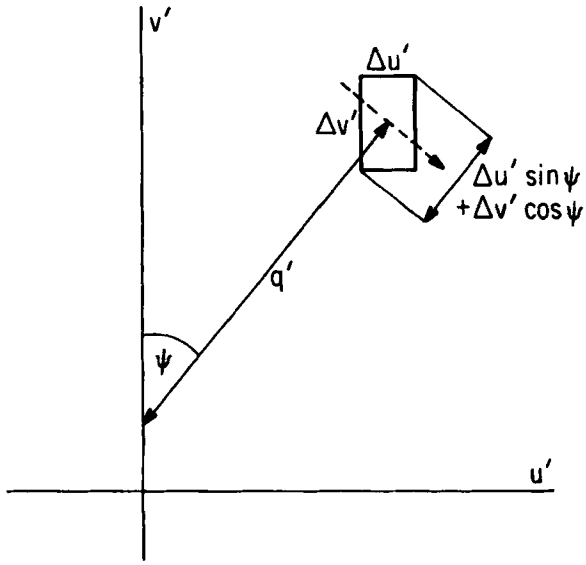


Figure 15.3 Derivation of the mean cell crossing time for the spatial frequency locus indicated by the broken line. The velocity of the spatial frequency vector in the (u', v') plane is $\omega_e q'$. The mean path length through the cell in the direction of the broken line is the cell area $\Delta u' \, \Delta v'$ divided by the cell width projected normal to that direction.

the synthesized field, which, except at long wavelengths, is unlikely to be more than 0.5°. We therefore assume that Δu is of order 100 or greater, which permits the following simplification. For $\Delta u = 100$ and $\delta < 70°$, f_1^2 goes from 1 to 10^{-3} as ψ goes from 0 to $< 17°$. Thus, most of the contribution to f_1^2 occurs for small ψ, and we can substitute $\nu_f \tau = \psi \, \Delta u \cos\delta$ in Eq. (15.5) and obtain

$$\langle f_1^2 \rangle = \frac{2}{\pi} \int_0^{\pi/2} \frac{\sin^2(\pi \psi \, \Delta u \cos\delta)}{(\pi \psi \, \Delta u \cos\delta)^2} d\psi \simeq \frac{1}{\pi \, \Delta u \cos\delta}. \tag{15.7}$$

Since Δu is large, we have used an upper limit of ∞ in evaluating the integral.

For the noise we again refer to Fig. 6.7b. The power spectral density of the noise near zero frequency is $H_0^4 k^2 T_S^2 \Delta\nu$, and an equivalent bandwidth τ^{-1}, including negative frequencies, is passed by the averaging process; see Eq. (6.44). The mean-squared component of the noise over the n_r grid points is thus

$$\sum_{n_r} \langle |r_n| \rangle^2 = H_0^4 k^2 T_S^2 \Delta\nu n_r \langle \tau^{-1} \rangle, \tag{15.8}$$

where $\langle \tau^{-1} \rangle$ is the mean value of τ^{-1}. From Fig. 15.3 the mean cell crossing time is

$$\tau = \frac{\Delta u \, |\operatorname{cosec}\delta|}{q' \omega_e \, (|\sin\psi| + |\operatorname{cosec}\delta| \, |\cos\psi|)}, \tag{15.9}$$

where $q' = \sqrt{X_\lambda^2 + Y_\lambda^2}$, and where X_λ and Y_λ are the components of antenna spacing projected onto the equatorial plane, as defined in Section 4.1. We have assumed that $\Delta u' = \Delta v' \sin\delta$ (i.e., $\Delta u = \Delta v$) and that, for all except a small number of cells, the path of the spatial frequency locus through a cell can be approximated by a straight line. The mean value of τ^{-1} around a locus in the (u', v') plane (see Section 4.2) is, from Eq. (15.9),

$$\frac{2}{\pi}\int_0^{\pi/2} \tau^{-1} d\psi = \frac{2\omega_e q'}{\pi \Delta u}(1 + |\sin\delta|), \tag{15.10}$$

and the mean for the n_r points in the (u, v) plane is

$$\langle \tau^{-1} \rangle = \frac{2\omega_e}{\pi \Delta u}(1 + |\sin\delta|)\frac{1}{n_r}\sum_{n_r} q'. \tag{15.11}$$

From Eqs. (15.6)–(15.8) and (15.11) the interference-to-noise ratio is

$$\frac{(|r_i|)_{\text{rms}}}{(|r_n|)_{\text{rms}}} = \frac{F_h c^2}{4\pi k T_S \nu^2 \sqrt{2\,\Delta\nu\omega_e}\cos\delta(1 + |\sin\delta|)}\frac{1}{\sqrt{\frac{1}{n_r}\sum_{n_r} q'}}. \tag{15.12}$$

By Parseval's theorem the ratio of the rms values of the interference and noise in the map is equal to the same ratio in the visibility domain, which is given by Eq. (15.12). To evaluate the harmful threshold F_h, we equate the right-hand side to 0.1 and obtain

$$F_h = \frac{0.4\pi k T_S \nu^2 \sqrt{2\,\Delta\nu\omega_e}}{c^2}\sqrt{\frac{1}{n_r}\sum_{n_r} q'}. \tag{15.13}$$

The factor $\sqrt{\cos\delta(1 + |\sin\delta|)}$ has been replaced by unity, the resulting error being less than 1 dB for $0 < |\delta| < 71°$, and 2.3 dB for $\delta = 80°$. Note that with fixed antenna positions q' is proportional to ν, so F_h is proportional to $\nu^{2.5}$. The number of points in the (u', v') plane to which an antenna pair contributes is proportional to q', so in evaluating Eq. (15.13) it is convenient to write

$$\frac{1}{n_r}\sum_{n_r} q' = \frac{\sum_{n_p} q'^2}{\sum_{n_p} q'}, \tag{15.14}$$

where n_p is the number of correlated antenna pairs in the array.

The interference threshold S_h, in units of dBW m^{-2} Hz^{-1}, is given by

$$S_h = \frac{F_h}{\Delta\nu} = \frac{0.4\pi k T_S \nu^2 \sqrt{2\omega_e}}{c^2\sqrt{\Delta\nu}}\sqrt{\frac{1}{n_r}\sum_{n_r} q'}. \tag{15.15}$$

Values of S_h for the VLA and the MERLIN array are shown in Fig. 15.2. Of the two curves for the VLA, the lower and the upper correspond to configurations in which the distance over which the antennas are distributed along each arm is 0.59 and 21 km, respectively (see Fig. 5.17b).

Since the averaging is ineffective in reducing the interference when u goes through zero, visibility values containing the greatest contributions from interference cluster around the v axis. Some degree of randomness in the occurrence of high values is to be expected, as a result of the varying sidelobe levels through which the interference enters. Because of the (u, v) distribution, the interference in the (l, m) domain takes the form of quasi-random structure that is elongated in the east–west direction; for an example see Thompson (1982). The clustering also suggests the possibility of reducing the interference response by deleting any suspect visibility data near the v axis. The resulting degradation of the (u, v) coverage would increase the sidelobes of the synthesized beam. The effect of such sidelobes could be mitigated to some degree by the deconvolution procedures discussed in Chapter 11.

The discussion above applies to cases where the observation is of sufficiently long duration that the (u, v) plane is well sampled, and where the strength of the interfering signal remains approximately constant during this time. If only a fraction α of the (u, v) loci cross the v axis, then a factor of $\sqrt{\alpha}$ should be introduced into the denominators of Eqs. (15.13) and (15.15). Strong, sporadic interference can produce different responses from that considered above.

Decorrelation of Broadband Signals

Since interfering signals are usually incident from directions other than that of the desired radiation, their time delays to the correlator inputs are generally not equal. Broadband interfering signals are thereby decorrelated, which further reduces their response. The reduction is not amenable to a general-case analysis like that resulting from averaging of the fringe frequency, but it can be computed for each particular antenna configuration and position of the interfering source. For this reason, and the fact that only broadband signals are reduced, the effect has not been included in the threshold equations (15.13) and (15.15).

At any instant during an observation, let θ_s be the angle between a plane normal to the baseline for a pair of antennas and the direction of the source under observation. θ_s defines a circle on the celestial sphere for which the delays are equalized. Similarly, let θ_i be the corresponding angle for the source of interference. The delay difference for the interfering signals at the correlator is

$$\tau_d = \frac{D|\sin\theta_s - \sin\theta_i|}{c}, \tag{15.16}$$

where D is the baseline length. Expressions for θ_s and θ_i can be derived from Eq. (4.3), since $\sin\theta_s = w\lambda/D$, where w is the third spacing coordinate as shown in Fig. 3.2, and λ is the wavelength. Suppose that the received interfering signal has an effectively rectangular spectrum of width $\Delta\nu$ and center frequency ν_0,

defined either by the signal itself or by the receiving passband. By the Wiener–Khinchin relation the autocorrelation function of the signal is equal to

$$\frac{\sin(\pi \Delta\nu\tau_d)}{\pi \Delta\nu\tau_d} \cos(2\pi \nu_0 \tau_d). \tag{15.17}$$

Expression (15.17) represents the real output of a complex correlator as a function of the differential delay τ_d. The imaginary output is represented by a similar expression in which the cosine function is replaced by a sine. Thus, the decorrelation of the modulus of the complex output for a delay τ_d is given by the factor

$$f_2 = \frac{\sin(\pi \Delta\nu\tau_d)}{\pi \Delta\nu\tau_d}. \tag{15.18}$$

For a fixed transmitter location, θ_i remains constant, but θ_s varies as the antennas track. Thus τ_d may go through zero, causing f_2 to peak, but unlike f_1, a peak in f_2 can occur at any point on the (u, v) plane. Those antenna pairs for which the f_1 and f_2 peaks overlap contribute most strongly to the interference in the map, and those for which the peaks are well separated contribute less. Therefore, for broadband signals, the fringe-frequency and decorrelation effects should be considered in combination. For example, in calculations for the response of the VLA to a geostationary satellite on the meridian, a factor

$$\sqrt{\frac{\sum q' f_1^2 f_2^2}{\sum q' f_1^2}} \tag{15.19}$$

was computed which represents the additional decrease in the rms interference resulting from decorrelation (Thompson 1982). The summations in (15.19) were taken over all antenna pairs for equal increments in hour angle, and the q' factors were inserted to compensate for the uneven density of points in the (u, v) plane resulting from this method of sampling. The antenna spacings of the VLA for both the most compact and most extended configurations were considered, with observing frequencies from 1.4 to 23 GHz and bandwidths of 25 and 50 MHz. The results indicate that suppression of broadband interference by decorrelation varies from 4 to 34 dB, with strong dependence on the observing declination. The interference was assumed to extend uniformly across the bandwidth, which would tend to overestimate the suppression in a practical situation.

15.3 VERY-LONG-BASELINE SYSTEMS

In VLBI arrays, in which the antenna spacings are hundreds or thousands of kilometers, the output resulting from correlated components of an interfering signal at the correlator inputs is usually negligible. This is because the natural fringe frequencies are higher than those in arrays with baselines up to a few tens of kilo-

meters, and the delay inequalities for signals that do not come from the direction of observation are also much greater. Furthermore, unless the interfering signal originates in a satellite or spacecraft, it is unlikely to be present at two widely separated locations.

Consider an interfering signal entering one antenna of a correlated pair. The interference reduces the measured correlation, and the overall effect is similar to an increase in the system noise for the antenna. In Fig. 15.4, $x(t)$ and $y(t)$ represent the signals plus system noise from two antennas in the absence of interference, and $z(t)$ represents an interfering signal at one antenna. The three waveforms have zero means, and the standard deviations are σ for x and y and σ_i for z. In the absence of interference, the measured correlation coefficient is

$$\rho_1 = \frac{\langle xy \rangle}{\sqrt{\langle x^2 \rangle \langle y^2 \rangle}} = \frac{\langle xy \rangle}{\sigma^2}. \tag{15.20}$$

When the interference is present, the correlation becomes

$$\rho_2 = \frac{\langle xy \rangle + \langle xz \rangle}{\sqrt{\langle x^2 \rangle \left(\langle y^2 \rangle + 2\langle yz \rangle + \langle z^2 \rangle \right)}}. \tag{15.21}$$

The interference is uncorrelated with x and y, so $\langle xz \rangle = \langle yz \rangle = 0$. Also, at the harmful threshold, $\sigma_i^2 \ll \sigma^2$. Thus, from Eqs. (15.20) and (15.21),

$$\rho_2 \simeq \rho_1 \left[1 - \tfrac{1}{2} \left(\frac{\sigma_i}{\sigma} \right)^2 \right]. \tag{15.22}$$

The interference reduces the measured correlation. In a system with automatic level control (ALC), the reduction in correlation can be envisaged as resulting

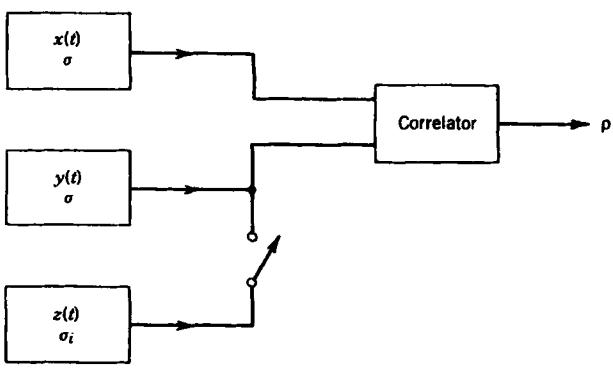


Figure 15.4 Components of the correlator input signals used in the discussion of the effects of interference on VLBI observations.

from a reduction in the system gain in response to the added power of the interference. The error introduced in the correlation measurement therefore takes the form of a multiplicative factor, rather than an additive error component. Interference causes additive errors in single antennas or arrays that have short enough baselines that the detector or correlator responds directly to the interfering signal. The different effects of these two types of error have been discussed in Section 10.6 under *Errors in Maps*. In principle, the change in the effective gain can be monitored by using a calibration signal, as discussed in Section 7.6. However, such a calibration process could be difficult if the strength of the interference varies rapidly. The harmful interference threshold should therefore be specified so that it is just small enough that the errors introduced do not significantly increase the level of uncertainty in the measurements. In general, a value of 1% for variations in the visibility amplitude resulting from interference is a reasonable choice. If we include the possibility of simultaneous but uncorrelated interference in both antennas, the resulting condition is

$$\left(\frac{\sigma_i}{\sigma}\right)^2 \leq 0.01. \tag{15.23}$$

It follows from Parseval's theorem, that a 1% rms error in the visibility introduces into the intensity an error of which the rms over the map is 1% of the corresponding rms of the true intensity distribution. The effect on the dynamic range of intensity within the map depends on the form of the intensity distribution and of the error distribution. For a map of a single point source, the rms intensity error would be about $10^{-2}\sqrt{f/n_r}$ times the peak intensity, where f is the fraction of the n_r gridded visibility data that contain interference. Here it is assumed that the fluctuations in the received interfering signal are sufficiently fast that the values of the interference level are essentially independent for each gridded visibility point. If this is not the case the resulting error will be greater.

To comply with the criterion in Eq. (15.23), the ratio of the powers of the interference to system noise as given by (15.1) must not exceed 0.01. Thus, for the harmful threshold, we have

$$F_h = \frac{0.04\pi k T_S \nu^2 \Delta\nu}{c^2}. \tag{15.24}$$

The interference threshold in units of W m^{-2} Hz^{-1} is

$$S_h = \frac{F_h}{\Delta\nu} = \frac{0.04\pi k T_S \nu^2}{c^2}. \tag{15.25}$$

Note that the interference-to-noise ratio of 0.01 here refers to the levels at the correlator input. In the case of total-power systems (single antennas) and the arrays considered in Section 15.2, for which the errors are additive, the criterion of an interference-to-noise ratio of 0.1 applies to the time-averaged output of the correlator or detector. This therefore results in lower (i.e., more stringent) thresholds

than those for VLBI in Eqs. (15.24) and (15.25). A curve for VLBI is shown in Fig. 15.2, using typical values for T_S. The harmful thresholds are approximately 40 dB less stringent than those for total-power systems.

15.4 INTERFERENCE FROM AIRBORNE AND SPACE TRANSMITTERS

In application of the F_h and S_h values obtained above, it should be remembered that they are derived with interference from stationary, ground-based transmitters in mind. It is often possible to make observations at sufficiently high angles of elevation that the antenna is pointed no closer than 19° to any such transmitter: 19° is the angle from the main beam at which most sidelobes fall below the isotropic level in the model in Fig. 15.1. Airborne and satellite transmitters present a special problem. Radio astronomy cannot share bands with space-to-earth (downlink) transmissions of satellites. However, because of the pressure for more spectrum for communications, allocations have been made in bands adjacent or close to those allocated to radio astronomy. Spurious emissions from satellite transmitters that fall outside the allocated band of the satellite arguably pose the most serious threat to radio astronomy. Motion of the transmitter across the sky is most likely to increase the fringe frequency at the correlator outputs of a synthesis array and thereby reduce the response to interference. On the other hand, these signals may be received in high-level sidelobes near the main beam. Transmitters on geostationary-orbit (GEO) satellites represent a particular hazard to radio astronomy because of their fixed locations at high elevation angles near the celestial equator. Interfering signals from a series of such satellites distributed along the geostationary orbit could result in a band of sky centered on the orbit in which high-sensitivity observations would be severely restricted.

Examples of spurious emissions that extend far outside the allocated band of the satellite system are described by Galt (1990) and Combrinck, West, and Gaylord (1994). In these cases the spurious emission resulted largely from the use of simple phase-shift keying for the modulation, and newer techniques [e.g., Gaussian-filtered minimum shift keying (GMSK)] provide much sharper reduction in spectral sidebands (Murota and Hirade 1981, Otter 1994). However, intermodulation products resulting from the nonlinearity of amplifiers carrying many communication channels remain a problem.

In some cases, operating requirements and limitations associated with space tend to make reduction of spurious emissions difficult. Some satellites use a large number of narrow beams to cover their area of operation, so that the same frequency channels can be used a corresponding number of times to accommodate a large number of customers. This requires phased-array antennas with many (of order one hundred or more) small radiating elements, each with its own power amplifier [see, e.g., Schuss et al. (1999)]. Because of power limitations from the solar cells, these amplifiers are operated at levels that maximize power efficiency but compromise linearity, resulting in spurious emissions from intermodulation

products. Filtering the individual outputs driving the radiating elements may be impractical because of weight limitations.

The recommended limits on spurious emissions (ITU-R 1997c) in effect require that, for space services, the power in spurious emissions measured in a 4-kHz band at the transmitter output should be no more than −43 dBW. Thus, for example, spurious emission at this level from a low-earth-orbit (LEO) satellite at 800 km height, and radiated from a sidelobe of 0 dBi gain, would produce a spurious spectral power flux density of -208 dBW m^{-2} Hz^{-1} at the earth's surface. This figure may be compared with the harmful interference thresholds for radio astronomy of -239 and -255 dBW m^{-2} Hz^{-1} for spectral line and continuum measurements, respectively, at 1.4 GHz. Although this very simple calculation considers only the worst-case situation, the differences of several tens of decibels show that the proposed limits do not protect radio astronomy. Thus radio astronomy is essentially regarded as a special problem to be studied on a case-by-case basis as new allocations are made and systems developed. The responsibility to ensure that such coordination takes place rests with radio astronomers.

APPENDIX 15.1 REGULATION OF THE RADIO SPECTRUM

Regulation of the usage of the radio spectrum is organized through the International Telecommunication Union (ITU), based in Geneva, which is a specialized agency of the United Nations Organization. Radio astronomy was first officially recognized as a radiocommunication service by the ITU in 1959. The Radiocommunication Sector of the ITU (ITU-R) was created in March 1993 and replaced the International Radio Consultative Committee (CCIR), an earlier entity within the ITU. A system of study groups within the ITU-R is responsible for technical matters. Study Group 7, entitled Science Services, includes radio astronomy, various aspects of space research, and standards for time and frequency. Study groups are subdivided into working parties that deal with specific areas. Their primary function is to study problems of current importance in frequency coordination, for example, specific cases of sharing of frequency bands between different services, and to produce documented Recommendations on the solutions. Decisions within the ITU are made largely by consensus. Recommendations must be approved by all of the radiocommunication study groups, and then effectively become part of the ITU Radio Regulations. For Recommendations specific to radio astronomy, see ITU-R (1997b) and other Recommendations in the RA series.

The ITU-R organizes meetings of study groups, working parties, and other groups required from time to time to deal with specific problems. It also organizes World Radiocommunication Conferences (WRCs) at intervals of two to three years, at which new spectrum allocations are made and the ITU Radio Regulations are revised as necessary. Administrations of many countries send delegations to WRCs, and the results of these conferences have the status of treaties. Participating countries can take exceptions to the international regulations so long as these do not impact spectrum usage in other countries. As a result, many administrations have their own system of radio regulations, based largely on the

ITU Radio Regulations, but with exceptions to accommodate their particular requirements. For further information see, for example, Pankonin and Price (1981), Thompson, Gergely, and Vanden Bout (1991), and ITU-R (1995).

BIBLIOGRAPHY

Crawford, D. L., Ed., *Light Pollution, Radio Interference, and Space Debris*, Astron. Soc. Pacific Conf. Series, **17**, ASP, San Francisco, CA, 1991.

ITU-R, *Handbook on Radio Astronomy*, International Telecommunication Union, Geneva, 1995 (or current revision).

Kahlmann, H. C., Interference: The Limits of Radio Astronomy, in *Review of Radio Science 1996–1999*, W. R. Stone, Ed., Oxford Univ. Press, Oxford, 1999, pp. 751–786.

Swenson, G. W. Jr., and A. R. Thompson, Radio Noise and Interference, in *Reference Data for Engineers: Radio, Electronics, Computer, and Communications*, Sams Indianapolis, 1993.

REFERENCES

Barnbaum, C. and R. F. Bradley, A New Approach to Interference Excision in Radio Astronomy: Real-Time Adaptive Cancellation, *Astron. J.*, **116**, 2598–2614, 1998.

Combrinck, W. L., M. E. West, and M. J. Gaylord, Coexisting with Glonass: Observing the 1612 MHz Hydroxyl Line, *Pub. Astron. Soc. Pacific*, **106**, 807–812, 1994.

Galt, J., Contamination from Satellites, *Nature*, **345**, 483, 1990.

ITU-R, *Handbook on Radio Astronomy*, International Telecommunication Union, Geneva, 1995 (or current revision).

ITU-R Recommendation SA.509-1, Generalized Space Research Earth Station Antenna Radiation Pattern for Use in Interference Calculations, Including Coordination Procedures, *ITU-R Recommendations, SA Series*, International Telecommunication Union, Geneva, 1997a (or current revision).

ITU-R Recommendation RA.769-1, Protection Criteria for Radioastronomical Measurements, *ITU-R Recommendations, RA Series*, International Telecommunication Union, Geneva, 1997b (or current revision).

ITU-R Recommendation SA.329-7, Spurious Emissions, *ITU-R Recommendations, SA Series*, International Telecommunication Union, Geneva, 1997c (or current revision).

Murota, K. and K. Hirade, GMSK Modulation for Digital Mobile Radio Telephony, *IEEE Trans. Commun.*, **COM-29**, 1044–1050, 1981.

Otter, M., *A Comparison of QPSK, OQPSK, BPSK and GMSK Modulation Schemes*, Report of the European Space Agency, European Space Operations Center, Darmstadt, Germany, June 1994.

Pankonin, V. and R. M. Price, Radio Astronomy and Spectrum Management: The Impact of WARC-79, *IEEE Trans. Electromagn. Compat.*, **EMC-23**, 308–317, 1981.

Schuss, J. J., J. Upton, B. Myers, T. Sikina, A. Rohwer, P. Makridakas, R. Francois, L. Wardle, and R. Smith, The IRIDIUM Main Mission Antenna Concept, *IEEE Trans. Antennas Propag.*, **AP-47**, 416–424, 1999.

Thompson, A. R., The Response of a Radio-Astronomy Synthesis Array to Interfering Signals, *IEEE Trans. Antennas Propag.*, **AP-30**, 450–456, 1982.

Thompson, A. R., T. E. Gergely, and P. Vanden Bout, Interference and Radioastronomy, *Physics Today*, 41–49, November 1991.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

16 Related Techniques

Concepts and techniques similar to those used in radio interferometry and synthesis mapping occur in various areas of astronomy. Here we introduce a few of them, including optical techniques, to leave the reader with a broader view. All of these subjects are described in detail elsewhere, so here the aim is mainly to outline the principles involved, and to make connections between them and the detailed material developed in previous chapters.

16.1 INTENSITY INTERFEROMETER

In long-baseline interferometry the intensity interferometer offers some technical simplifications that were mainly of importance in radio astronomy during the early development of the subject. As mentioned in Section 1.3 under *Early Measurements of Angular Width*, its practical applications in radio astronomy have been limited (Jennison and Das Gupta 1956, Carr et al. 1970, Dulk 1970) because, in comparison with a conventional interferometer, it requires a much larger signal-to-noise ratio in the receiving system, and only the modulus of the visibility function is measured. The intensity interferometer was devised by Hanbury Brown, who has described its development and application (Hanbury Brown 1974).

In the intensity interferometer, the signals from the antennas are amplified and then passed through square-law (power linear) detectors before being applied to a correlator, as shown in Fig. 16.1. As a result, the rms signal voltages at the correlator inputs are proportional to the powers delivered by the antennas, and therefore also proportional to the intensity of the signal. No fringes are formed because the phase of the radio frequency (RF) signals is lost in the detection, but the correlator output indicates the degree of correlation of the detected waveforms. Let the voltages at the detector inputs be V_1 and V_2. The outputs of the detectors are V_1^2 and V_2^2 and each consists of a dc component, which is removed by a filter, and a time-varying component, which goes to an input of the correlator. From the fourth-order moment relation [Eq. (6.36)] the correlator output is

$$\begin{aligned}\langle (V_1^2 - \langle V_1^2\rangle)(V_2^2 - \langle V_2^2\rangle)\rangle &= \langle V_1^2 V_2^2\rangle - \langle V_1^2\rangle\langle V_2^2\rangle \\ &= 2\langle V_1 V_2\rangle^2.\end{aligned} \tag{16.1}$$

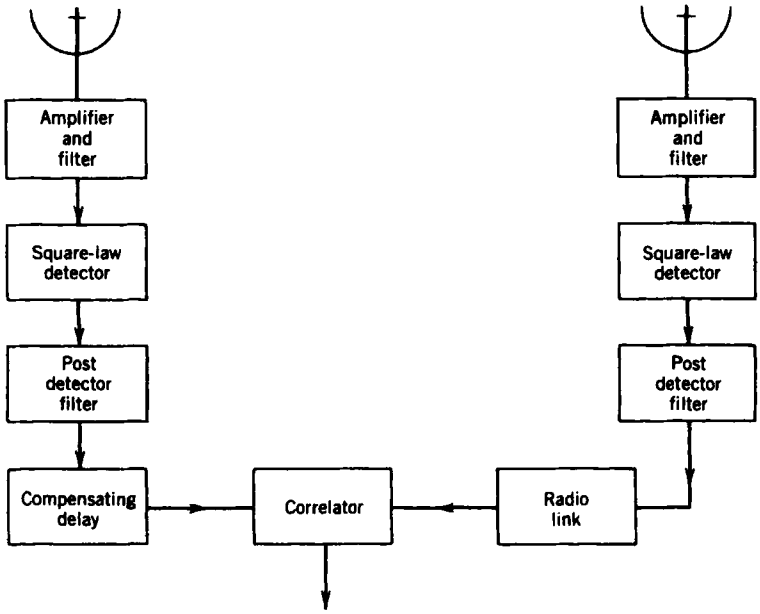


Figure 16.1 The intensity interferometer. The amplifier and filter block may also incorporate a local oscillator and mixer. The compensating delay equalizes the time delays of signals from the source to the correlator inputs. The post-detector filters remove dc and radio frequency components.

The correlator output is proportional to the square of the correlator output of a conventional interferometer, and measures the squared modulus of the visibility of a source under observation.

We now give an alternative derivation of the response, which provides a physical picture of how the signals from different parts of the source combine within the instrument. The source is represented as a one-dimensional intensity distribution in Fig. 16.2. We suppose that it can be considered as a linear distribution of many small regions, each of which is large enough to emit a signal with the characteristics of stationary random noise, but of angular width small compared with $1/u$, which defines the angular resolution of the interferometer. The source is assumed to be spatially incoherent, so the signals from different regions are uncorrelated. Consider two regions of the source, k and ℓ, at angular positions θ_k and θ_ℓ and subtending angles $d\theta_k$ and $d\theta_\ell$, as in Fig. 16.2. Each radiates a broad spectrum, but we first consider only the output resulting from a Fourier component at frequency ν_k from region k and similarly a component at ν_ℓ from region ℓ. Let $A_1(\theta)$ be the power reception pattern of the two antennas and $I_1(\theta)$ the intensity distribution of the source, these two functions being one-dimensional representations. Then the detector output of the first receiver is equal to

$$[V_k \cos 2\pi \nu_k t + V_\ell \cos(2\pi \nu_\ell t + \phi_1)]^2 , \tag{16.2}$$

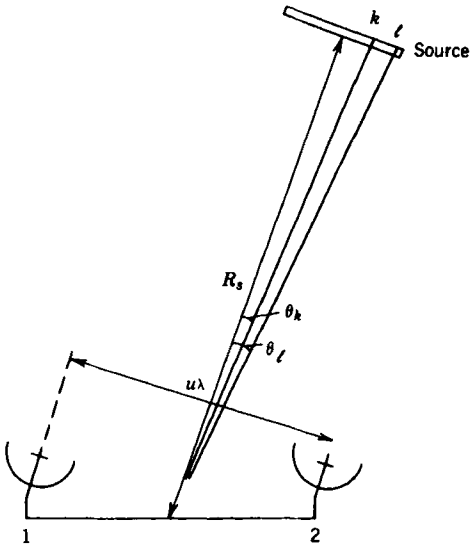


Figure 16.2 Distances and angles used in the discussion of the intensity interferometer.

where ϕ_1 is a phase term resulting from path-length differences, and the signal voltages V_k and V_ℓ are given by

$$V_k^2 = A_1(\theta_k) I_1(\theta_k)\, d\theta_k\, d\nu_k \tag{16.3}$$

$$V_\ell^2 = A_1(\theta_\ell) I_1(\theta_\ell)\, d\theta_\ell\, d\nu_\ell. \tag{16.4}$$

After expanding (16.2) and removing the dc and RF terms, we obtain for the detector output from receiver 1:

$$V_k V_\ell \cos\left[2\pi(\nu_k - \nu_\ell)t - \phi_1\right]. \tag{16.5}$$

Similarly, the detector output from receiver 2 is

$$V_k V_\ell \cos\left[2\pi(\nu_k - \nu_\ell)t - \phi_2\right]. \tag{16.6}$$

The correlator output is proportional to the time-averaged product of (16.5) and (16.6), that is, to

$$\langle A_1(\theta_k) A_1(\theta_\ell) I_1(\theta_k) I_1(\theta_\ell)\, d\theta_k\, d\theta_\ell\, d\nu_k\, d\nu_\ell \cos(\phi_1 - \phi_2)\rangle. \tag{16.7}$$

The change in the phase term with respect to frequency is small so long as the fractional bandwidth is much less than the ratio of the resolution to the field of view [see Eq. (6.69) and related discussion]. With this restriction expression (16.7) is effectively independent of the frequencies ν_k and ν_ℓ, so that if we

integrate it with respect to ν_k and ν_ℓ over a rectangular receiving passband of width $\Delta\nu$, $d\nu_k\, d\nu_\ell$ is replaced by $\Delta\nu^2$.

The phase angles ϕ_1 and ϕ_2 result from the path differences kk' and $\ell\ell'$ shown in Fig. 16.3. Note that ϕ_1 and ϕ_2 have opposite signs since the excess path length to antenna 1 is from point ℓ and that to antenna 2 is from point k. If R_s is the distance of the sources from the antennas, the distance $k\ell$ in the source is approximately equal to $R_s(\theta_k - \theta_\ell)$. The angle $\alpha_k + \alpha_\ell$ is approximately equal to $u\lambda/R_s$, since u represents the antenna spacing projected normal to the source and measured in wavelengths. The preceding approximations are accurate if α_k, α_ℓ, and the angle subtended by the source are all small. Thus the difference of the phase angles is

$$\begin{aligned}\phi_1 - \phi_2 &= 2\pi R_s(\theta_k - \theta_\ell)\frac{(\sin\alpha_k + \sin\alpha_\ell)}{\lambda} \\ &\simeq 2\pi u(\theta_k - \theta_\ell).\end{aligned} \tag{16.8}$$

From (16.7) the output of the correlator now becomes

$$\langle A_1(\theta_k)A_1(\theta_\ell)I_1(\theta_k)I_1(\theta_\ell)\,\Delta\nu^2\cos[2\pi u(\theta_k - \theta_\ell)]\,d\theta_k\,d\theta_\ell\rangle. \tag{16.9}$$

To obtain the output from all pairs of regions within the source expression (16.9) can, with the assumption of spatial incoherence, be integrated with respect to θ_k and θ_ℓ over the source, giving

$$\begin{aligned}&\left\langle\left[\Delta\nu\int A_1(\theta_k)I_1(\theta_k)\cos(2\pi u\theta_k)\,d\theta_k\right]\left[\Delta\nu\int A_1(\theta_\ell)I_1(\theta_\ell)\cos(2\pi u\theta_\ell)\,d\theta_\ell\right]\right. \\ &\quad\left.+\left[\Delta\nu\int A_1(\theta_k)I_1(\theta_k)\sin(2\pi u\theta_k)\,d\theta_k\right]\left[\Delta\nu\int A_1(\theta_\ell)I_1(\theta_\ell)\sin(2\pi u\theta_\ell)\,d\theta_\ell\right]\right\rangle \\ &= A_0^2\,\Delta\nu^2\left[\mathcal{V}_\mathrm{R}^2 + \mathcal{V}_\mathrm{I}^2\right] = A_0^2\,\Delta\nu^2\,|\mathcal{V}|^2,\end{aligned} \tag{16.10}$$

where we assume that the antenna response $A_1(\theta)$ has a constant value A_0 over the source, and the subscripts R and I denote the real and imaginary parts of the visibility. This result follows from the definition of visibility that is given for a

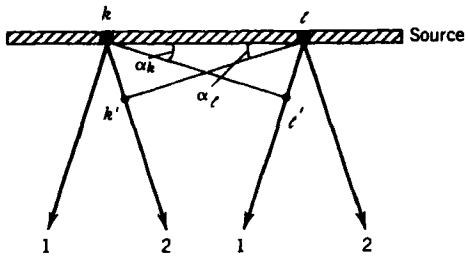


Figure 16.3 Relative delay paths kk' and $\ell\ell'$ from regions k and ℓ of the source for rays traveling in the directions of antennas 1 and 2.

two-dimensional source in Section 3.1. Thus, the correlator output is proportional to the square of the modulus of the complex visibility. For a more detailed discussion following the same approach, see Hanbury Brown and Twiss (1954). An analysis based on the mutual coherence of the radiation field is given by Bracewell (1958).

Some characteristics of the intensity interferometer offer advantages over the conventional interferometer. The intensity interferometer is much less sensitive to atmospheric phase fluctuations, because each signal component at the correlator input is generated as the difference between two radio frequency components that have followed almost the same path through the atmosphere. The phase fluctuations in the difference-frequency components at the detectors are less than those in the radio frequency signals by the ratio of the difference frequency to the radio frequency, which may be of order 10^{-5}. In the conventional interferometer such phase fluctuations can make the amplitude, as well as the phase, of the visibility difficult to measure. Similarly, fluctuations in the phases of the local oscillators, in the two receivers do not contribute to the phases of the difference-frequency components. Thus, it is not necessary to synchronize the local oscillators, or even to use high-stability frequency standards as in VLBI. These advantages were helpful, although by no means essential, in the early radio implementation of the intensity interferometer. Had the diameters of the sources under investigation then been of order of arcseconds, rather than arcminutes, the characteristics of the intensity interferometer would have played a more essential role.

The serious disadvantage of the intensity interferometer is its relative lack of sensitivity. Because of the action of the detectors in the receivers, the ratio of the signal power to the noise power at the correlator inputs is proportional to the square of the corresponding ratio in the RF (predetector) stages [see Eq. (9.73)], the exact value being dependent on the bandwidths of these and the postdetector stages (Hanbury Brown and Twiss 1954). In a conventional interferometer, it is possible to detect signals that are ~60 dB below the noise at the correlator inputs. In the intensity interferometer, a similar signal-to-noise ratio at the correlator output would require signal-to-noise ratios greater by ~30 dB in the RF stages. This effect, together with the lack of sensitivity to the visibility phase, has greatly restricted the radio usage of the intensity interferometer. Intensity interferometry played a similar role in the early days of optical interferometry (see Section 16.4 under *Optical Intensity Interferometer*), before the development of the modern Michelson interferometer.

16.2 LUNAR OCCULTATION OBSERVATIONS

Measurement of the light intensity from a star as a function of time during occultation by the moon was suggested by MacMahon (1909) as a means of determining the star's size and position. His analysis, which was based on a simple consideration of geometric optics, was criticized by Eddington (1909) who stated that diffraction effects would mask the detail at the angular scale of the star. Eddington's paper probably discouraged observations of lunar occultations for some

time. The first occultation measurements were reported 30 years later by Whitford (1939), who observed the stars β Capricorni and ν Aquarii and obtained clear diffraction patterns.

What was not realized by Eddington and others at the time was that although the temporal response to an occultation is not a simple step function, as it would be for the case of geometrical optics and a point source, the Fourier transform of the point-source response, which represents the sensitivity to spatial frequency on the sky, has the same amplitude as that of a step function and differs only in the phase. Hence, the lunar occultation is sensitive to all Fourier components, and there is no intrinsic limit to the resolution that can be obtained, except for that imposed by the finite signal-to-noise ratio. This equality of the amplitudes was recognized by Scheuer (1962), who devised a method of deriving the one-dimensional intensity distribution I_1 from the occultation curve. By that time, the concept of spatial frequency had become widely understood through application to radio interferometry. Since, in lunar occultations, the diffraction occurs outside the earth's atmosphere, the high angular resolution is not corrupted significantly by atmospheric effects, as it is in the case of ground based interferometry. Furthermore, the only dependence of the obtainable resolution on the telescope size results from the signal-to-noise ratio. An early radio application of the technique was the measurement of the position and size of 3C273 by Hazard, Mackey, and Shimmins (1963), which led to the identification of quasars. As mentioned in Section 12.1 under *Requirements for Astrometry*, this position measurement was used for many years as the right ascension reference for VLBI position catalogs. Radio occultation measurements have been most important at meter wavelengths, since at shorter wavelengths the high thermal flux density from the moon presents a difficulty. Radio observations have been largely superseded by interferometry, but lunar occultations are still used at optical and infrared wavelengths.

Figure 16.4 shows the geometrical situation and the form of an occultation record. The departure of the moon's limb from a straight edge, as a result of curvature and roughness, is small compared with the size of the first Fresnel zone at radio frequencies. Thus the point-source response is the well known diffraction pattern of a straight edge, which is derived in most texts on physical optics. The main change in the received power in Fig. 16.4b corresponds to the covering or uncovering of the first Fresnel zone by the moon, and the oscillations result from higher-order zones. The critical scale is the size of the first Fresnel zone, $\sqrt{(\lambda R_m/2)}$, where $R_m \simeq 3.84 \times 10^5$ km is the earth–moon distance. This corresponds to 4400 m at 10 cm wavelength and 10 m at 0.5 μm, or 2.3 arcsec and 5 mas, respectively, in angle as seen from the earth. The maximum velocity of the occulting edge of the moon is approximately 1 km s^{-1}, but the effective velocity depends on the position of the occultation on the moon's limb, and we use 0.6 km s^{-1} as a typical figure. Thus, the coverage time of the first Fresnel zone, which determines the characteristic fall time and oscillation period, is typically about 7 s at a wavelength of 10 cm and 16 ms at 0.5 μm.

In the case of the hypothetical geometrical optics occultation, the observed curve would be the integral of I_1 as a function of θ, the angle between the source and the moon's limb as measured in Fig. 16.4a. Then I_1 could be obtained by

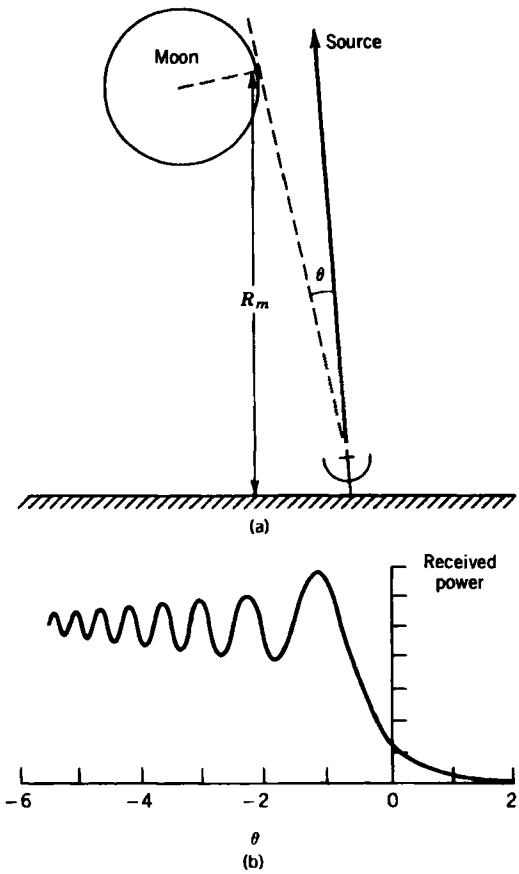


Figure 16.4 Occultation of a radio source by the moon: (a) the geometrical situation, in which θ is measured clockwise from the direction of the source, and is negative as shown; (b) the occultation curve for a point source, which is proportional to $\mathcal{P}(\theta)$. The units of θ on the abscissa are equal to $\sqrt{\lambda/2R_m}$, where λ is the wavelength and R_m the moon's distance.

differentiation. In the actual case the observed occultation curve $\mathcal{G}(\theta)$ is equal to convolution of $I_1(\theta)$ with the point-source diffraction pattern of the moon's limb $\mathcal{P}(\theta)$. This convolution is $I_1(\theta) * \mathcal{P}(\theta)$. Differentiation with respect to θ yields

$$\mathcal{G}'(\theta) = I_1(\theta) * \mathcal{P}'(\theta), \tag{16.11}$$

where the primes indicate derivatives. Fourier transformation of the two sides of Eq. (16.11) gives

$$\overline{\mathcal{G}}'(u) = \overline{I}_1(u)\,\overline{\mathcal{P}}'(u), \tag{16.12}$$

where the bar indicates the Fourier transform, the prime indicates a derivative in the θ domain, and u is the conjugate variable of θ.

Now in the geometrical-optics case $\mathcal{P}(\theta)$, would be a step function, and thus $\mathcal{P}'(\theta)$ would be a delta function for which the Fourier transform is a constant. For the diffraction-limited case the function $\overline{\mathcal{P}}(u)$ (adapted from Cohen 1969) is given by

$$\overline{\mathcal{P}}(u) = \frac{j}{u} \exp\left[j2\pi\theta_F^2 u^2 \mathrm{sgn} u\right], \tag{16.13}$$

where θ_F is the angular size of the first Fresnel zone, $\sqrt{\lambda/2R_m}$, and sgn is the sign function, which takes values ± 1 to indicate the sign of u. It follows from the derivative theorem of Fourier transforms that $\overline{\mathcal{P}}'(u) = j2\pi u\overline{\mathcal{P}}(u)$, which has a constant amplitude with no zeros and can be divided out from Eq. (16.12). Thus $I_1(\theta)$ is equal to $\mathcal{G}'(\theta)$ convolved with a function whose Fourier transform is $1/\overline{\mathcal{P}}'(u)$. Scheuer (1962) shows that this last function is proportional to $\mathcal{P}'(-\theta)$, which can be used as a restoring function as follows:

$$\begin{aligned} I_1(\theta) &= \mathcal{G}'(\theta) * \mathcal{P}'(-\theta) \\ &= \mathcal{G}(\theta) * \mathcal{P}''(-\theta). \end{aligned} \tag{16.14}$$

The second form on the right-hand side is more useful since it avoids the practical difficulty of differentiating a noisy occultation curve. In principle, this restoration provides I_1 without limit on the angular resolution, in contrast to the performance of an array. Remember, however, that the amplitude of the spatial frequency sensitivity of the occultation curve, which is given by Eq. (16.13), is proportional to $1/u$. Thus in the restoration in Eq. (16.14) the amplitudes of the Fourier components, which also include the noise, are increased in proportion to u. The increase of the noise sets a limit to the useful resolution. This limit can be conveniently introduced by replacing $\mathcal{P}''(\theta)$ in Eq. (16.14) by $\mathcal{P}''(\theta)$ convolved with a Gaussian function of θ with a resolution $\Delta\theta$. One then derives I_1 as it would be observed with a beam of the same Gaussian shape. In practice, the introduction of the Gaussian function is essential to the method, since it ensures the convergence of the convolution integral in Eq. (16.14). The optimum choice of $\Delta\theta$ depends on the signal-to-noise ratio. Examples of restoring functions for various resolutions can be found in von Hoerner (1964).

The discussion above follows the classical approach to reduction of moon-occultation observations, which developed from the geometrical optics analogy. One can envisage the reduction more succinctly as taking the Fourier transform of the occultation curve, dividing by $\overline{\mathcal{P}}(u)$ (with suitable weighting to control the increase of the noise), and retransforming to the θ domain. This process is mathematically equivalent to that in Eq. (16.14).

An estimate of the noise-imposed limit on the angular resolution can be obtained using the geometrical optics model, since the signal-to-noise ratio of the Fourier components is the same as for the actual point-source response. Consider the region of an occultation curve (see Fig. 16.4b) in which the main change in the received power occurs, and let τ be a time interval in which the change in the record level is equal to the rms noise. Then if v_m is the rate of angular motion of the moon's limb over the radio source, the obtainable angular resolution is

approximately

$$\Delta\theta = v_{\mathrm{m}}\tau. \tag{16.15}$$

During the interval τ, the flux density at the antenna changes by ΔS. Let θ_s be the width of the main structure of the source in a direction normal to the moon's limb, and let S be the total flux density of the source. Then for simple source structures the average intensity is approximately S/θ_s^2, and the change in solid angle of the covered part of the source in time τ is $\theta_s\,\Delta\theta$. Thus we have

$$\frac{\Delta\theta}{\theta_s} \simeq \frac{\Delta S}{S}. \tag{16.16}$$

The signal-to-noise ratio at the receiver output for a component of flux density ΔS is

$$\mathcal{R}_{\mathrm{sn}} = \frac{A\Delta S\sqrt{\Delta\nu\tau}}{2kT_S}, \tag{16.17}$$

where A is the collecting area of the antenna, $\Delta\nu$ and T_S are the bandwidth and system temperature of the receiving system, and k is Boltzmann's constant. Note that the thermal contribution from the moon can contribute substantially to T_S. The conditions that we are considering correspond to $\mathcal{R}_{\mathrm{sn}} \simeq 1$, and from Eqs. (16.15)–(16.17) we obtain

$$\Delta\theta = \left(\frac{2kT_S\theta_s}{AS}\right)^{2/3}\left(\frac{v_{\mathrm{m}}}{\Delta\nu}\right)^{1/3}. \tag{16.18}$$

Note that frequency (or wavelength) does not enter directly into Eq. (16.18), but the values of several parameters, for example, S, $\Delta\nu$, and T_S, depend upon the observing frequency. As an example, consider an observation at a frequency in the 100–300 MHz range for which we use $A = 2000\ \mathrm{m}^2$, $T_S = 200$ K, and $\Delta\nu = 2$ MHz. For a fairly weak radio source we take $S = 10^{-26}\ \mathrm{W\ m^{-2}\ Hz^{-1}}$ (1 Jy) and $\theta_s = 5$ arcsec. v_{m} is typically 0.3 arcsec s^{-1}. With these values, Eq. (16.18) gives $\Delta\theta = 0.7$ arcsec. Although Eq. (16.18) is derived using a geometrical optics approach, this does not limit its applicability. For an observed occultation curve, the equivalent curve for geometrical optics can be obtained by adjustment of the phases of the Fourier components, which does not affect the signal-to-noise ratio.

The bandwidth of the receiving system has the effect of smearing out angular detail in an occultation observation in a manner similar to that for an array. Thus, since the signal-to-noise ratio increases with bandwidth, for any observation there exists a bandwidth with which the sensitivity to fine angular structure is maximized. This bandwidth is approximately $\nu^2\Delta\theta^2 R_m/c$, which can be derived from the requirement that the phase term in Eq. (16.13) not change significantly over the bandwidth. This result can be compared to the bandwidth limitation for an array [given by Eq. (6.70)] by noting that a measurement by lunar occultation with resolution $\Delta\theta$ involves examination of the wavefront, at the distance of the moon, on a linear scale of $\lambda/\Delta\theta$. Such an interval subtends an angle $\lambda/\Delta\theta R_m$

at the earth. Further discussion of such details, and of the practical implementation of Scheuer's restoration technique, is given by von Hoerner (1964), Cohen (1969), and Hazard (1976). Note that a source may undergo a number of occultations within a period of a few months, with the moon's limb traversing the source at different position angles. If a sufficient range of position angles is observed, the one-dimensional intensity distributions can be combined to obtain a two-dimensional image of the source [see, e.g., Taylor and De Jong (1968)].

The method of lunar occultation has been widely used in optical and infrared astronomy to measure the size and the limb darkening of stars, and the separation of close binary stars. Consistency of the results with those of optical interferometry proves that the lunar occultation method is not corrupted by variations in the lunar topography, which can be expected to become important when the size of the variations is comparable to the Fresnel scale. Angular sizes have been routinely measured down to about 1 mas. The analysis of stellar occultation curves is usually done by fitting parameterized models, rather than the reconstruction methods used in radio observations described above. A review of special considerations for lunar occultation observations at optical and infrared wavelengths can be found in Richichi (1994). Extensive measurement of stellar diameters [see, e.g., White and Feierman (1987)] and binary star separations [see, e.g., Evans et al. (1985)] have been made. Other applications include the measurement of subarcsecond dust shells surrounding Wolf-Rayet stars [see, e.g., Ragland and Richichi (1999)].

16.3 MEASUREMENTS ON ANTENNAS

Measurement of the electric field distribution over the aperture of an antenna is an important step in optimizing the aperture efficiency, especially in the case of a reflector antenna for which such results indicate the accuracy of the surface adjustment. The Fourier transform relationship between the voltage response pattern of an antenna and the field distribution in the aperture has been derived in Section 14.1 under *Diffraction at an Aperture and the Response of an Antenna*. If x and y are axes in the aperture plane, the field distribution $\mathcal{E}(x_\lambda, y_\lambda)$ is the Fourier transform of the far-field voltage radiation (reception) pattern $V_A(l, m)$ (see Section 3.3 under *Antennas*), where l and m are here the direction cosines measured with respect to the x and y axes and the subscript λ indicates measurement in wavelengths. Thus

$$V_A(l, m) \propto \iint_{-\infty}^{\infty} \mathcal{E}(x_\lambda, y_\lambda) e^{j2\pi(x_\lambda l + y_\lambda m)}\, dx_\lambda\, dy_\lambda. \tag{16.19}$$

Direct measurement of $\mathcal{E}$ can be made by moving a probe across the aperture plane, but care must be taken to avoid disturbing the field, which may be difficult. Such a technique is useful for characterizing horn antennas for millimeter wavelengths (Chen et al. 1998). However, in many applications, especially for large antennas on fully steerable mounts, it is easier to measure V_A. It is neces-

sary to measure both the amplitude and phase of $V_A(l, m)$ in order to perform the Fourier transform for $\mathcal{E}(x_\lambda, y_\lambda)$. To accomplish this, the beam of the antenna under test can be scanned over the direction of a distant transmitter, and a second, nonscanning antenna can be used to receive a phase reference signal. The function $V_A(l, m)$ is obtained from the product of the signals from the two antennas. This technique resembles the use of a reference beam in optical holography, and antenna measurements of this type have been described as holographic (Napier and Bates 1973, Bennett et al. 1976).

The holographic technique is readily implemented for measurements of antennas in interferometers and synthesis arrays. If the instrumental parameters (baselines, etc.) and the source position are accurately known, and the phase fluctuations introduced by the atmosphere are negligible, then for an unresolved source, calibrated visibility values will have a real part corresponding to the flux density of the source and an imaginary part equal to zero (except for the noise). If one antenna of a correlated pair is scanned over the source, while the other antenna continues to track the source, the corresponding visibility values will be proportional to the amplitude and phase of $V_A(l, m)$ for the scanning antenna. Measurement of synthesis array antennas as outlined above was first described by Scott and Ryle (1977), whose analysis, and that of D'Addario (1982), we largely follow below.

It is convenient to visualise the data in the aperture plane $\mathcal{E}(x_\lambda, y_\lambda)$ and in the sky plane $V_A(l, m)$ as discrete measurements at grid points in two $N \times N$ arrays to be used in the discrete Fourier transformation. For simplicity, consider a square antenna aperture with dimensions $d_\lambda \times d_\lambda$. Since $\mathcal{E}(x_\lambda, y_\lambda)$ is zero outside a range $\pm d_\lambda/2$, the sampling theorem of Fourier transforms indicates that the response must be sampled at intervals in (l, m) no greater than $1/d_\lambda$. [This interval is twice the sampling interval for the power beam because the power beam is the Fourier transform of the *autocorrelation function* of $\mathcal{E}(x_\lambda, y_\lambda)$.] If the $V_A(l, m)$ samples are spaced at $1/d_\lambda$, the aperture data just fill the $\mathcal{E}(x_\lambda, y_\lambda)$ array. The spacing of the measurements in the aperture is d_λ/N. Therefore N is usually chosen so that the sample interval provides several measurements on each surface panel. In the (l, m) plane the range of angles over which the scanning takes place is N times the pointing interval, that is, N/d_λ. This scan range is approximately N beamwidths. The procedure is to scan with the antenna under test in N^2 discrete pointing steps and thereby obtain the $V_A(l, m)$ data to fill the sky-plane array.

As a measure of the strength of the signal, let $\mathcal{R}_{\rm sn}$ be the signal-to-noise ratio obtained in time τ_a with the beams of both antennas pointed directly at the source. Now suppose that the (x_λ, y_λ) aperture plane is divided into square cells (as in Fig. 5.3) with sides d_λ/N centered on the measurement points. Consider the contribution to the correlator output of the signal from one such aperture cell, of area $(d_\lambda/N)^2$, in the antenna under test. The effective beamwidth of such an aperture cell is N times the antenna beamwidth, that is, approximately the total scan width required. Such an area contributes a fraction $1/N^2$ to the signal at the correlator output, so relative to the noise at the correlator output the component resulting from one aperture cell is $\mathcal{R}_{\rm sn}/N^2$ in time τ_a, or $\mathcal{R}_{\rm sn}/N$ in time $N^2\tau_a$, which is the total measurement time. The accuracy of the phase measurement for the signal component from one aperture cell, $\delta\phi$, is the reciprocal of $\sqrt{2}$ times

the corresponding signal-to-noise ratio, that is, $N/(\sqrt{2}\mathcal{R}_{sn})$. The factor $\sqrt{2}$ is introduced because only the component of the system noise that is normal to the signal (visibility) vector introduces error in the phase measurement; see Fig. 6.8. Now a displacement ε in the surface of the aperture cell causes a change of phase $4\pi\varepsilon/\lambda$ in the reflected signal. Thus an uncertainty $\delta\phi$ in the phase of this signal component results in an uncertainty in ε of $\delta\varepsilon = \lambda\delta\phi/(4\pi) = \lambda N/(4\sqrt{2}\pi\mathcal{R}_{sn})$. From the accuracy $\delta\varepsilon$ desired for the surface measurement, we determine that the signal strength should be such that the signal-to-noise ratio in time τ_a, with both beams on source, is

$$\mathcal{R}_{sn} = \frac{N\lambda}{4\sqrt{2}\pi\delta\varepsilon}. \tag{16.20}$$

Having determined $\mathcal{R}_{sn}$, we can use Eqs. (6.48) and (6.49) to obtain values of antenna temperature or flux density (W m^{-2} Hz^{-1}) for the signal. If the two antennas used are not of the same size, then in Eqs. (6.48) and (6.49) A, T_A, and T_S are replaced by the geometric means of the corresponding quantities. Several simplifying approximations have been made. The statement that one aperture cell contributes $1/N^2$ of the antenna output implies the assumption that the field strength is uniform over the aperture. If the aperture illumination is tapered, a higher value of $\mathcal{R}_{sn}$ will be required to maintain the accuracy at the outer edges. Consideration of a square antenna overestimates $\mathcal{R}_{sn}$ for a circular aperture of diameter d_λ by $4/\pi$. The situation can be significantly different when the signal used in the holography measurement is a cw (continuous wave) tone, often from a satellite. The received signal power P can be large compared with the receiver noise $kT_R\,\Delta\nu$ (D'Addario 1982). In that case the noise in the correlator output is dominated by the cross products formed by the signal and the receiver noise voltages. The resulting signal-to-noise ratio in time τ is $\sqrt{P\,\Delta\nu\tau/(kT_R\,\Delta\nu)}$, which is independent of the receiver bandwidth.

An example of holographic measurements on an antenna of a submillimeter-wavelength synthesis array is shown in Fig. 16.5. Some practical points are listed below.

- The source used in a holographic measurement is ideally strong enough to allow a high signal-to-noise ratio to be obtained. Usually either a signal from a transmitter on a satellite or a cosmic maser source is used. Morris et al. (1988) describe measurements on the 30-m antenna at Pico de Veleta in which a measurement accuracy (repeatability) of 25 μm was achieved using the 22.235 GHz water maser in Orion. For holography with interferometer elements, sources that are partially resolved can be used (Serabyn, Phillips, and Masson 1991).
- If the test antenna is on an altazimuth mount, the beam will rotate relative to the sky as the observation proceeds. In determining the pointing directions, the (l, m) axes of the sky plane should remain aligned with the local horizontal and vertical directions. If the antenna is on an equatorial mount, the (l, m) axes should be the directions of east and north on the sky plane [i.e., the usual (l, m) definition].

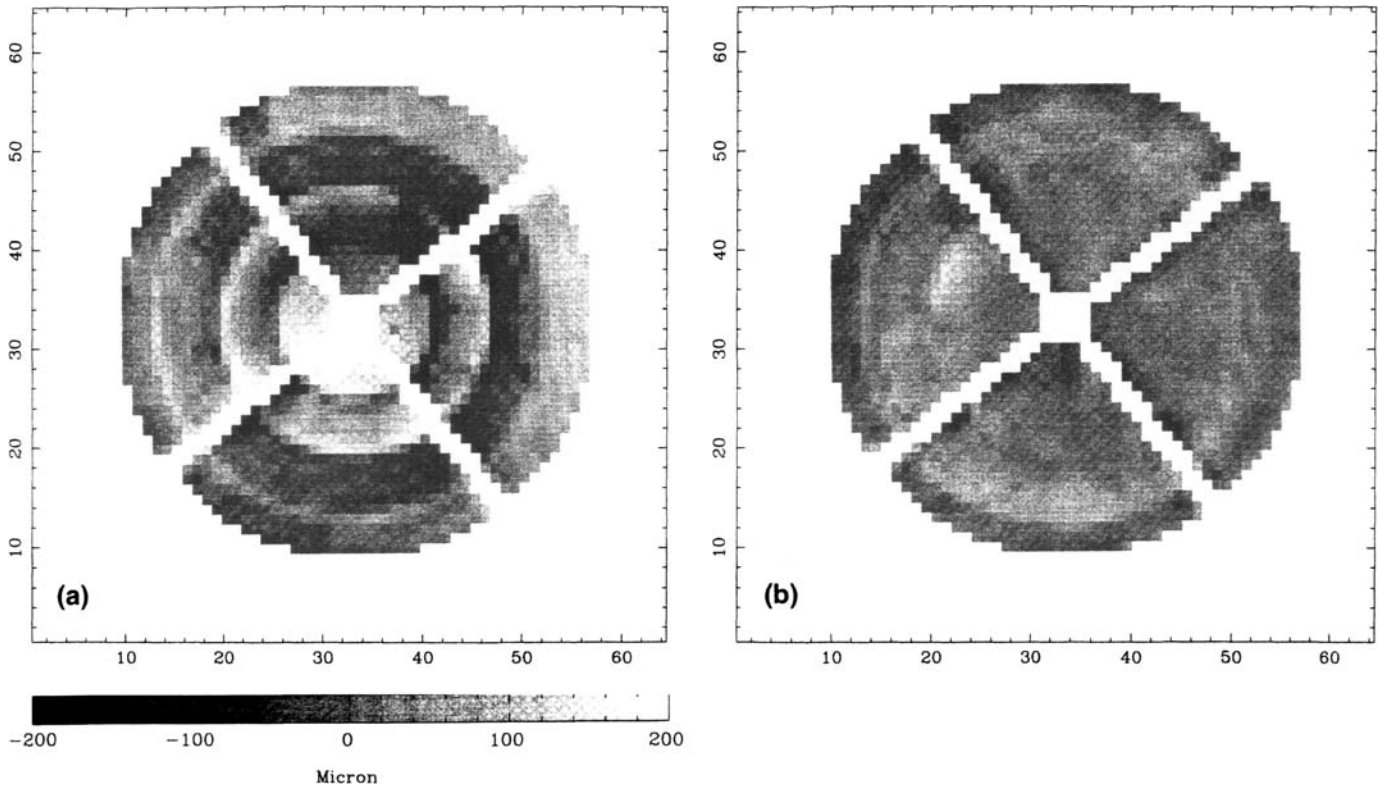


Figure 16.5 (a) Surface deviations of a 6-m SMA reflector measured by holography at 92 GHz with a cw beacon located at a distance of 250 m. The image has 64 × 64 pixels and a resolution of about 9 cm. The pixels outside the dish, and under the quadrupod legs and secondary mirror, have been blanked. The rms deviation is 73 μm. There are 64 panels set in four rows. A systematic setting error in the radial direction can be seen. (b) The result after one iteration of setting the panels. The rms surface deviation is 30 μm. From measurements by S. K. Tirupati.

- If the source is strongly linearly polarized and the antennas are on altazimuth mounts, it may be necessary to compensate for rotation of the beam. This is possible if the antennas receive on two orthogonal polarizations.
- When using two separate antennas, differences in the signal paths resulting from tropospheric irregularities can cause phase errors. It may be necessary to make periodic recordings with both beams centered on the source to determine the magnitude of such effects. In the case of measurement on a single large antenna, a small antenna mounted on the feed support structure of the large one, and pointing in the same direction as the large antenna's beam, is sometimes used to provide the on-source reference signal. Tropospheric effects on the phase should then cancel.
- An antenna may be rotated (through a limited angular range) about any axis through its phase center without varying the phase of the received signal. The phase center of a parabolic reflector lies on the axis of the paraboloid, and is roughly near the mid-point between the vertex and the aperture plane.* In the scanning, the maximum angle through which the antenna is turned from the on-source direction is $N/(2d_\lambda)$. If the axis about which it is turned is a distance r from the phase center, the phase path length to the antenna will be increased by $r[1 - \cos(N/2d_\lambda)]$. If this distance is a significant fraction of a wavelength, a phase correction must be applied to the signal at the correlator output.
- For an antenna in a radome, structural members of which can cause scattering of the incident radiation, corrections are necessary. Rogers et al. (1993) describe such corrections for measurements on the Haystack 37-m antenna.
- In measurements on the antennas of a correlator array in which the number of antennas n_a is large, a possible procedure would be to use one antenna to track the source and provide the reference signal, and scan all the others over the source. However, a better procedure would be to use $n_a/2$ antennas to track the source while the other $n_a/2$ antennas are scanned. The averaging time would be half that of the first procedure to allow the roles of the two sets of antennas to be interchanged at the midpoint of the observation. However, there would be $n_a/2$ different measurements for each antenna, so compared with the first procedure, the sensitivity would be increased by a factor $\sqrt{n_a/4}$. Also, cross-correlation of the signals from the tracking antennas would provide information about the phase stability of the atmosphere, which would be useful in interpreting the measurements.

*Consider transmission from an antenna in which the parabolic surface is formed by rotation of the parabola $x = ay^2$ around the x axis. Radiation from a ring-shaped element of the surface between the planes $x = x'$ and $x = x' + dx$ has an effective phase center on the x axis at x'. The area of such an element projected onto the aperture plane (i.e., normal to the x axis) is independent of x'. If the aperture illumination is uniform, each surface element between planes normal to the x axis and separated by the same increment makes an equal contribution to the electric vector in the far field. Thus the effective phase center of the total radiation should be on the x axis, midway between the vertex and the aperture plane. Note that this is an approximate analysis based on geometrical optics.

A method that requires only measurement of the amplitude of the far-field pattern has been developed by Morris (1985). In such a procedure the reference antenna is not required. The method is based on the Misell algorithm (Misell 1973), and the procedure can be outlined as follows. Input requirements are an initial "first guess" model of the amplitude and phase of the field distribution across the antenna aperture, and two measurements of the far-field amplitude pattern, one with the antenna correctly focused, and the other with the antenna defocused sufficiently to produce phase errors of a few radians at the antenna edge. The model aperture distribution is used to calculate the in-focus far-field pattern in amplitude and phase, and the calculated in-focus amplitude is replaced by the measured amplitude. The measured in-focus amplitude and the calculated phase are then used to calculate the corresponding aperture amplitude and phase, which then become the new aperture model. This new model is then used to calculate the defocused far-field pattern. In calculating the defocused pattern, it is assumed that, in the aperture, the defocusing affects only the phase, and that it introduces a component that varies in the aperture as the radius squared. The calculated defocused amplitude pattern is then replaced by the measured defocused pattern, and the corresponding in-focus aperture distribution is calculated and becomes the new model. In the continuing iterations, the in-phase and defocused amplitudes are calculated alternately. After each calculation the amplitude pattern is replaced by the corresponding measured pattern, and the result is used to upgrade the model. The required solution to which the procedure should converge is a model that fits both the in-focus and defocused responses. This technique requires a higher signal-to-noise ratio than when phase measurements are made. For measurements near nulls in the beam, the required signal-to-noise ratio is approximately equal to the square of that when the phase is measured (Morris 1985).

A holographic method involving only one antenna, suitable for a large submillimeter wavelength telescope, is described by Serabyn, Phillips, and Masson (1991). Measurements are made in the focal plane using a shearing interferometer, an adaptation of a technique used for optical instruments.

16.4 OPTICAL INTERFEROMETRY

The principles of optical interferometry are essentially identical to those at radio frequencies, but accurate measurements are more difficult to make. One difficulty arises because irregularities in the atmosphere introduce variations in the effective path length that are large compared with the wavelength, and thus cause the phase to vary irregularly by many rotations. Also, obtaining the mechanical stability of an instrument required to obtain fringes at a wavelength of order 500 nm presents a formidable problem. Thus it is difficult to calibrate the instrumental phase response, and in many cases only the visibility amplitude is measured. However, the practicality of synthesis imaging in the optical spectrum has been demonstrated by Haniff et al. (1987) and Baldwin et al. (1996), using phase closure techniques; see Section 10.3. In the absence of visibility phase, the amplitude data can be interpreted in terms of models, or the autocorrelation of the intensity

distribution as explained in Section 11.4 under *Mapping with Visibility Amplitude Data Only*. Techniques for two-dimensional reconstruction without phase data [see, e.g., Bates (1984)] are also applicable. Optical interferometry is an active and growing field, and here we attempt only to give an overview of some basic principles. A general review is given by Shao and Colavita (1992) and a detailed review of the theory is given by Tango and Twiss (1980). A very useful collection of some of the most important publications in optical interferometry has been compiled by Lawson (1997): see bibliography.

Before discussing instruments, we briefly review some relevant atmospheric parameters. The irregularities in the atmosphere give rise to random variations in the refractive index over a large range of linear scales. For any particular wavelength, there exists a scale size over which portions of a wavefront remain substantially plane compared with the wavelength, that is, atmospheric phase variations are small compared with 2π. This scale size is represented by a parameter, the Fried length r_0 (Fried 1966); see the discussion following Eq. (13.102). The Fried length is equal to $3.2d_0$, where d_0 is the spacing between paths through the atmosphere for which the rms phase difference is one radian; see Eq. (13.102). Regions for which the uniformity of the phase path lies within this range are sometimes referred to as seeing cells. The scale size r_0 and the height at which the dominant irregularities occur define an isoplanatic angle (or isoplanatic patch size), that is, an angular range of the sky within which the incoming wavefronts from different points encounter similar phase shifts. Within an isoplanatic patch the point-spread function remains constant, so the convolution relationship between source and image holds. Typical figures for the 50th percentile value of r_0, which scales as $\lambda^{6/5}$ [see Eq. (13.102)], and the isoplanatic angle, are given in Table 16.1. Also included for comparison are the corresponding values of the diffraction-limited resolution of a telescope of 1-m aperture. Optical interferometers provide a powerful means of studying the structure functions of the atmosphere at infrared and optical wavelengths; see, for example, Bester et al. (1992) and Davis et al. (1995). Note that techniques involving correction of atmospheric distortion of the wavefront by means of the telescope hardware are referred to as adaptive optics [see, e.g., Roggemann, Welch, and Fugate (1997), Milonni (1999)].

TABLE 16.1 Atmospheric and Instrumental Parameters at Visible and Infrared Wavelengths

Wavelength (μm)	r_0 (m)	Isoplanatic Angle at Zenith	Resolution of 1-m Diameter Aperture	Atmospheric Resolution (λ/r_0)
0.5 (visible)	0.14	5.5″	0.13″	0.70″
2.2 (infrared)	0.83	33″	0.55″	0.55″
20 (infrared)	11.7	8′	5.0″	0.35″

Source: Woolf (1982)

Modern Michelson Interferometer

The original Michelson instrument was briefly discussed in Section 1.3, and it was pointed out that the instability of the fringes was a limiting factor in estimation of the visibility. The timescale of the atmospheric fluctuations is of the order of 10 ms, which can be accommodated by using an electronic system for control and measurement of the fringes. A simplified diagram illustrating the basic features of a modern Michelson type of interferometer is shown in Fig. 16.6. The two mirrors S are mounted as siderostats and track the optical source under study. The positions of the retroreflectors R are continuously adjusted to equalize the lengths of the paths from the source to the combination point B. This delay compensation is usually implemented in evacuated tubes because the geometric delay of the interferometer largely occurs above the atmosphere. If air delay lines were used, a separate mechanism would be needed to compensate for the dispersive component of the delay, which is difficult to implement in wide bandwidth systems [see, e.g., Benson et al. (1997)]. The siderostats are mounted on stable foundations, and the rest of the system is usually mounted on a system of optical benches within a controlled environment. The apertures of the interferometer, determined by the mirrors S, are made no larger than the scale size r_0. Thus the wavefront across the mirror remains essentially plane, and the effect of the irregularities is to produce a variation in the angle of arrival of the wavefront. The variation cannot be tolerated since the angles of the beams at the combination point B must be correct to within 1 arcsec. To mitigate this effect, the polarizing beamsplitter cubes P reflect light to quadrant detectors Q, which produce a voltage proportional to any displacement of the angle of the light beam. These voltages are then used to control the tilt angles of the mirrors T, to compensate for the wavefront variation. A servo loop with bandwidth ~1 kHz is required to follow the fastest atmospheric

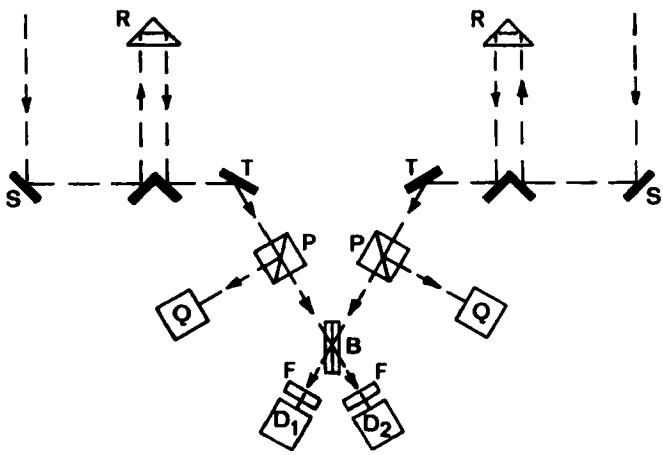


Figure 16.6 Basic features of a modern, direct detection implementation of the Michelson interferometer. The broken line represents the light path from a star. From Davis and Tango, *Proc. Astron. Soc. Australia*, **6**, 34–38, 1985.

effects. The filters F define the operating wavelength. The two detectors D_1 and D_2 respond to points on the fringe pattern spaced by one-quarter of a fringe cycle, and their outputs provide a measure of the instantaneous amplitude and phase of the fringes. This method is described, for example, by Rogstad (1968), who has also pointed out that with a multielement system the phase information can be utilized by means of closure relationships as discussed in Section 10.3.

Optical interferometers can be built with very wide bandwidths, that is, $\Delta\lambda/\lambda \simeq 0.1$ or possibly more, so the central, or white light, fringe is readily identifiable. If such a system is made to operate at two such wide wavelength bands simultaneously, the effects of the atmosphere, which is slightly dispersive, can be removed. Ground-based optical astrometry with dual-wavelength phase-tracking interferometers can yield accurate positions of stars (Colavita, Shao, and Staelin 1987). As examples of earlier interferometry, Currie, Knapp, and Liewer (1974) made measurements using two apertures on a single large telescope, and Labeyrie (1975) obtained the first successful measurements using two telescopes. For descriptions of later, more complex, instruments see, for example, Davis and Tango (1985), Shao et al. (1988), Baldwin et al. (1994), Mourard et al. (1994), Armstrong et al. (1998), and Davis et al. (1999a,b). For use in space where the earth's atmosphere is avoided, optical interferometry holds great promise. The Space Interferometry Mission (Shao 1998, Allen and Böker 1998, Böker and Allen 1999) is a space-based interferometer for the wavelength band 0.4–1.0 μm with variable baseline up to 10 m, intended to provide synthesis imaging with a resolution of 10 mas, and to measure fringe phases with sufficient accuracy to provide positions of stars to within 4 μarcsec. An application of space interferometry to the detection of planets around distant stars is discussed by Bracewell and MacPhie (1979). The ratio of the signal from the planet to that from the star is maximized by choosing an infrared wavelength on the long-wavelength side of 20 μm, and by placing a fringe-pattern null in the direction of the star. A demonstration of the nulling technique using ground-based telescopes is described by Hinz et al. (1998).

In the systems mentioned above, the fringes are formed by combining the incoming radiation at the same wavelength as it is received, as in the classical Michelson stellar interferometer. They are therefore also referred to as *direct detection systems*. An alternative to the direct detection system is the *heterodyne system*, in which the light from each aperture is mixed with coherent light from a central laser to produce an intermediate frequency. The IF waveforms are then amplified and correlated in an electronic system, in a manner basically identical to that used in radio interferometry. In comparison with a direct detection system, the sensitivity is greatly limited by the quantum effects mentioned in Section 1.4. It is also limited by the bandwidth that can be handled by the electronic amplifiers, unless the mixer outputs are split into many frequency channels, each of which is processed in parallel. A large bandwidth can then be processed using a correspondingly large number of amplifiers and correlators. The bandwidth division also has the effect of increasing the path length difference over which the signals remain coherent. The heterodyne technique has been used in infrared interferometry; see, for example, Johnson, Betz, and Townes (1974), Assus et al. (1979), and Bester, Danchi, and Townes (1990). Possible application to large multiele-

ment telescopes with multiband processing in the infrared and visible ranges has been discussed by Swenson, Gardner, and Bates (1986).

From the submillimeter radio range to the optical is a factor of $\sim 10^3$ in wavelength, and a further factor of $\sim 10^3$ takes one to the X-ray region. X-ray astronomy could benefit tremendously by the potentially high angular resolution obtainable through interferometry. The viability of X-ray interferometry, suitable for astronomical imaging, has been demonstrated in the laboratory by Cash et al. (2000). It holds promise of providing extremely high angular resolution in observations above the atmosphere. At a wavelength of 2 nm, a baseline of 1 m provides a fringe spacing of 4×10^{-4} arcsec. In the laboratory instrument, the apertures are defined by flat reflecting surfaces, which are used at grazing incidence to minimize the requirement for surface accuracy. Direct detection is the only available technique, and if the fringes are formed by simply allowing the reflected beams to converge on a detector surface, a long distance is required to obtain sufficient fringe spacing. With 4×10^{-4} arcsec angular spacing of the fringes, adjacent maxima would be separated by only 1 μ at 500-m distance. Thus more complicated systems are likely to be required for practical astronomical interferometry.

Sensitivity of Direct Detection and Heterodyne Systems

Factors that determine the sensitivity of optical systems, such as losses due to scattering, partial reflection, and absorption, are different from corresponding effects at radio wavelengths. However, in heterodyne systems the most important difference is the role of quantum effects. The energy of optical photons is five or more orders of magnitude greater than that of radio photons, and quantum effects are largely negligible in the radio domain at frequencies lower than $\sim$100 GHz. In the optical range (wavelength $\sim$500 μm) the frequency is of order 600 THz, and the bandwidth could be as high as 100 THz. In a typical heterodyne system in the infrared, the wavelength of 10 μm corresponds to 30 THz, and the bandwidth is $\sim$3 GHz [see, e.g., Townes et al. (1998)].

In direct detection systems the detector or photon counter does not preserve the phase of the signal, and thus the noise resulting from the uncertainty priciple, discussed in Section 1.4, does not occur. The noise is principally shot noise resulting from the random arrival times of the signal photons. The number of photons received from a source of intensity I is

$$N = \frac{I \Omega_s A \Delta\nu}{h\nu} \quad \text{(photons s}^{-1}\text{)}, \tag{16.21}$$

where Ω_s is the solid angle of the source (with no atmospheric blurring), A is the collecting area of the telescope, $\Delta\nu$ is the bandwidth, ν is the frequency, and h is Planck's constant. If the source is a blackbody at temperature T, Planck's formula gives

$$I = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT} - 1}. \tag{16.22}$$

Note that for direct detection we are considering the signal in both polarizations. Thus we have

$$N = \frac{2\Omega_s A \Delta\nu}{\lambda^2} \frac{1}{e^{h\nu/kT} - 1} \quad \text{(photons s}^{-1}\text{)}. \tag{16.23}$$

The received power is

$$P = h\nu N. \tag{16.24}$$

The fluctuations in the power, ΔP_D, are caused by photon shot noise and therefore are proportional to $\sqrt{N}$. Thus

$$\Delta P_D = h\nu\sqrt{N}. \tag{16.25}$$

ΔP_D is known as the noise equivalent power (NEP). The signal-to-noise ratio in one second is $P/\Delta P_D = \sqrt{N}$, and therefore for an integration time τ_a the signal-to-noise ratio for direct detection is

$$\mathcal{R}_{\text{snD}} = \left[\left(\frac{2\Omega_s A}{\lambda^2}\right)\frac{\Delta\nu\tau_a}{e^{h\nu/kT} - 1}\right]^{1/2}, \tag{16.26}$$

where the subscript D indicates direct detection. Note that $\mathcal{R}_{\text{snD}}$ is proportional to $\sqrt{A}$, because of the shot noise, rather than to A as in the radio case.

In a heterodyne system the noise is determined by the uncertainty principle, since the mixer is a linear device that preserves phase. The minimum noise is one photon per mode (one photon per hertz per second), as noted in the discussion following Eq. (1.14). This is equivalent to saying that the system temperature is $h\nu/k$ [see, e.g., Heffner (1962), Caves (1982)]. Hence, in a period of one second the uncertainty in power is

$$\Delta P_H = h\nu\sqrt{\Delta\nu}. \tag{16.27}$$

The heterodyne detector responds only to the component of the radiation to which its polarization is matched, and the received power is half of that in Eq. (16.24). The signal-to-noise ratio for a heterodyne system (indicated by subscript H) is therefore $P/(2\Delta P_H)$ in one second, and in time τ_a it is

$$\mathcal{R}_{\text{snH}} = \left(\frac{\Omega_s A}{\lambda^2}\right)\frac{\sqrt{\Delta\nu\tau_a}}{e^{h\nu/kT} - 1}. \tag{16.28}$$

Note that Eq. (16.28) reduces to the usual radio form in Eq. (1.7) when $h\nu/kT \ll 1$. In that case, $T_A = T\Omega_s A/\lambda^2$ and the minimum value of $h\nu/k$ can be used for system temperature. The ratio of signal-to-noise ratios for the direct detection and heterodyne systems, when parameters other than the bandwidth are the same, is

$$\frac{\mathcal{R}_{\mathrm{snH}}}{\mathcal{R}_{\mathrm{snD}}} \simeq \sqrt{\left(\frac{\Omega_s A}{2\lambda^2}\right)\frac{1}{e^{h\nu/kT}-1}\left(\frac{\Delta\nu_{\mathrm{H}}}{\Delta\nu_{\mathrm{D}}}\right)}. \tag{16.29}$$

As indicated earlier, $\sqrt{\Delta\nu_{\mathrm{H}}/\Delta\nu_{\mathrm{D}}}$ could be as low as $\sim 4 \times 10^{-3}$. However, for direct detection, the propagation delays through the different siderostats to the fringe forming point must be maintained constant to $\sim 1/10$ of the reciprocal bandwidth. This requirement restricts the bandwidths that can practicably be used, especially with baselines of hundreds of meters. The heterodyne system offers simpler hardware that provides useful sensitivity at 10 μm wavelength, and possibly to the next atmospheric window at 5 μm. It also allows the amplified IF signals to be split without loss in sensitivity, to provide multiple simultaneous correlations in multielement arrays. Relative advantages of the heterodyne and direct detection systems are discussed by Townes and Sutton (1981) and de Graauw and van de Stadt (1981).

Optical Intensity Interferometer

The use of the intensity interferometer for optical measurements on stars was demonstrated by Hanbury Brown and Twiss (1956a), shortly after the success of the radio intensity interferometer described in Section 1.3 under *Early Measurements of Angular Width* and Section 16.1. At that time the possibility of coherence between photons in different light rays from the same source was questioned, and the physical basis and consistency with quantum mechanics is explained by Hanbury Brown and Twiss (1956c) and Purcell (1956). The laboratory demonstration of the correlation of intensity fluctuations of light by Hanbury Brown and Twiss (1956b) led to the broader development of quantum statistical studies, and their application to particle beams as well as electromagnetic radiation (Henny et al. 1999).

In the optical intensity interferometer, a photomultiplier tube at the focus of each telescope mirror replaces the RF and IF stages and the detectors of the radio instrument. The photomultiplier outputs are amplified and fed to the inputs of the correlator. The optical intensity interferometer is largely insensitive to atmospheric phase fluctuations, as explained for the radio case in Section 16.1. The size of the light-gathering apertures is therefore unrestricted by the scale size of the irregularities. Also, it is not necessary that the reflecting mirrors produce a diffraction-limited image, and their accuracy need only be sufficient to deliver all the light to the photomultiplier cathodes. This is fortunate since the low sensitivity mentioned earlier for the radio case necessitates the use of large light-gathering areas. Hanbury Brown (1974) has given an analysis of the response of the optical instrument and has shown that it is proportional to the square of the visibility modulus as in the radio case. Either a correlator or a photon coincidence counter can be used to combine the photomultiplier outputs.

The intensity interferometer constructed at Narrabri, Australia (Hanbury Brown, Davis, and Allen 1967; Hanbury Brown 1974), used two 6.5-m diameter reflectors and a bandwidth of 60 MHz for the signals at the correlator inputs. The resulting limiting magnitude of +2.5 enabled measurements of 32 stars

to be made. Davis (1976) has discussed the relative merits of the intensity interferometer and modern implementations of the Michelson interferometer for development of more sensitive instruments.

Speckle Imaging

The image of an unresolved point source observed with a telescope of which the width of the aperture is large compared with the Fried length r_0 depends on the exposure time over which the image is averaged. An exposure no longer than 10 ms shows a group of bright speckles, each of which is of the approximate size of the Airy disk (i.e., the diffraction-limited point-source image) of the telescope. If the exposure is much longer, the pattern is blurred into a single patch (the "seeing" disk) of typical diameter 1 arcsec, determined by the atmosphere. The characteristic fluctuation time of 10 ms in the optical range corresponds to the time taken for a cell of size $r_0 \simeq 0.14$ m to move past any point in the telescope aperture at a typical wind speed of 10–20 m s^{-1}. The use of sequences of short-exposure images to obtain information at the diffraction limit of a large telescope is known as speckle imaging. Speckle patterns reflect the random distribution of atmospheric irregularities over the aperture, and differ from one exposure to the next on the 10-ms timescale. Reduction of many exposures is required to observe faint objects by this technique.

For the theory of the speckle response see, for example, Dainty (1973), Bates (1982), or Goodman (1985). Here we note that the high-resolution image represented by a single speckle can be understood if one considers each speckle as resulting from several seeing cells of the wavefront, located at points distributed across the telescope aperture. These cells are the ones that present approximately equal phase shifts in the ray paths from the wavefront to the speckle image (Worden 1977). Then, by analogy with an array of antennas, the resolution corresponds to the maximum spacing of the cells, that is, it is of the order λ/d, where d is the telescope aperture. Aberrations in the reflector do not significantly degrade the speckle pattern as long as the dominant phase irregularities are those of the atmosphere. The area of the image over which the speckles are spread corresponds to λ/r_0 on the sky and becomes the seeing disk in a long exposure. The seeing cells can be regarded as subapertures within the main telescope aperture, the responses of which combine with random phases in the image. The number of speckles is of the order of the number of subapertures, that is, $(d/r_0)^2$. With a large telescope ($d \sim 1$ m) this number is of the order of 50 at optical wavelengths. Also, the size of the seeing cells increases with wavelength, and in the infrared only a few speckles appear in the image.

A rather simple image restoration technique called the "shift-and-add" algorithm can be applied to speckle images (Christou 1991). It works best when there is a point source in the field, and at infrared wavelengths where there are relatively few speckles per frame and the isoplanatic patch is relatively large (see Table 16.1). The short exposure speckle frames are aligned on their brigthest speckles and summed. The point spread function ("dirty beam"), which can be obtained from the image of a point source within the field, will have a diffraction-

limited component and a much broader component composed of the fainter speckles. This step can be followed by other restoration algorithms such as CLEAN (see Section 11.2) to improve the image quality further [see, e.g., Eckart et al. (1994)].

When the shift-and-add algorithm is not applicable, the modulus of the visibility can be obtained by the technique of speckle interferometry, which originated with Labeyrie (1970). This procedure can be understood from the following simplified discussion. On a single image of short exposure, a number of approximately diffraction-limited speckles appear at random locations within the seeing disk. The speckle image $I_s(l, m)$ can be described as the convolution of the actual intensity distribution $I(l, m)$ with the speckle point spread function $\mathcal{P}(l, m)$. Thus

$$I_s(l, m) = I(l, m) ** \mathcal{P}(l, m). \tag{16.30}$$

The function $\mathcal{P}(l, m)$ is a random function that cannot be specified exactly. As a first approximation, we will assume that $\mathcal{P}(l, m)$ is the point spread function of the telescope in the absence of atmospheric effects, $b_0(l, m)$, replicated at the position of each speckle. Thus we can write

$$\mathcal{P}(l, m) = \sum b_0(l - l_i, m - m_i), \tag{16.31}$$

where l_i and m_i are the locations of the speckles, all of which are assumed to have the same intensity. From Eqs. (16.30) and (16.31), we obtain

$$I_s(l, m) = \sum I(l, m) ** b_0(l - l_i, m - m_i). \tag{16.32}$$

If the Fourier transform of $b_0(l, m)$ is $\bar{b}_0(u, v)$, then the Fourier transform of $b_0(l - l_i, m - m_i)$ is $\bar{b}_0(u, v)\exp[j2\pi(ul_i + vm_i)]$. Hence, the Fourier transform of Eq. (16.32) can be written as

$$\bar{I}_s(u, v) = \sum \mathcal{V}(u, v)\bar{b}_0(u, v)e^{j2\pi(ul_i + vm_i)}, \tag{16.33}$$

where $\mathcal{V}$ and $\bar{I}$ are the Fourier transforms of I and I_s, respectively. The speckle transforms $\bar{I}_s$ cannot be summed directly because of random phase factors in Eq. (16.33). To eliminate these phase factors, we calculate $|\bar{I}_s|^2$ (i.e., $\bar{I}_s\,\bar{I}_s^*$), which is

$$\begin{aligned} |\bar{I}_s(u, v)|^2 &= \sum_i \sum_k |\mathcal{V}(u, v)|^2 |\bar{b}_0(u, v)|^2 e^{j2\pi[u(l_i - l_k) + v(m_i - m_k)]} \\ &= |\mathcal{V}(u, v)|^2 |\bar{b}_0(u, v)|^2 \left[N + \sum_{i \neq k} e^{j2\pi[u(l_i - l_k) + v(m_i - m_k)]} \right], \end{aligned} \tag{16.34}$$

where N is the number of speckles. Since the expectation of the summation term in the second line of Eq. (16.34) is zero, the expectation of Eq. (16.34) is

$$\langle|\bar{I}_s(u, v)|^2\rangle = N_0|\mathcal{V}(u, v)|^2|\bar{b}_0(u, v)|^2, \tag{16.35}$$

where N_0 is the average number of speckles. Hence, the average of a series of measurements of $|I_s(u, v)|^2$, estimated from short exposures, is proportional to the squared modulus of $\mathcal{V}(u, v)$ times the squared modulus of $\bar{b}_0(u, v)$. Since $b_0(u, v)$ is nonzero for $|u|, |v| < D/\lambda$, the function $|\mathcal{V}(u, v)|^2$ can be determined over the same range of u and v, if $\bar{b}_0(u, v)$ is known. In practice the speckles cannot be accurately modeled by Eq. (16.31). However, we can write

$$\langle|\bar{I}_s(u, v)|^2\rangle = |\mathcal{V}(u, v)|^2\langle|\overline{\mathcal{P}}(u, v)|^2\rangle, \tag{16.36}$$

where $\overline{\mathcal{P}}(u, v)$ is the Fourier transform of $\mathcal{P}(l, m)$. From Eqs. (16.35) and (16.36), $\langle|\overline{\mathcal{P}}(u, v)|^2\rangle$ should be approximately proportional to $|\bar{b}_0(u, v)|^2$. It can be estimated by observing a point source under the same conditions as those for the source under study.

The phase information can be extracted from the speckle frames, but with considerably more computational effort. Most phase retrieval algorithms are variations of two basic methods: the Knox–Thompson, or cross-spectral, method [Knox and Thompson (1974), Knox (1976)], and the bispectrum method (Lohmann, Weigelt, and Wirnitzer 1983). These methods are described in detail by Roggeman, Welch, and Fugate (1997).

BIBLIOGRAPHY

Lawson, P. R., Ed., *Selected Papers on Long Baseline Stellar Interferometry*, SPIE Milestone Ser. MS139, SPIE, Bellingham, WA, 1997.

Lawson, P. R., Ed., *Principles of Long Baseline Stellar Interferometry*, Course Notes from the 1999 Michelson Summer School, Jet Propulsion Laboratory, Pasadena, CA, 2000.

Léna, P. J., and A. Quirrenbach, Eds., *Interferometry in Optical Astronomy*, Proc. SPIE, **4006**, SPIE, Bellingham, WA, 2000.

Reasenberg, R. D., Ed, *Astronomical Interferometry*, Proc. SPIE, **3350**, SPIE, Bellingham, WA, 1998.

Robertson, J. G., and W. J. Tango, Eds., Very High Angular Resolution Imaging, IAU Symp. 158, Kluwer, Dordrecht, 1994.

REFERENCES

Allen, R. J. and T. Böker, Optical Interferometry and Aperture Synthesis in Space with the Space Interferometry Mission, in *Astronomical Interferometry*, R. D. Reasenberg, Ed., Proc. SPIE, **3350**, 561–570, 1998.

Armstrong, J. T., D. Mozurkewich, L. J. Rickard, D. J. Hutter, J. A. Benson, P. F. Bowers, N. M. Elias II, C. A. Hummel, K. J. Johnston, D. F, Buscher, J. H. Clark III, L. Ha, L.-C. Ling, N. M. White, and R. S. Simon, The Navy Prototype Optical Interferometer, *Astrophys. J.*, **496**, 550–571, 1998.

Assus, P., H. Choplin, J. P. Corteggiani, E. Cuot, J. Gay, A. Journet, G. Merlin, and Y. Rabbia, L'Interféromètre Infrarouge du C.E.R.G.A., *J. Opt.* (Paris), **10**, 345–350, 1979.

Baldwin, J. E., M. G. Beckett, R. C. Boysen, D. Burns, D. F. Buscher, G. C. Cox, C. A. Haniff, C. D. Mackay, N. S. Nightingale, J. Rogers, P. A. G. Scheuer, T. R. Scott, P. G. Tuthill, P. J. Warner, D. M. A. Wilson, and R. W. Wilson, The First Images from an Optical Aperture Synthesis Array: Mapping of Capella with COAST at Two Epochs, *Astron. Astrophys.*, **306**, L13–L16, 1996.

Baldwin, J. E., R. C, Boysen, G. C. Cox, C. A. Haniff, J. Rogers, P. J. Warner, D. M. A. Wilson, and C. D. Mackay, Design and Performance of COAST, *Amplitude and Intensity Spatial Interferometry II*, J. B. Breckinridge, Ed., Proc. SPIE, **2200**, 118–128, 1994.

Bates, R. H. T., Astronomical Speckle Imaging, *Phys. Rep.*, **90**, 203–297, 1982.

Bates, R. H. T., Uniqueness of Solutions to Two-Dimensional Fourier Phase Problems for Localized and Positive Images, *Comp. Vision, Graphics Image Process.*, **25**, 205–217, 1984.

Bennett, J. C., A. P. Anderson, and P. A. McInnes, Microwave Holographic Metrology of Large Reflector Antennas, *IEEE Trans. Antennas Propag.*, **AP-24**, 295–303, 1976.

Benson, J. A., D. J. Hutter, N. M. Elias, P. F. Bowers, K. J. Johnston, A. R. Haijian, J. T. Armstrong, D. Mozurkewich, T. A. Pauls, L. J. Rickard, C. A. Hummel, N. A. White, D. Black, and C. S. Denison, Multichannel Optical Aperture Synthesis Imaging of Eta 1 Ursae Majoris with Navy Optical Prototype Interferometer, *Astron. J.*, **114**, 1221–1226, 1997.

Bester, M., W. C. Danchi, C. G. Degiacomi, L. J. Greenhill, and C. H. Townes, Atmospheric Fluctuations: Empirical Structure Functions and Projected Performance of Future Instruments, *Astrophys. J.*, **392**, 357–374, 1992.

Bester, M., W. C. Danchi, and C. H. Townes, Long Baseline Interferometer for the Mid-Infrared, *Amplitude and Intensity Spatial Interferometry*, J. B. Breckinridge, Ed., Proc. SPIE, **1237**, 40–48, 1990.

Böker, T. and R. J. Allen, Imaging and Nulling with the Space Interferometer Mission, *Astrophys. J. Supl.*, **125**, 123–142, 1999.

Bracewell, R. N., Radio Interferometry of Discrete Sources, *Proc. IRE*, **46**, 97–105, 1958.

Bracewell, R. N. and R. H. MacPhie, Searching for Nonsolar Planets, *Icarus*, **38**, 136–147, 1979.

Carr, T. D., M. A. Lynch, M. P. Paul, G. W. Brown, J. May, N. F. Six, V. M. Robinson, and W. F. Block, Very Long Baseline Interferometry of Jupiter at 18 MHz, *Radio Sci.*, **5**, 1223–1226, 1970.

Cash, W., A. Shipley, S. Osterman, and M. Joy, Laboratory Detection of X-ray Fringes with a Grazing-Incidence Interferometer, *Nature*, **407**, 160–162, 2000.

Caves, C. M., Quantum Limits on Noise in Linear Amplifiers, *Phys. Rev.*, **26D**, 1817–1839, 1982.

Chen, M. T., C.-Y. E. Tong, R. Blundell, D. C. Papa, and S. Paine, Receiver Beam Characterization for the SMA, in *SPIE Conf. Advanced Technology MMW, Radio, and Terahertz Telescopes*, Kona, Hawaii, March 1998, T. G. Phillips, Ed., Proc. SPIE **3357**, 106–113, 1998.

Christou, J. C., Infrared Speckle Imaging: Data Reduction with Application to Binary Stars, *Experimental Astronomy*, **2**, 27–56, 1991.

Cohen, M. H., High Resolution Observations of Radio Sources, *Ann. Rev. Astron. Astrophys.*, **7**, 619–664, 1969.

Colavita, M. M., M. Shao, and D. H. Staelin, Two-color Method for Optical Astrometry: Theory and Preliminary Measurements with the Mark III Stellar Interferometer, *Appl. Opt.*, **26**, 4113–4122, 1987.

Currie, D. G., S. L. Knapp, and K. M. Liewer, Four Stellar-Diameter Measurements by a New Technique: Amplitude Interferometry, *Astrophys. J.*, **187**, 131–134, 1974.

D'Addario, L. R., *Holographic Antenna Measurements: Further Technical Considerations*, 12 Meter Millimeter Wave Telescope Memo. 202, National Radio Astronomy Observatory, Charlottesville, VA, 1982.

Dainty, J. C., Diffraction-Limited Imaging of Stellar Objects Using Telescopes of Low Optical Quality, *Opt. Commun.*, **7**, 129–134, 1973.

Davis, J., High Angular Resolution Stellar Interferometry, *Proc. Astron. Soc. Aust.*, **3**, 26–32, 1976.

Davis, J., P. R. Lawson, A. J. Booth, W. J. Tango, and E. D. Thorvaldson, Atmospheric Path Variations for Baselines up to 80 m Measured with the Sydney University Stellar Interferometer, *Mon. Not. R. Astron. Soc.*, **273**, L53–L58, 1995.

Davis, J. and W. J. Tango, The Sydney University 11.4 m Prototype Stellar Interferometer, *Proc. Astron. Soc. Aust.*, **6**, 34–38, 1985.

Davis, J., W. J. Tango, A. J. Booth, T. A. ten Brummelaar, R. A. Minard, and S. M. Owens, The Sydney University Stellar Interferometer—I. The Instrument, *Mon. Not. R. Astron. Soc.*, **303**, 773–782, 1999a.

Davis, J., W. J. Tango, A. J. Booth, E. D. Thorvaldson, and J. Giovannis, The Sydney University Stellar Interferometer—II. Commissioning Observations and Results, *Mon. Not. R. Astron. Soc.*, **303**, 783–791, 1999b.

de Graauw, T. and H. van de Stadt, Coherent Versus Incoherent Detection for Interferometry at Infrared Wavelengths, *Proc. ESO Conf. Scientific Importance of High Angular Resolution at Infrared and Optical Wavelengths*, M. H. Ulrich and K. Kjär, Eds., European Southern Observatory, Garching, 1981.

Dulk, G. A., Characteristics of Jupiter's Decametric Radio Source Measured with Arc-Second Resolution, *Astrophys. J.*, **159**, 671–684, 1970.

Eckart, A., R. Genzel, R. Hofmann, B. J. Sams, L. E. Tacconi-Garman, and P. Cruzalebes, Diffraction Limited Near-Infrared Imaging of the Galactic Center, in *The Nuclei of Normal Galaxies*, R. Genzel and A. Harris, Eds., Kluwer, Dordrecht, 1994, pp. 305–315.

Eddington, A. S., Note on Major MacMahon's paper 'On the Determination of the Apparent Diameter of a Fixed Star,' *Mon. Not. R. Astron. Soc.*, **69**, 178–180, 1909.

Evans, D. S., D. A. Edwards, M. Frueh, A. McWilliam, and W. Sandmann, Photoelectric Observations of Lunar Occultations. XV, *Astron. J.*, **90**, 2360–2371, 1985.

Fried, D. L., Optical Resolution through a Randomly Inhomogenious Medium for Very Long and Very Short Exposures, *J. Opt. Soc. Am.*, **56**, 1372–1379, 1966.

Goodman, J. W., *Statistical Optics*, Wiley, New York, 1985, pp. 441–459.

Hanbury Brown, R., *The Intensity Interferometer*, Taylor and Francis, London, 1974.

Hanbury Brown, R. and R. Q. Twiss, A New Type of Interferometer for Use in Radio Astronomy, *Philos. Mag.*, ser. 7, **45**, 663–682, 1954.

Hanbury Brown, R. and R. Q. Twiss, A Test of a New Type of Stellar Interferometer on Sirius, *Nature*, **178**, 1046–1048, 1956a.

Hanbury Brown, R. and R. Q. Twiss, Correlation between Photons in Two Coherent Light Beams, *Nature*, **177**, 27–29, 1956b.

Hanbury Brown, R. and R. Q. Twiss, A Question of Correlation Between Photons in Coherent Light Rays, *Nature*, **178**, 1447–1448, 1956c.

Hanbury Brown, R., J. Davis, and L. R. Allen, The Stellar Interferometer at Narrabri Observatory-I, *Mon. Not. R. Astron. Soc.*, **137**, 375–392, 1967.

Haniff, C. A., C. D. Mackay, D. J. Titterington, D. Sivia, J. E. Baldwin, and P. J, Warner, The First Images from Optical Aperture Synthesis, *Nature*, **328**, 694–696, 1987.

Hazard, C., Lunar Occultation Measurements, in *Methods of Experimental Physics*, Vol. 12C, M. L. Meeks, Ed., Academic Press, New York, 1976.

Hazard, C., M. B. Mackey, and A. J. Shimmins, Investigation of the Radio Source 3C273 by the Method of Lunar Occultations, *Nature*, **197**, 1037–1039, 1963.

Heffner, H., The Fundamental Noise Limit of Linear Amplifiers, *Proc. IRE*, **50**, 1604–1608, 1962.

Henny, M., S. Oberholzer, C. Strunk, T. Heinzel, K. Esslin, M. Holland, and C. Schönenberger, The Fermionic Hanbury Brown and Twiss Experiment, *Science*, **284**, 296–298, 1999.

Hinz, P. M., J. R. P. Angel, W. F. Hoffman, D. W. McCarthy Jr., P. C. McGuire, M. Cheselka, J. L. Hora, and N. J. Woolf, Imaging Circumstellar Environments with a Nulling Interferometer, *Nature*, **395**, 251–253, 1998.

Jennison, R. C. and M. K. Das Gupta, The Measurement of the Angular Diameter of Two Intense Radio Sources, *Philos. Magn.*, Ser. 8, **1**, 55–75, 1956.

Johnson, M. A., A. L. Betz, and C. H. Townes, 10-μm Heterodyne Stellar Interferometer, *Phys. Rev. Lett.*, **33**, 1617–1620, 1974.

Knox, K. T., Image Retrieval from Astronomical Speckle Patterns, *J. Opt. Soc. Am.*, **66**, 1236–1239, 1976.

Knox, K. T. and B. J. Thompson, Recovery of Images from Atmospherically Degraded Short-Exposure Photographs, *Astrophys. J.*, **193**, L45–L48, 1974.

Labeyrie, A., Attainment of Diffraction Limited Resolution in Large Telescopes by Fourier Analysing Speckle Patterns in Star Images, *Astron. Astrophys.*, **6**, 85–87, 1970.

Labeyrie, A., Interference Fringes Obtained on Vega with Two Optical Telescopes, *Astrophys. J.*, **196**, L71–L75, 1975.

Lohmann, A. W., G. Weigelt, and B. Wirnitzer, Speckle Masking in Astronomy: Triple Correlation Theory and Applications, *Applied Optics*, **22**, 4028–4037, 1983.

MacMahon, P. A., On the Determination of the Apparent Diameter of a Fixed Star, *Mon. Not. R. Astron. Soc.*, **69**, 126–127, 1909.

Milonni, P. W., Resource Letter: AOA-1: Adaptive Optics in Astronomy, *Am J. Phys.*, **67**, 476–485, 1999.

Misell, D. L., A Method for the Solution of the Phase Problem in Electron Microscopy, *J. Phys. D.*, **6**, L6–L9, 1973.

Morris, D., Phase Retrieval in the Radio Holography of Reflector Antennas and Radio Telescopes, *IEEE Trans. Antennas Propag.*, **AP-33**, 749–755, 1985.

Morris, D., J. W. M. Baars, H. Hein, H. Steppe, C. Thum, and R. Wohlleben, Radio-Holographic Reflector Measurements on the 30-m millimeter Radio Telescope at 22 GHz with a Cosmic Signal Source, *Astron. Astrophys.*, **203**, 399–406, 1988.

Mourard, D., I. Tallon-Bosc, A. Blazit, D. Bonneau, G. Merlin, F. Morand, F. Vakili, and A. Labeyrie, The G12T Interferometer on Plateau de Calern, *Astron. Astrophys.*, **283**, 705–713, 1994.

Napier, P. J. and R. H. T. Bates, Antenna-Aperture Distributions from Holographic Type of Radiation-Pattern Measurements, *Proc. IEEE*, **120**, 30–34, 1973.

Purcell, E. M., A Question of Correlation Between Photons in Coherent Light Rays, *Nature*, **178**, 1449–1450, 1956.

Ragland, S. and A. Richichi, Detection of a Sub-Arcsecond Dust Shell around the Wolf-Rayet Star WR112, *Mon. Not. R. Astron. Soc.*, **302**, L13–L16, 1999.

Richichi, A., Lunar Occultations, in *Proc. Very High Angular Resolution Imaging, IAU Symp. 158*, J. G. Robertson and W. J. Tango, Eds., Kluwer, Dordrecht, 1994, pp. 71–81.

Rogers, A. E. E., R. Barvainis, P. J. Charpentier, and B. E. Corey, Corrections for the Effect of a Radome on Antenna Surface Measurements Made by Microwave Holography, *IEEE Trans. Antennas. Propag.*, **AP-41**, 77–84, 1993.

Roggemann, M. C., B. M. Welch, and R. Q. Fugate, Improving the Resolution of Ground-Based Telescopes, *Rev. Mod. Phys.*, **69**, 437–505, 1997.

Rogstad, D. H., A Technique for Measuring Visibility Phase with an Optical Interferometer in the Presence of Atmospheric Seeing, *Appl. Opt.*, **7**, 585–588, 1968.

Scheuer, P. A. G., On the Use of Lunar Occultations for Investigating the Angular Structure of Radio Sources, *Aust. J. Phys.*, **15**, 333–343, 1962.

Scott, P. F. and M. Ryle, A Rapid Method for Measuring the Figure of a Radio Telescope Reflector, *Mon. Not. R. Astron. Soc.*, **178**, 539–545, 1977.

Serabyn, E., T. G. Phillips, and C. R. Masson, Surface Figure Measurements of Radio Telescopes with a Shearing Interferometer, *Applied Optics*, **30**, 1227–1241, 1991.

Shao, M., SIM the Space Interferometry Mission, in *Astronomical Interferometry*, R. D. Reasenberg, Ed., Proc. SPIE, **3350**, 536–540, 1998.

Shao, M. and Colavita, M. M., Long-Baseline Optical and Stellar Interferometry, *Ann. Rev. Astron. Astrophys.*, **30**, 457–498, 1992.

Shao, M., M. M. Colavita, B. E. Hines, D. H. Staelin, H. J. Hutter, K. J. Johnston, D. Mozurkewich, R. S. Simon, J. L. Hershey, J. A. Hughes, and G. H. Kaplan, The Mark III Stellar Interferometer, *Astron. Astrophys.*, **193**, 357–371, 1988.

Swenson, G. W., Jr., C. S. Gardner, and R. H. T. Bates, Optical Synthesis Telescopes, *Proc. SPIE*, **643**, 129–140, 1986.

Tango, W. J, and R. Q. Twiss, Michelson Stellar Interferometry, *Prog. Opt.*, **17**, 239–277, 1980.

Taylor, J. H. and M. L. De Jong, Models of Nine Radio Sources from Lunar Occultation Observations, *Astrophys. J.*, **151**, 33–42, 1968.

Townes, C. H., M. Bester, W.C. Danchi, D. D. S. Hale, J. D. Monnier, E. A. Lipman, P. G. Tuthill, M. A. Johnson, and D. Walters, Infrared Spatial Interferometer, in *Astronomical Interferometry*, R. D. Reasonberg, Ed., Proc. SPIE, Vol. 3350, 908–932, 1998.

Townes, C. H. and E. C. Sutton, Multiple Telescope Infrared Interferometry, *Proc. ESO Conf. on Scientific Importance of High Angular Resolution at Infrared and Optical Wavelengths*, M. H. Ulrich and K. Kjär, Eds., European Southern Observatory, Garching, 1981, pp. 199–223.

von Hoerner, S., Lunar Occultations of Radio Sources, *Astrophys. J.*, **140**, 65–79, 1964.

White, N. M. and B. H. Feierman, A Catalog of Stellar Angular Diameters Measured by Lunar Occultation, *Astrophys. J.*, **94**, 751–770, 1987.

Whitford, A. E., Photoelectic Observation of Diffraction at the Moon's Limb, *Astrophys. J.*, **89**, 472–481, 1939.

Woolf, N. J., High Resolution Imaging from the Ground, *Ann. Rev. Astron. Astrophys.*, **20**, 367–398, 1982.

Worden, S. P., Astronomical Image Reconstruction, *Vistas in Astron.*, **20**, 301–318, 1977.

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

PRINCIPAL SYMBOLS

Listed below are the principal symbols used throughout the book. Locally defined symbols with restricted usage are selectively included.

a	Model dimension, scale size, atmospheric model constant (Section 13.1), scale size of ionospheric irregularities (Section 13.4)
A	Antenna collecting area (reception pattern)
$\mathbf{A}$	Antenna polarization matrix (Chapter 4)
A_1	One-dimensional reception pattern
A_0	Antenna collecting area on axis
A_N	Normalized reception pattern
$\mathcal{A}$	Mirror-image reception pattern, azimuth
b	Galactic latitude (Section 13.6)
b_0	Snythesized beam pattern, point-source response
b_N	Normalized synthesized beam pattern
B	Magnetic field magnitude
$\mathbf{B}$	magnetic field vector
c	Velocity of light
C	Coherence function (Chapter 9), convolving function (Chapter 10)
C_n^2, C_{ne}^2	Turbulence strength parameters for refractive index, electron density (Chapter 13)
$\mathcal{C}$	Amplitude of a complex signal (Appendix 3.1)
d	Distance, antenna diameter, baseline declination, projected baseline (Chapter 13)
d_{tc}	Distance between ray paths to target and calibrator sources in turbulent region
d_0	Distance over which rms phase deviation = 1 rad (Chapter 13)
D	Baseline (antenna spacing), polarization leakage (Chapter 4)
$\mathbf{D}$	Baseline vector

D_λ, $\mathbf{D}_\lambda$	Baseline measured in wavelengths
D_a, $\mathbf{D}_a$	Interaxis distance of antenna mount (Chapter 4)
D_E	Equatorial component of baseline
D_m	Dispersion measure (Chapter 13)
D_R	Delay resolution function [Eq. (9.161)]
$\mathcal{D}$	Dispersion in optical fiber (Section 7.1, Appendix 7.2), sensitivity degradation factor (Section 7.3)
$\mathcal{D}_\tau$	Structure function of phase (temporal) (Chapter 13)
$\mathcal{D}_\phi$	Structure function of phase (spatial) (Chapters 12, 13)
$\mathcal{D}_n$	Structure function of refractive index (spatial) (Chapter 13)
e	Electronic charge = $-e$ (Chapter 13)
E, $\mathbf{E}$	Electric field (usually in the measurement plane), spectral components of electric field, energy
E_x, E_y	Components of electric field
$\mathcal{E}$	Electric field at a source or aperture (Chapters 3, 14, 16), elevation angle
f	Frequency of Fourier components of power spectrum (Chapters 9, 13)
f_i	Oscillator strength at resonance i (Chapter 13)
f_m, f_n	Phase switching waveforms (Chapter 7)
F	Power flux density (W m^{-2}), fringe function
F_h	Threshold of harmful interference (W m^{-2}) (Chapter 15)
$F(\beta)$	Faraday dispersion function (Chapter 13)
F_1, F_2	See Eqs. (9.17)
F_1, F_2, F_3	Entropy measures (Chapter 11)
F_B	Bandwidth pattern (Chapter 2)
$\mathcal{F}_R$, $\mathcal{F}_I$	Quantized fringe-rotation functions (Chapter 9)
g	Voltage gain constant for an antenna, gravitational acceleration (Chapter 13)
G	Gravitational constant
G_i	Power gain of receiver for one antenna (Chapter 7)
G_{mn}	Gain factor for a correlated antenna pair
G_0	Gain factor (Chapter 7)
$\mathcal{G}$	Occultation response function (Chapter 16)
h	Planck's constant, impulse response of a filter (Section 3.3), hour angle, height
h_0	Atmospheric scale height (Chapter 13)
H	Hour angle, voltage–frequency response, Hadamard matrix (Section 7.5)
H_0	Gain constant

i	Electric current
$\mathbf{i}$	Unit vector in direction of polar or azimuth axes (Chapter 4), current vector (Chapter 13)
I	Intensity, Stokes parameter
I^2	Variance of fractional frequency deviation (Chapter 9)
I_s	Speckle intensity (Chapter 16)
I_v	Stokes visibility
I_0	Peak intensity of a point-source map, derived (synthesized) intensity distribution, modified Bessel function of zero order (Chapters 6, 9)
I_1	One-dimensional intensity function, modified Bessel function of first order (Chapter 9)
$\mathcal{I}m$	Imaginary part
j	$\sqrt{-1}$
$\mathbf{J}$	Jones matrix (Chapter 4)
j_ν	Volume emissivity of a source (Chapter 13)
J	Mutual intensity (Chapter 14)
J_0	Bessel function of first kind and zero order
J_1	Bessel function of first kind and first order
k	Boltzmann's constant, propagation constant $2\pi/\lambda$ (Chapter 13)
$\mathbf{k}$	Propagation vector with magnitude $2\pi/\lambda$ (Chapter 9)
l	Direction cosine with respect to baseline component u, lapse rate (Chapter 13)
L	Length of a transmission line, loss factor in a transmission line (Chapter 7), probability integral [Eq. (8.70)], path length, likelihood function (Chapter 12), thickness of turbulent atmospheric layer or screen (Chapter 13)
L_{inner}, L_{outer}	Scales of turulence (Chapter 13)
ℓ	Length, galactic longitude (Chapter 13)
ℓ_λ	Unit spacing (in wavelengths) in a grating array (Chapters 1, 5)
$\mathcal{L}$	Latitude, *excess* path length (Chapter 13)
$\mathcal{L}_D$, $\mathcal{L}_V$	Excess path length of dry air, water vapor
m	Direction cosine with respect to baseline compoment v, modulation index (Appendix 7.2), measured quantity (Appendix 12.1), electron mass (Chapter 13)
m_ℓ, m_c, m_t	Degree of linear, circular, and total polarization
M	Frequency multiplication factor (Chapter 9), model function (Chapter 10), mass, complex degree of linear polarization (Chapter 13)
$\mathcal{M}$, $\mathcal{M}_D$, $\mathcal{M}_V$	Molecular weight; total, dry air, water vapor (Chapter 13)

n	Direction cosine with respect to baseline component w, weighting factor in quantization (Chapter 8), noise component, index of refraction (Chapter 13)
$n = n_R + jn_I$	Complex refractive index
n_a	Number of antennas
n_d	Number of data points
n_e, n_i, n_n, n_m	Density of electrons, ions, neutral particles, and molecules (Chapter 13)
n_p	Number of antenna pairs
n_s	Number of sources
n_r	Number of points in a rectangular array (grid points)
n_0	Refractive index at earth's surface (Chapter 13)
N	Number of samples (Chapter 8), total refractivity (Chapter 13)
N_b	Number of bits per sample (Chapter 8)
N_D, N_V	Refractivity of: dry air, water vapor (Chapter 13)
N_N	Number of Nyquist-rate samples (Chapter 8)
$\mathcal{N}$	$2\mathcal{N}$ and $(2\mathcal{N} + 1)$ Are even and odd numbers of quantization levels (Chapter 8)
p	Probability density or probability distribution [i.e., $p(x)\,dx$ is the probability that the random variable lies between x and $x + dx$], bivariate normal probability function (Chapter 8), number of model parameters (Chapter 10), partial pressure (Section 13.1), impact parameter (Section 13.5)
p_D	Partial pressure of dry air (Chapter 13)
p_V	Partial pressure of water vapor (Chapter 13)
P	Power, cumulative probability, total atmospheric pressure (Chapter 13)
P_0	Atmospheric pressure at earth's surface (Chapter 13)
$\mathbf{P}$	Dipole moment per unit volume
P_3	Triple product (bispectrum)
P_{mnp}	Instrumental polarization factor
P_{ne}	Spectrum of electron density fluctuations
$\mathcal{P}$	Point-source response at moon's limb (Section 16.2), speckle point-spread function (Section 16.4)
q	Distance in (u, v) plane
q'	Distance in (u', v') plane
q_x, q_y	Components in the spatial frequency (cycles per meter) plane (Chapter 13)
Q	Stokes parameter, quality factor of a line or cavity (Section 9.5), number of quantization levels (Section 9.6)
Q_v	Stokes visibility

r	Correlator output, distance in the (l, m) plane, radial distance
$\mathbf{r}$	Position vector of antenna relative to center of earth
r_e	Classical electron radius (Chapter 13)
r_ℓ	Correlator output resulting from lower sideband
r_0	Radius of the earth (Chapter 13), Fried length (Chapter 16)
r_u	Correlator output resulting from upper sideband
R	Autocorrelation function, correlator output, Rademacher function (Section 7.5), distance, gas constant (Chapter 13)
$\mathbf{R}$	Correlator output matrix (Chapter 4)
R_a	Response with visibility averaging (Chapter 6)
R_b	Response with finite bandwidth (Chapter 6)
R_e	Radius of electron orbit (Chapter 13)
R_{ff}	Far-field distance
R_m	Rotation measure (Chapter 13), distance of the moon's limb (Chapter 16)
R_n	Autocorrelation for n-level quantizarion (Chapter 8)
R_y	Autocorrelation function of fractional frequency deviation (Chapter 9)
R_ϕ	Autocorrelation function of phase (Chapters 9, 13)
$\mathcal{R}e$	Real part
$\mathcal{R}_{\mathrm{sn}}$	Signal-to-noise ratio
s	Signal component, smoothness measure (Chapter 11)
$\mathbf{s}$	Unit position vector (Chapter 3)
$\mathbf{s}_0$	Unit position vector of field center (Chapter 3)
S	(Spectral) power flux density (W m^{-2} Hz^{-1})
S_c	Flux density of a calibrator
S_E	System equivalent flux density
S_h	Threshold of harmful interference (W m^{-2}Hz^{-1}) (Chapter 15)
$\mathcal{S}$	Cross power spectrum (Chapter 9)
$\mathcal{S}_I$	Power spectrum of intensity fluctuations (Chapter 13)
$\mathcal{S}_y, \mathcal{S}'_y$	Single-sided and double-sided power spectra of fractional frequency deviation (single-sided power spectrum used only in Section 9.4)
$\mathcal{S}_\phi$, $\mathcal{S}'_\phi$	Single-sided and double-sided power spectra of phase fluctuations (single-sided power spectrum used only in Section 9.4)
$\mathcal{S}_2$	Two-dimensional power spectrum of phase (Chapter 13)
t	Time
t_e	Period of the earth's rotation (Chapter 12)

t_{cyc}	Cycle period for target and calibrator sources
T	Temperature, time interval, transmission factor (Chapter 14)
T_{at}	Atmospheric temperature (Chapter 13)
T_A	Component of antenna temperature resulting from wanted signal
T'_A	Total antenna temperature
T_B	Brightness temperature
T_c	Noise temperature of calibration signal
T_g	Gas temperature (Chapter 9)
T_R	Receiver temperature
T_S	System temperature
$\mathcal{T}$	Time interval
u	Antenna spacing coordinate in units of wavelength (spatial frequency)
u'	Projection of u coordinate onto the equatorial plane
U	Stokes parameter
U_v	Stokes visibility
$\mathcal{U}$	Unwanted response (Section 7.5)
v	Antenna spacing coordinate in units of wavelength (spatial frequency), phase velocity in a transmission line (Chapter 8)
v'	Projection of v coordinate onto the equatorial plane
v_g	Group velocity (Chapter 13)
v_m	Rate of angular motion of moon's limb (Chapter 16)
v_p	Phase velocity (Chapter 13)
v_r	Radial velocity
v_s	Velocity of scattering screen (parallel to baseline, if relevant) (Chapters. 12, 13)
v_0	Quantization level (Chapter 8), particle velocity (Chapter 9)
V	Voltage, Stokes parameter
V_A	Voltage response of an antenna
V_v	Stokes visibility (Chapter 4)
$\mathcal{V}$, $\boldsymbol{\mathcal{V}}$	Complex visibility, vector visibility
$\mathcal{V}_m$	Measured complex visibility
$\mathcal{V}_M$	Michelson's fringe visibility
$\mathcal{V}_N$	Normalized complex visibility
w	Antenna spacing coordinate in units of wavelength (spatial frequency), weighting function, column height of precipitable water (Chapter 13)

w'	w coordinate measured in the polar direction
w_a	Atmospheric weighting function (Chapter 13)
w_{mean}	Mean of weighting factors (Chapter 6)
w_{rms}	Root-mean-square of weighting factors (Chapter 6)
w_t	Visibility tapering function (Chapter 10)
w_u	Function that adjusts visibility amplitude for effective uniform weighting (Chapter 10)
W	Spectral sensitivity function (spatial transfer function); propagator (Chapter 14)
x	General position coordinate, coordinate in antenna aperture, signal voltage
x_λ	x coordinate measured in wavelengths
X	Coordinate of antenna spacing [see Eq. (4.1)], signal waveform measured in units of rms amplitude (Section 8.4), coordinate within a source or an aperture (Chapters 3, 14), signal spectrum (Section 8.7)
X_λ	X coordinate measured in wavelengths
y	General position coordinate, coordinate in antenna aperture, signal voltage, distance along a ray path (Chapter 13)
y_k	Fractional frequency deviation (Chapter 9)
y_λ	y coordinate measured in wavelengths
Y	Coordinate of antenna spacing [Eq. (4.1)], Y factor (Chapter 7), coordinate within a source or aperture (Chapters 3, 14) signal waveform measured in units of rms amplitude (Section 8.4), signal spectrum (Section 8.7)
Y_λ	Y coordinate measured in wavelengths
z	General position coordinate, signal voltage, zenith angle (Chapter 13)
z_λ	z coordinate measured in wavelengths
Z	Coordinate of antenna spacing [Eq. (4.1)], visibility plus noise in correlator output (Chapters 6, 9)
$\mathcal{Z}_D$, $\mathcal{Z}_V$	Compressibility factors for dry air and water vapor (Chapter 13)
$\mathbf{Z}$	Visibility-plus-noise vector (Chapters 6, 9)
Z_λ	Z coordinate measured in wavelengths
α	Right ascension, power attenuation coefficient, quantization threshold in units of σ (Chapter 8), spectral index (Chapter 11), absorption coefficient and power-law exponent in Table 13.2 and related text (Section 13.1), exponent in electron density fluctuation [Eq. (13.164)] (Section 13.4), solar elongation (Section 13.5)

β	Fractional length change in transmission line (Chapter 7), oversampling factor (Chapter 8), exponent of distance in rms phase fluctuation [Eq. (13.79a)] (Sections 12.2, 13.1), exponent in solar electron density [Eq. (13.178)] (Section 13.5), Faraday depth (Section 13.6)
γ	Instrumental polarization factor (Section 4.8), maser relaxation rate (Chapter 9), loop gain in CLEAN (Chapter 11), source coherence function (Chapter 14)
Γ	Damping factor (Chapter 13), mutual coherence function (Chapter 14), gamma function
Γ_{12}	Mutual coherence function (Chapter 14)
δ	Declination, increment prefix, (Dirac) delta function, instrumental polarization factor (Section 4.8)
${}^2\delta$	Delta function in two dimensions
Δ	Small length, increment prefix
$\Delta\nu$	Bandwidth, Doppler shift (Appendix 10.2)
$\Delta\nu_{\mathrm{IF}}$	Intermediate-frequency bandwidth
$\Delta\nu_{\mathrm{LF}}$	Low-frequency bandwidth
$\Delta\nu_{\mathrm{LO}}$	Frequency difference of local oscillators
$\Delta\tau$	Delay error
Δu, Δv	Increments in (u, v) plane
Δl, Δm	Increments in (l, m) plane
ϵ	Width of quantization level in units of σ (Chapter 8), noise component in IF signal (Chapter 9), permittivity (Chapter 13)
ϵ_a	Amplitude error (Chapter 11)
ϵ_0	Permittivity of free space (Chapter 13)
ε	Noise component of correlator output (Chapters 6, 9), residual, error component, dielectric constant (Chapter 13)
$\boldsymbol{\varepsilon}$	Noise vector (Chapter 6)
η	Loss factor
η_D	Discrete delay step loss factor
η_Q	Efficiency (loss) factor for Q-level quantization
η_R	Fringe rotation loss factor
η_S	Fringe sideband rejection loss factor
θ	General angle, angle measured from a plane normal to the baseline, instrumental phase angle
θ_0	Angular position of source or field center
θ_b	Width of synthesized beam, bending angle (Chapter 13)
θ_f	Width of synthesized field
θ_F	Width of first Fresnel zone
θ_{LO}	Local oscillator phase

θ_m, θ_n	Local oscillator phase at antennas m and n (Chapter 6)
θ_s	Effective beamwidth resulting from atmospheric fluctuations (Chapter 13), width of source (Chapter 16)
Θ	Variation in earth-rotation angle (UT1–UTC) (Chapter 12)
λ	Wavelength
λ_{opt}	Wavelength of optical carrier (Appendix 7.2)
Λ	Reflected amplitude in a transmission line (Chapter 7)
μ	Power-law exponent in Allan variance (Chapter 9)
ν	Frequency
ν'	Frequency measured with respect to center frequency or local oscillator frequency (Chapter 9)
ν_b	Bit rate
ν_{B}	Gyrofrequency (Chapter 13)
ν_c	Collision frequency (Chapter 13)
ν_C	Cavity frequency (Chapter 9)
ν_d	Intermediate frequency at which delay is inserted
ν_{ds}	Delay step frequency (Chapter 9)
ν_f	Fringe frequency
ν_{in}	Instrumental component of fringe frequency (Chapter 12)
ν_{IF}	Intermediate frequency
ν_{LO}	Local oscillator frequency
ν_ℓ	Frequency of a correlator channel (Chapter 9)
ν_m	Frequency of modulation on optical carrier (Chapter 7)
ν_{RF}	Radio frequency
ν_{opt}	Frequency of optical carrier (Appendix 7.2)
ν_p	Plasma frequency (Chapter 13)
ν_0	Center frequency of an IF or RF band, frequency of absorption peak (Chapter 13)
ρ	Autocorrelation function, cross-correlation coefficient, reflection coefficient (Chapter 7), gas density (Chapter 13)
ρ_D, ρ_V, ρ_T	Density: dry air, water vapor, total (Chapter 13)
ρ_{mn}	Cross-correlation
ρ_σ	Area density in the (u, v) plane (Chapter 10)
ρ_m, ρ_n	Reflection coefficients in transmission line (Chapter 7)
ρ_w	Density of water (Chapter 13)
σ	Standard deviation, rms noise level
$\boldsymbol{\sigma}$	Position vector on the unit sphere
σ_y	Allan standard deviation (σ_y^2 = Allan variance)
σ_τ	Root-mean-square uncertainty in delay (Chapter 9)
σ_ϕ	Root-mean-square deviation of phase

τ	Time interval
τ_a	Averaging (integration) time
τ_{at}	Atmospheric delay error (Chapter 12)
τ_c	Coherent integration time (Chapter 9)
τ_e	Clock error
τ_g	Geometric delay
τ_i	Instrumental delay
τ_0	Unit increment of instrumental delay, duration of an observation (Chapter 6), zenith optical depth (opacity) of the atmosphere (Chapter 13)
τ_s	Sampling interval in time
τ_{or}	Minimum period of orthogonality (Chapter 7)
τ_{sw}	Interval between switch transitions (Chapter 7)
τ_v	Optical depth (opacity) (Chapter 13)
ϕ	Phase angle
ϕ_m	Phase of signal received by antenna m
ϕ_v	Visibility phase
ϕ_G, ϕ_{in}	Instrumental phase for correlated antenna pair
ϕ_{pp}	Peak-to-peak phase error (Chapter 9)
Φ	Phase of a complex signal (Appendix 3.1), probability integral [Eq. (8.44)] (Chapter 8), phase of a signal (Section 13.1)
χ	Arctangent of axial ratio of polarization ellipse
χ^2	Statistical parameter
ψ	Position angle, phase angle
ψ_p	Parallactic angle
ω_e	Angular rotation velocity of the earth
Ω	Solid angle
Ω_s	Solid angle subtended by source
Ω_0	Solid angle of main lobe of synthesized beam

FREQUENTLY USED SUBSCRIPTS

1, 2	Antenna designation
2, 3, 4, ∞	Quantization levels (Chapter 8)
A	Antenna
d	Delay, double sideband
D	Dry component (Chapter 13)
I	Imaginary part
IF	Intermediate frequency

ℓ	Left circular polarization, lower sideband
LO	Local oscillator
0	Center of frequency band or angular field, earth's surface (Chapter 13)
m, n	Antenna designation
N	Normalized, Nyquist rate (Sections 8.2, 8.3)
r	Right circular polarization
R	Real part
S	System
u	Upper sideband
V	Water vapor (Chapter 13)
λ	Measured in wavelengths

OTHER SYMBOLS

Π	Unit rectangle function
$\prod$	Product symbol
III	Shah function in one dimension
^{2}III	Shah function in two dimensions
$\rightleftharpoons$	"Is the Fourier transform of"
$*$	Convolution in one dimension
$**$	Convolution in two dimensions
$\star$	Cross-correlation in one dimension
$\star\star$	Cross-correlation in two dimensions
$\langle\ \rangle$	Expectation (or approximation by a finite average)
Dot ($\dot{}$)	First derivative with respect to time
Double dot ($\ddot{}$)	Second derivative with respect to time
Overline ($\bar{}$), ($\bar{\ }$)	Average (Chapters 1, 9, Section 14.1); Fourier transform of function (Chapters 3, 5, 8, 10, 11, 13, Section 14.2)
Circumflex ($\hat{}$)	Quantized variable (Chapter 8)
Circumflex ($\widehat{\ }$)	Function of frequency (Chapter 3)

FUNCTIONS

For definitions, see, for example, Abramowitz, M., and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1964, reprinted by Dover, New York, 1965.

erf	Error function
J_0	Bessel function of first kind and zero order

J_1	Bessel function of first kind and first order
I_0	Modified Bessel function of zero order
I_1	Modified Bessel function of first order
Γ	Gamma function [Note that $\Gamma(x+1) = x\Gamma(x)$]
δ	Dirac delta function

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

AUTHOR INDEX

A reference to a footnote of a page or table is indicated by *n*.

Aarons, J., 562
Ables, J. G., 432, 433
Abramowitz, M., 255
Adatia, N. A., 124
Agrawal, G. P., 219*n*
Alder, B., 40
Allan, D. W., 334, 353, 376
Alef, W., 477
Allen, C. W., 572
Allen, L. R., 220, 647
Allen, R. J., 644
Altenhoff, W. J., 539
Ananthakrishnan, S., 580
Anantharamaiah, K. R., 187, 576, 607, 609, 610
Anderson, B., 231
Anderson, S. B., 260, 276
Apostol, T. M., 71*n*
Appleton, E. V., 16, 554
Archer, J. W., 218, 249
Armstrong, J. T., 644
Armstrong, J. W., 249, 538, 540*n*
Arsac, J., 142
Ash, M. E., 419
Ashby, N., 353, 376
Askne, J. I. H., 543
Assus, P., 644
Audoin, C., 335
Azoubib, J., 353

Baade, W., 21
Baars, J. W. M., 32, 144, 212, 420, 540*n*
Backer, D. C., 35, 475, 583
Bagri, D. S., 238, 353
Bailey, D. K., 559
Bajaja, E., 451
Balanis, C. A., 164
Baldwin, J. E., 439, 445, 460, 583, 641, 644
Ball, J. A., 290, 419
Bally, J., 545*n*
Bare, C., 34, 356
Barnbaum, C., 615
Barnes, J. A., 332, 333, 339*n*, 344
Barnett, M. A. F., 554
Bartel, N., 469, 477
Bates, R. H. T., 445, 637, 642, 645, 648
Batty, M. J., 212
Bean, B. R., 511, 528
Beasley, A. J., 477, 478
Beauchamp, K. G., 242
Becker, R. H., 402
Belcora, L., 38
Bennett, A. S., 25
Bennett, J. C., 637
Benson, J. A., 643
Benson, J. M., 290
Beran, M. J., 604, 608, 611
Berkeland, D. J., 346
Berkner, L. V., 40
Bernier, L. G., 343
Bester, M., 642, 644
Betz, A. L., 644
Bevington, P. R., 404, 490
Beynon, W. J. G., 554
Bieging, J. H., 540*n*
Bignell, R. C., 106, 115
Bilitza, D., 560
Biraud, F., 40, 377
Blackman, R. B., 287
Blair, B. E., 332
Blake, G. A., 7
Block, W. F., 34
Blum, E. J., 27, 187
Blythe, J. H., 25, 137
Boboltz, D. A., 486
Bohlander, R. A., 507
Boischot, A., 27
Böker, T., 644
Bolton, J. G., 18, 19, 21
Booker, H. G., 562, 564, 600
Booth, R. S., 486, 502

Borella, M. S., 219
Boriakoff, V., 577
Born, M., 69*n*, 82, 83, 97, 594, 607, 611
Bos, A., 405
Bourgois, G., 575, 580, 582
Bowers, F. K., 260, 269, 270, 278
Bowyer, S., 499
Braccesi, A., 26
Bracewell, R. N., 27, 40, 41, 59, 61, 65, 66, 69*n*, 75, 82, 85, 126, 127, 128, 129, 133, 135, 142, 143, 145, 146, 201, 212, 256, 283, 388, 391, 392, 393, 403, 412, 422, 427, 451, 458, 533, 574, 599, 600, 631, 644
Bradley, R. F., 615
Braude, S. Ya., 154
Braun, R., 432, 434, 435, 438
Bregman, J. D., 102, 109, 113, 114, 115, 287
Breit, G., 554
Bridle, A. H., 41, 413, 422
Briggs, D. S., 392, 397, 432, 434, 435, 438
Brigham, E. O., 128
Broten, N. W., 27, 34, 356
Brooks, J. W., 41, 114, 154
Brouw, W. N., 76, 144, 396, 450
Brown, G. W., 34
Brown, L. F., 115
Brown, R. L., 33
Bruck, Y. M., 445
Bryan, R. K., 434
Budden, K. G., 554
Burdick, H. M., 543
Burke, B. F., 34, 39, 40, 158
Burn, B. J., 579
Burns, W. R., 260, 264
Burton, W. B., 411
Butler, B., 548

Calame, O., 502
Callen, H. B., 215
Caloccia, E. M., 218
Campbell, R. M., 480, 579
Cannon, W. H., 40, 316, 356
Carilli, C. L., 41, 422, 539, 540*n*, 544, 551, 552, 553*n*
Carlson, B. R., 297, 377
Carpenter, J., 554
Carr, T. D., 34, 627
Carswell, R. F., 461
Carter, A. W. L., 401
Carter, W. E., 485, 574
Carver, K. R., 97
Cash, W., 645
Casse, J. L., 213
Caves, C. M., 646
Cernicharo, J., 547
Chamberlin, R. A., 545*n*, 548
Champeney, D. C., 59
Chandler, S. C., 482
Charlot, P., 470
Chen, M. T., 636
Chesalin, L. S., 356*n*
Chi, A. R., 377
Chicada, Y., 290
Chie, C. M., 352
Chivers, H. J. A., 562
Chow, Y. L., 149, 150
Christiansen, W. N., 26, 27, 30, 41, 421
Christou, J. C., 648
Chu, T.-S., 124
Chylek, P., 553
Clark, B. G., 28, 182, 247, 248, 326, 356, 358, 431
Clark, T. A., 35, 347, 353, 356, 485
Clarke, M., 574
Clegg, J. A., 41
Clemence, G. M., 481
Clemmow, P. C., 600
Clifford, S. F., 529
Cocke, W. J., 420
COESA, 516, 519
Cohen, M. H., 35, 315, 366, 475, 570, 575, 576, 580, 634, 636
Colavita, M. M., 642, 644
Cole, T., 260, 422
Coles, W. A., 575, 580, 582
Collin, R. E., 164
Colvin, R. S., 28, 181, 187
Combrinck, W. L., 624
Condon, J. J., 9
Conway, J. E., 453, 454, 477, 478
Conway, R. G., 4, 104, 105, 115
Cooper, B. F. C., 260, 264, 269, 270
Cordes, J. M., 566, 569, 577*n*, 579, 580, 583, 610, 611
Cornwell, T. J., 41, 151, 164, 431, 432, 434, 435, 437, 438, 441, 442, 444, 448, 449, 450, 452, 453, 454, 455, 456, 457, 458, 576, 607, 609, 610
Cotton, W. D., 115, 326, 327, 328, 439
Coulman, C. E., 534, 538
Counselman, C. C., III, 480, 485, 572
Covington, A. E., 27
Cowan, J. J., 33, 443
Cox, A. N., 419*n*, 529
Crane, P. C., 398, 399
Crane, R. K., 507, 512, 562
Crawford, D. L., 626
Cronyn, W. M., 570, 580
Currie, D. G., 644
Cutler, L. S., 351

D'Addario, L. R., 210, 234, 238, 242, 277, 279, 280, 281, 296, 355, 374, 376, 405, 422, 433, 637, 638
Dainty, J. C., 648
Danchi, W. C., 644
Daniell, G. J., 433, 434
Daniell, R. E., 560
Das Gupta, M. K., 22, 23, 627
Davies, J. G., 231
Davies, K., 554
Davis, J., 642, 643, 644, 647, 648
Davis, J. L., 518
Davis, M. M., 345, 353
Davis, R. J., 502
Davis, W. F., 278
de Graauw, T., 647
de Vegt, C., 469, 502
Debye, P., 528
Delgado, G., 545*n*
DeJong, M. L., 636
Dennison, B., 580
Dewdney, P. E., 297
Dewey, R. J., 371
Dhawan, V., 315
Diamond, P. J., 115, 377, 486
Dicke, R. H., 11
Dillinger, W. H., 574
Doeleman, S. S., 329*n*, 330, 331
Douglas, J. N., 200
Downes, D., 421, 539, 540*n*
Drane, C. J., 604, 611
Dreher, J. W., 33, 443
Dreyer, J. L. E., 9
Drullinger, R. E., 346
Duffet-Smith, P. J., 580
Dulk, G. A., 627
Dutta, P., 339
Dutton, E. J., 528

Eckart, A., 649
Eddington, A. S., 631
Edge, D. O., 9, 25
Edson, W. A., 332, 351
Ekers, R. D., 32, 148, 149, 151, 212, 282, 404, 406, 412, 442, 447, 459, 572
Elgaroy, O., 23
Elgered, G., 543
Elitzur, M., 6, 486
Ellithorpe, J. D., 462
Elsmore, B., 21, 31, 212, 468
Emerson, D. T., 245, 246
Enge, P., 377, 502
Erb, K., 155, 212
Erickson, W. C., 155, 212, 460, 572, 575
Escoffier, R. P., 297, 298
Eshleman, V. R., 574
Evans, D. S., 636
Evans, J. V., 554, 555*n*, 556, 560, 562, 564, 572
Evans, K. F., 434, 449

Fanselow, J. L., 316, 475, 485, 500
Farley, D. T., 260, 264, 266, 269, 271
Feierman, B. H., 636
Fejer, B. G., 562
Feldman, M. J., 40, 181, 216, 217
Felli, M., 377
Fenstermacher, D. L., 213
Fernbach, S., 40
Fey, A. L., 470
Fiebig, D., 115
Fiedler, R. L., 582
Fienup, J. R., 445
Findlay, J. W., 41
Fomalont, E. B., 28, 29, 109, 403, 469, 561, 571, 572
Forman, P., 347
Fort, D. N., 438
Fowle, F. F., 243
Frail, D. A., 368, 567
Frank, R. L., 353
Frater, R. H., 41, 114, 154, 220
Freeman, R. L., 542
Frehlich, R. G., 122
Fridman, P., 373
Fried, D. L., 536, 642
Frieden, B. R., 432, 433
Fugate, R. Q., 642, 650

Gabor, D., 82
Gallagher, J. J., 507
Galt, J., 624
Garcia-Barreto, J. A., 486
Gardner, C. S., 645
Gardner, F. F., 579
Gardner, F. M., 230, 376
Gauss, K. F., 491
Gaylord, M. J., 624
Geldart, D. J. W., 553
Genee, R., 41, 144
Genzel, R., 35, 486, 487
Gergely, T. E., 626
Gilbert, S. W., 353
Ginat, M., 27
Gold, B., 128
Gold, T., 35
Goldenberg, H. M., 348, 350, 351
Goldsmith, P. F., 41
Goldstein, H., 524, 542
Goldstein, R. M., 254
Goldstein, S. J., Jr., 200

Goodman, J. W., 82, 582, 607, 611, 648
Gordon, M. A., 419, 420
Gower, J. F. R., 26
Gradshteyn, I. S., 415
Graham-Smith, F., 40
Granlund, J., 228, 247, 248, 249, 293
Grossi, M. D., 560
Guilloteau, S., 33
Guinot, B., 482
Guiraud, F. O., 537, 543
Gull, S. F., 433
Gundermann, E. J., 576
Gupta, Y., 564
Gurwell, M., 416
Güsten, R., 115
Gwinn, C. R., 475, 481, 486, 500, 579, 580, 582

Haddock, F. T., 41
Hagen, J. B., 260, 264, 266, 269, 271
Hagfors, T., 38, 554, 555*n*, 556, 562, 572
Hall, T., 491
Hamaker, J. P., 102, 109, 113, 114, 115, 142, 538
Hamilton, W. C., 490, 497
Hanbury Brown, R., 16, 22, 23, 627, 631, 647
Haniff, C. A., 641
Hankins, T. H., 296
Hanson, D. W., 377, 502
Hanson, R. J., 435
Hargrave, P. J., 32
Harmuth, H. F., 242
Harris, D. E., 576, 580
Harris, F. J., 287
Harris, R. W., 213
Haskell, R. E., 557
Hazard, C., 25, 468, 632, 636
Heffner, H., 646
Heiles, C., 277, 577
Hellwig, H., 345*n*, 349
Henny, M., 647
Herbig, T., 125
Herring, T. A., 357, 362, 475, 481, 485, 500
Herrnstein, J. R., 37
Hess, S. L., 510, 512
Hewish, A., 24, 25, 137, 562, 563, 574, 576, 580
Hewitt, J. N., 462
Hey, J. S., 562
Hill, R. J., 520, 528, 529
Hills, R. E., 547, 554
Hinder, R. A., 539, 540*n*, 564
Hinteregger, H. F., 356, 475
Hinz, P. M., 644
Hirabayashi, H., 38, 158, 375, 377
Hirade, K., 624
Hjellming, R. M., 148, 162
Ho, C. M., 560, 562
Hocke, K., 479, 563
Högbom, J. A., 41, 144, 427, 428, 430, 433
Hogg, D. C., 537, 543
Hogg, D. E., 31
Holdaway, M. A., 164, 449, 452, 539, 540*n*, 544, 545*n*, 550, 551, 552, 553*n*
Hollweg, J. V., 574
Holt, E. H., 557
Holt, J. M., 564
Hooghoudt, B. G., 144
Hopkins, A., 32
Horn, P. M., 339
Howard, J., 543
Hueckstaedt, R. M., 527
Hughes, M. P., 28
Humphreys, W. J., 510
Hunsucker, R. D., 563

IAU, 9, 100, 101
IEEE, 6, 100
Imbriale, W. A., 164
Ingalls, R. P., 125
Inoue, M., 377
Ishiguro, M., 41, 142, 540*n*, 547
Itano, W. M., 346
ITU-R, 614, 615, 616, 625, 626

Jackson, J. D., 524*n*
Jacobs, C. S., 316, 485
Jacobs, I. M., 184*n*
Jacquinot, P., 392
Jaeger, J. C., 573
James, G. L., 117
Jansky, K. G., 16
Janssen, M. A., 583
Jasik, H., 164
Jaynes, E. T., 433
Jenet, F. A., 260, 276
Jennison, R. C., 22, 23, 400, 627
Jespersen, J., 377, 502
Johnson, D. R., 5
Johnson, M. A., 644
Johnson, R. C., 164
Johnston, K. J., 469, 480, 502
Jones, R. C., 109

Kahlmann, H. C., 626
Kalachev, P. D., 26
Kanzawa, T., 540*n*, 547
Kaplan, G. H., 475, 561
Kardashev, N. S., 34
Kartashoff, P., 344
Kassim, N. E., 460
Kasuga, T., 540*n*, 547
Kawabe, R., 540*n*

Kawaguchi, N., 356
Keihm, S. J., 520
Kelder, H., 563
Kellermann, K. I., 4, 156, 611
Kelley, M. C., 562
Kemball, A. J., 115
Kenderdine, S., 212
Kerr, A. R., 40, 216, 217
Kerr, F. J., 419*n*
Keshner, M. S., 339
Kesteven, M. J. L., 106, 107, 114, 115, 118
Keto, E., 140, 150, 151, 152, 154, 246
Killeen, E. B., 106, 107, 114, 115, 118
Klemperer, W. K., 306
Klepczynski, W. J., 353
Kleppner, D., 348, 350, 351
Klingler, R. J., 260, 269, 270
Knapp, S. L., 644
Knowles, S. H., 34
Knox, K. T., 650
Ko, H. C., 99
Kobayashi, H., 377
Kochanek, C. S., 461, 462
Kogan, L. R., 132, 356*n*
Kojima, M., 122
Kokkler, A. B. J., 373
Koles, W. A., 122
König, A., 72
Kraus, J. D., 41, 97, 214
Krichbaum, T., 41
Krishnan, T., 27
Kronberg, P. P., 104, 105, 115, 116, 578
Kroupa, V. F., 377
Kulkarni, S. R., 277, 330
Kundu, M. R., 571

Labeyrie, A., 644, 649
Labrum, N. R., 27
Lambeck, K., 481
Lampton, M., 499
Lane, A. P., 545*n*
Lanyi, G. E., 537
Latham, V., 22
Lawrence, C. R., 125
Lawrence, R. S., 528, 562
Lawson, C. L., 435
Lawson, J. L., 184*n*
Lawson, P. R., 642, 650
Lay, O. P., 537, 550, 554
Lazio, T. J. W., 566
Lebach D. E., 573
Leech, J., 143
Leick, A., 353
Léna, P. J., 650
Lesage, P., 335
Lestrade, J.-F., 470, 475
Levine, M. W., 351
Levy, G. S., 38
Lewandowski, W., 353
Lewis, L. L., 342
Liebe, H. J., 518, 519, 520, 526, 527, 543, 544
Lieske, J. H., 481
Liewer, K. M., 644
Lightman, A. P., 3, 4, 417, 521
Lindsey, W. C., 352
Lineweaver, C. H., 419
Linfield, R., 37, 38
Little, A. G., 26, 229, 401
Little, C. G., 562
Little, L. T., 576
Liu, C. H., 562
Lo, K. Y., 580, 581
Lo, W. F., 175, 283
Lohmann, A. W., 650
Long, R. J., 4
Longair, M. S., 3
Lonsdale, C. J., 446
Louden, R., 39, 524
Lovas, F. J., 5
Love, A. W., 164
Lovelace, R. V., 569, 579, 580, 611
Lovell, A. C. B., 41, 562
Lynden-Bell, D., 419*n*
Lyne, A. G., 579

McCarthy, D. D., 484, 502
McClean, D. J., 458
McCready, L. L., 18
McGilchrist, M. M., 458
MacKay, J. R., 485
McKenzie, A. A., 353
MacMahon, P. A., 631
McMillan, R. W., 507
MacPhie, R. H., 607, 611, 644
Ma, C., 9, 35, 470, 475, 480, 500
Mackey, M. B., 632
Mahoney, M. J., 155, 212
Maltby, P., 28, 403
Mandel, L., 611
Mannucci, A. J., 560
Marcaide, J. M., 477
Margon, B., 499
Marini, J. W., 518
Markowitz, W., 482
Marscher, A. P., 28
Martin, D. H., 208
Masson, C. R., 402, 530, 537, 538, 540*n*, 545*n*, 546, 547, 549, 638, 641
Mathur, N. C., 80, 149, 150, 155, 200, 560, 603
Matsakis, D. N., 486

Matsushita, S., 547
Mattes, H., 41
Matveenko, L. I., 34
Mauzy, R. E., 249
Mayer, C. E., 125
Meeks, M. L., 41
Meinel, A. B., 370
Melchior, P., 485
Mercier, R. P., 568
Michelson, A. A., 13
Mickelson, R. L., 198
Middleton, D., 184*n*, 262*n*
Milligan, T. A., 164
Mills, B. Y., 21, 22, 24, 26, 137
Milonni, P. W., 642
Miner, G. F., 543
Minkowski, R., 21
Misell, D. L., 641
Misner, C. W., 574
Misra, P., 377, 502
Miyoshi, M., 37
Moffet, A. T., 28, 142, 143, 403
Moore, E. M., 28
Moran, J. M., 6, 33, 34, 38, 155, 194, 292, 315, 316, 329*n*, 330, 331, 340, 345, 356, 372, 373, 486, 488, 502, 538, 541, 543, 562, 571
Morison, I., 213
Morita, K.-I., 33
Morris, D., 24, 102, 638, 641
Morris, D., 377
Morrison, N., 500
Mould, J. R., 420
Mourard, D., 644
Mueller, H., 112
Mueller, I. I., 469
Muhleman, D. O., 572
Mullaly, R. F., 27
Murota, K., 624

Napier, P. J., 32, 35, 148, 149, 151, 156*n*, 212, 282, 307, 377, 398, 399, 435, 442, 445, 637
Narayan, R., 433, 434, 461, 462, 570, 576, 582, 583, 607, 609, 610
NASA, 502
Neville, A. C., 31
Newhall, X X, 419
Niell, A. E., 518
Nityananda, R., 40, 433, 434
Norris, R. P., 486
NRAO, 148, 228, 540*n*
Nyquist, H., 10, 213, 256*n*

Okumura, S. K., 290
Oliver, B. M., 40
Olmi, L., 540*n*
O'Neill, E. L., 112
Oosterloo, T. A., 427
Oppenheim, A. V., 128, 256
O'Sullivan, J. D., 293
Otter, M., 624
Owens, J. C., 528

Padin, S., 160, 220
Paine, S., 547
Paley, R. E., 244
Palmer, H. P., 23
Pan, S.-K., 40, 217
Pankonin, V., 626
Papoulis, A., 128, 193, 317
Pardo, J. R., 547
Parkinson, B. W., 353
Parrent, G. B., Jr., 604, 606, 608, 609, 611
Parsons, S. J., 562
Patnaik, A. R., 480
Pauliny-Toth, I. I. K., 4
Pawsey, J. L., 18, 25, 41
Payne, J. M., 208, 213, 230
Payne-Scott, R., 18
Pearlman, M. R., 560
Pearson, T. J., 35, 36, 403, 441
Pease, F. G., 13, 16
Peebles, P. J. E., 420
Percival, B. D., 292
Perley, R. A., 33, 41, 109, 422, 443, 445, 455, 456, 457, 458, 475
Petley, B. W., 467
Phillips, J. A., 38
Phillips, J. W., 562
Phillips, T. G., 213, 638, 641
Pi, X., 562
Picken, J. S., 27
Pidwerbetsky, A., 569, 579, 611
Pierce, J. A., 353
Pilkington, J. D. H., 484, 502
Pol, S. L. C., 520
Pollak, H. O., 397
Ponsonby, J. E. B., 432
Pospieszalski, M. W., 213
Press, W. H., 339, 499
Preston, R. A., 158, 159
Price, R., 266
Price, R. M., 626
Priestly, J. T., 528
Puplett, E., 208
Purcell, E. M., 524, 647

Quirrenbach, A., 650

Rabiner, L. R., 128
Rademacher, H., 153

Radford, S. J. E., 547, 548
Radhakrishnan, V., 28, 40, 102
Ragland, S., 636
Raimond, E., 41, 103, 144
Räisänen, A. V., 11, 191, 213
Ramsey, N. F., 348, 350, 351
Rankin, J. M., 572, 577
Ratcliffe, J. A., 554, 558, 564, 568
Rawer, K., 554
Ray, P. S., 542
Read, R. B., 28, 181, 212
Readhead, A. C. S., 125, 160, 438, 439, 441, 580
Reasenberg, R. D., 650
Reber, E. E., 512
Reber, G., 16
Reid, M. J., 6, 35, 315, 406, 476, 486, 502
Reid, M. S., 213
Reiland, G., 547
Resch, G. M., 543
Rhodes, D. R., 397
Rice, S. O., 193
Richichi, A., 636
Rickett, B. J., 579, 580, 582, 583
Riddle, A. C., 66, 422
Roberts, D. H., 115, 564
Roberts, J. A., 61, 127, 133, 392, 427, 462
Robertson, D. S., 485, 574
Robertson, J. G., 650
Robinson, D. K., 404, 490
Roddier, F., 534
Roelfsema, P., 410, 411
Roger, R. S., 28
Rogers, A. E. E., 35, 195, 311, 329*n*, 330, 331, 340, 345, 356, 365, 367, 438, 439, 538, 543, 640
Roggemann, M. C., 642, 650
Rogstad, D. H., 644
Rohlfs, K., 5, 41, 97
Roizen-Dossier, B., 392
Rolston, S. L., 346
Romney, J. D., 296
Rönnäng, B. O., 543
Ros, E., 477, 480
Rosen, B. R., 541, 543
Rosenkranz, P. W., 553
Rotenberg, M., 40
Rots, A. H., 447
Rowson, B., 24, 31, 88
Rudge, A. W., 124
Ruf, C. S., 520
Rutman, J., 332, 333
Ruze, J., 125
Rybicki, G. B., 3, 4, 417, 521
Ryle, M., 12, 17, 18, 20, 21, 24, 25, 30, 31, 32, 137, 143, 212, 468, 540*n*, 564, 637
Ryzhik, I. M., 415

Saastamoinen, J., 516, 518
Sakurai, T., 116, 575
Salpeter, E. E., 575
Sault, R. J., 102, 106, 107, 109, 113, 114, 115, 118, 427, 450, 454
Schafer, R. W., 128, 256
Schaper, L. W., Jr., 541
Scheuer, P. A. G., 564, 632, 634
Schilke, P., 7
Schlegel, K., 479, 563
Schuss, J. J., 624
Schwab, F. R., 41, 300, 326, 327, 328, 392, 394, 395, 397, 422, 432, 438, 441, 455
Schwarz, U. J., 429
Scott, P. F., 26, 574, 637
Scott, S. L., 575
Scoville, N. J., 33, 554
Searle, C. L., 351
Seidelmann, P. K., 9, 91, 469, 481
Seielstad, G. A., 102, 156
Serabyn, E., 547, 638, 641
Serna, R., 218
Shaffer, D. B., 315, 475
Shakeshaft, J. R., 24, 25, 137
Shannon, C. E., 256*n*
Shao, M., 642, 644
Shapiro, I. I., 316, 475, 477, 481, 485, 500, 561, 571, 574
Shillue, B., 547
Shimmins, A. J., 632
Shimoda, K., 348
Shinn, D. H., 564
Shklovsky, I. S., 41
Sholomitskii, G. B., 34
Sieber, W., 580
Siegman, A. E., 350
Silver, S., 599
Simard-Normandin, M., 578
Sinclair, M. W., 212
Skilling, J., 434
Slee, O. B., 18, 24
Slepian, D., 392, 397
Smart, W. M., 515, 572
Smegal, R. J., 115
Smith, E. K., Jr., 529
Smith, F. G., 21, 22, 401, 468, 562, 579
Smith, H. M., 482
Smolders, A. B., 32
Smoot, G. F., 11
Snider, J. B., 543
Snyder, L. E., 5

Sodin, L. G., 445
Solheim, F., 543
Southworth, G. C., 16
Sovers, O. J., 316, 485
Spangler, S. R., 116, 575, 579
Spencer, R. E., 377
Spitzer, L., 577
Spoelstra, T. A. T., 560, 563
Sramek, R. A., 35, 392, 397, 455, 475, 479, 538, 540*n*, 561, 571
Staelin, D. H., 520, 541, 542, 644
Standish, E. M., 419
Stanley, G. J., 18, 19, 21
Stark, A. A., 545*n*
Stavely-Smith, L., 450
Stegun, I. A., 255
Stutzman, W. L., 164
Subrahmanya, C. R., 433
Subramanian, S., 534
Sullivan, W. T., III, 41
Sutton, E. C., 527, 534, 647
Swarup, G., 27, 32, 155, 163, 222
Sweezy, W. B., 537
Swenson, G. W., Jr., 80, 155, 156, 198, 200, 603, 626, 645
Swope, J. R., 512

Taff, L. G., 481
Tahmoush, D. A., 543
Tango, W. J., 642, 643, 644, 650
Tatarski, V. I., 479, 534, 535, 583
Taylor, G. B., 41, 422
Taylor, G. I., 537
Taylor, J. H., 469, 580, 636
Taylor, J. R., 404
Têtu, M., 343
Thayer, G. D., 528
Thiele, G. A., 164
Thomas, C., 353
Thomas, J. B., 357
Thomasson, P., 24, 155, 444
Thompson, A. R., 4, 23, 27, 28, 32, 129, 145, 146, 148, 149, 151, 210, 212, 231, 234, 238, 240, 247, 248, 273, 282, 353, 388, 393, 422, 442, 616, 620, 621, 626
Thompson, B. J., 650
Thompson, M. C., 222
Thorburn, M., 164
Thorne, K. S., 574
Thornton, D. D., 451, 543
Tirupati, S. K., 639
Tiuri, M. E., 11, 187, 191, 213
Toeplitz, O., 153
Townes, C. H., 348, 534, 644, 645, 647
Treuhaft, R. N., 537
Tucker, J. R., 211, 216, 217
Tukey, J. W., 287
Turrin, R. H., 124
Tuve, M. A., 554
Twiss, R. Q., 22, 401, 631, 642, 647

Uhlenbeck, G. E., 184*n*
Unser, M., 127
USAF, 221
Uson, J. M., 164, 449, 452

van Albada, G. D., 451
van Ardenne, A., 373
van de Stadt, H., 647
van Gorkom, J. H., 404, 406, 459
van Haarlem, M. P., 32
van Schooneveld, C., 462
Van Vleck, J. H., 262*n*, 524
Vanden Bout, P., 626
Vander Vorst, A. S., 181
Vanier, J., 343
Verschuur, G. L., 611
Vessot, R. F. C., 338, 343, 344, 347, 348, 349, 351, 376
Vinokur, M., 193
Visser, J. J., 213
Vitkevich, V. V., 26
von Hoerner, S., 634, 636
Vonberg, D. D., 12

Wade, C. M., 95, 469
Wahr, J. M., 481, 482
Walden, A. T., 292
Waldram, E. M., 458
Walker, R. C., 156, 157, 328, 486, 489, 490
Wallington, S., 461, 462
Walsh, D., 25, 461
Walsh, J. L., 242
Wand, R. H., 564
Wang Shougun, 583
Wang, T. C., 348
Warburton, J. A., 31, 421
Wardle, J. F. C., 115, 116
Warner, P. J., 439, 445
Waters, J. W., 519, 520, 523*n*, 541, 545*n*
Webber, J. C., 156, 213
Weigelt, G., 650
Weiler, K. W., 102, 103, 104, 115
Weimer, R., 358
Weinberg, S., 573
Weinreb, S., 213, 218, 254, 260, 297, 358
Weintraub, S., 529
Weisberg, J. M., 577
Welch, B. M., 642, 650
Welch, P. D., 292

Welch, W. J., 33, 41, 125, 164, 212, 451, 540*n*, 543, 552, 554
Welton, T. A., 215
Wengler, M. J., 216
Wernecke, S. J., 422, 433
West, M. E., 624
Westfold, K. C., 573
Westwater, E. R., 541, 542, 543, 583
Weymann, R. J., 461
Wheeler, J. A., 574
White, M., 404
White, N. M., 636
White, R. L., 402
Whiteoak, J. B., 154, 579
Whitford, A. E., 632
Whitney, A. R., 35, 314, 352, 356
Wiedner, M. C., 554
Wieringa, M. H., 454
Wietfeldt, R. D., 355, 356, 368
Wild, J. P., 41, 155
Wilkinson, P. N., 439, 441, 444, 453, 454
Williams, W. F., 123
Willis, A. G., 287
Wills, D., 26, 574
Wilson, R. W., 28
Wilson, T. L., 5, 41, 97
Winterhalter, D., 571
Wirnitzer, B., 650
Woestenburg, E. E. M., 213
Wohlleben, R., 41
Wolf, E., 69*n*, 82, 83, 97, 594, 607, 611
Wolszczan, A., 583, 610
Woodward, R. H., 353
Woody, D. P., 213, 216, 554
Woolard, E. W., 481
Woolf, N. J., 534, 642*n*
Worden, S. P., 648
Wozencraft, J. M., 184*n*
Wright, M. C. H., 180, 212, 403, 538*n*, 540*n*
Wu, S. C., 543

Yang, K. S., 222
Yao, S. S., 260, 264
Yee, H. K. C., 438
Yeh, K. C., 562
Yen, J. L., 34, 41, 290, 304
Young, A. C., 212
Young, A. T., 575

Zeissig, G. A., 580
Zensus, J. A., 377
Zorin, A. B., 216

Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA

SUBJECT INDEX

A reference to a footnote of a page or table is indicated by *n*.

Aberration
  diurnal, 9, 316, 481
  chromatic, 409
Absorption
  bound oscillator model, 524–527
  chopper wheel calibration, 523
  in clouds (liquid water), 542
  coefficient, 508, 521, 526–527
  ionospheric, 555, 562
  spectra, 28
  tipping scan calibration, 552
  tropospheric, 518–528
  by water vapor, 519, 541–543
Academia Sinica, 12, 155
Adaptive calibration, 438–444. *See also* Hybrid mapping; Self-calibration
  comparison of methods, 442–444
  limitations of, 444
Adaptive optics, 642
Airy disk (diffraction pattern of circular aperture), 389*n*, 600, 648
Algol, 468
Aliasing, 192, 256, 414, 459. *See also* Ringlobes
  definition, 127
  suppression in maps, 394–399
Allan variance, 334–335, 344
  atmosphere, 537–538
  frequency standards, 334–346, 350–351
Allen–Baumbach model, 572
ALMA, 12, 543, 547. *See also* Chajnantor, Chile
Altazimuth mount, 94–96, 114, 116
Amplifiers, at antennas, 132
Amplitude closure, 400–401. *See also* Adaptive calibration
Amplitude scintillation, 610
Analog processing
  comparison with digital, 254
  Fourier transformation, 422
Analytic signal, 82–84
Anomalous refraction, 539–540
Antenna(s), 78, 122–125, 599
  angular resolution (beamwidth), 1, 599
  aperture efficiency, 10, 125
  aperture illumination, 133, 134, 451–452, 636
  asymmetric feed geometry, 124
  axis offsets, 94–96, 499
  Cassegrain focus, 123, 125
  collecting area, 10, 79, 125
  cost scaling, 163
  design, 122–125
  feed displacement, 116
  feeds, bandwidth of, 117
  measurements, holographic, 636–641
    Misel algorithm, 640–641
    practical considerations, 638–640
    required sensitivity, 638
  millimeter wavelength, 163
  minimum number, 440
  mounts, 94–96
  Naysmith focus, 123
  offset-Cassegrain focus, 123
  polarization
    circularly polarized, 104–105
    cross polarized, 103–105
    identically polarized, 102–103
    oppositely polarized, *see* cross polarized
    polarization model, 99–102
  prime focus, 123
  received power, 10, 58
  reflections in structure, 288
  response (reception) pattern, 58, 134, 599
  shaping of reflector, 123
  sidelobe model, 614
  in space, 37–39, 158–159, 373–377
  spacing loci, 89–90
  surface accuracy, 124–125
  surface measurements, 636–641
  temperature, 10
  tracking effect on fringe frequency, 138–139
  unblocked aperture, 123

Antenna(s) (*Continued*)
voltage reception (response) pattern, 78–79, 134, 599, 636
Antenna-spacing coordinates, 86–89
Aperture, radiation pattern of, 389*n*, 599–600
Aperture efficiency, 10, 125
Aperture synthesis, 137, 140
Apodization (weighting), 392
Area density, 388
Arecibo, Puerto Rico, 610
Arrays
Arsac, 142
Bracewell, 142
circular (ring), 150–153
closed configurations, 150–153
correlator, 129–132
cross-shaped, 24–26, 137, 148. *See also* Mills cross
grating, 26–27
linear, 142–147
minimum redundancy, 142–143
mixed, 96
mosaicking, 451–453
nontracking, 122, 133
one-dimensional, *see* Arrays, linear
open-ended configurations, 148–150
phased, *see* Phased array
planar, 159–161
Reuleaux triangle, 122, 152–153, 155
sensitivity, 162
tracking, transfer function of, 138–139
T-shaped, 137, 148
two-dimensional, 147–161
VLBI, 155–159
Y-shaped, 148–150
Astrometry, 1, 316, 467–506, 561
accuracy, 475
nutation, 481
polar motion, 482, 484–485
position measurements, 470–472
precession, 481
reference frames, 469–470
VLBI, 472–476
Atacama Large Millimeter Array, *see* ALMA
Atmosphere, neutral, 507–554
absorption, 518–528
brightness temperature, 541–543
calibration, *see* Calibration, atmosphere
effects
on astrometry, 475
of clouds, 542
on fringe frequency and delay, 475
on visibility, 530–539
on VLBI, 534–539
excess path length, 516–518
Fried length in, 536, 641
opacity, 545–546. *See also* absorption
oxygen, 518, 520, 542
phase fluctuations, 530–540
refraction, 513–518
water vapor, *see* Water vapor
Attenuation in cable, 227
Attenuation in optical fiber, 218
Australia telescope, 114, 117, 154, 450
Autocorrelation function, 54, 133–135, 136, 184, 257, 333, 621
definition, 54, 133
of intensity distribution (map), 444, 641
Autocorrelator, 254
Automatic level control (ALC), 248, 278, 384, 622
Automatic phase-correction, 228–229
Azimuth, 88, 117

Bandpass calibration, 404–406
Bandwidth
correlation, 567
effect in maps, 199–205, 234
Gaussian, 55
output (postcorrelator), 185
pattern, 55, 56, 601
rectangular, 55, 234
rms, 366, 474, 495
synthesis, 366–368, 473
Bartlett weighting (smoothing), 286
Baseband response, 218, 256, 364
Baseline
calibration, 93–94, 470–472, 499
coordinates, 86–88
definition, 16, 50
equatorial component, 472
reference point, 95
retarded, 315–316
surveying, 2, 88, 468
Baselines, non-coplanar, 76–77, 454–458
Beam
clean, 427–429
fan, 24, 26, 145, 421
fringe frequency, 490
pencil, 24, 145
synthesized (dirty), 148, 201–202, 378–392, 412, 501, 620, 648
Beam-shape effects, 96–97, 389–391
Beamwidth effects, 96–97, 389–391
Besselian year, 481
Bias
in MEM, 438
in polarization measurements (Rice distribution), 116
in variance of mean, 492
in visibility measurements, 319, 320, 322

Bispectrum, 330–331
Bivariate (joint) Gaussian probability distribution, 255–256, 497, 568
Blackbody, *see* Planck formula
Blackman weighting (smoothing), 286
Bologna, Italy, 26
Borrego Springs (California), 155
Brewster angle, 97
Brightness, 8. *See also* Intensity
  temperature, 9
Brunt-Väisälä frequency, 563
Bureau International de l'Heure (BIH), 484
Burst mode, VLBI, 368–369
Burst radiation, Jupiter, 34, 304

C_n^2, 535
C_{ne}^2, 569
Cables
  attenuation, 227
  reflections, 224, 227, 235
  velocity dispersion, 224
Calibration, 383–387
  adaptive, 438–444
  of atmospheric
    delay, 560–562
    phase, by fast switching, 550–551
    phase, by paired antennas, 551
    phase, by water vapor measurement, 552–554
    phase, in VLBI, 353
  bandpass, 404–406, 409
  baseline, *see* Baseline calibration
  cables, 79, 460
  gain, 386
  polarization, 112–116
  sources, 21, 308, 385–387
  spectral-line, 404–409
Calibrator (source), 385–387. *See also* Calibration sources
Callen and Welton formula, 215–217
Cambridge (England), 24, 540, 563
  Five-Kilometer Radio Telescope, 32, 143
  fourth survey, 26
  Low-Frequency Synthesis Telescope, 458
  One-Mile Radio Telescope, 30, 31, 143
  third survey, 9, 25
Canadian VLBI array, 34, 356
Cassegrain focus, 123
Causal function, 84
Cavity pulling, 350, 351
CCIR, 625
Celestial
  coordinates, 88, 117
  equator, 87, 88, 147
  sphere, 69, 87
Cell
  averaging, 129, 239, 398, 617
  crossing time, 618
Chajnantor, Chile, 33, 540, 545, 546, 548
Chandler wobble, 482
Chi-squared parameter (χ^2), 404, 491–493
Chopper wheel method, 523
Circular array, 150
Circular polarization, 115
  degree, 98
  IEEE definition, 100
Circulator, 223
Classical electron radius, 564
Clausius–Clapeyron equation, 512
CLEAN algorithm, 62, 146, 402, 427–432
  application, 429–432
  Clark's algorithm, 431–432
  comparison with MEM, 434–435
  extended sources, 431, 434–435
  hybrid mapping, 439
  loop gain, 427, 430
  self-calibration, 441
  speckle imaging, 649
  spectral line data, 459
Clipping (clipped noise), 261, 262. *See also* Quantization
Closure relationships, 23, 399–401, 439–440, 641
  amplitude, 400
  phase, 23, 330, 400
Clouds, atmospheric
  absorption, 542
  index of refraction of, 542
CMB, *see* Cosmic microwave background
COESA (Committee on the Extension to the Standard Atmosphere), 516, 519
Coherence
  complex degree of, 604
  function
  source, 69, 600, 603–607
    temporal, 340–342. *See also* Autocorrelation function
  of hydrogen maser, 344
  mutual, 594–597, 631
  of oscillator, 340–344
  partial, 604
  propagation of, 607–611
  pulsars and masers, 611
  self, 605
  time, 304, 308, 319, 324, 340–344, 473
Coherency matrix (polarization), 99
Coherent source, 606–607
Collecting area, 10
Comb spectrum, 231, 352
Compensating delay, *see* Delay, instrumental
Complex correlator, *see* Correlator, complex

Complex degree of coherence, 604
Complex visibility, 27, 61, 69
  definition, 69
Compound interferometer, 27
Compton loss, 306
Confusion of radio sources, 24, 141, 455, 459
Connected-element array (definition), 35*n*
Continuum radiation, 3
Conventional International Origin (CIO), 482. *See also* Polar motion
Conversion
  frequency, *see* Frequency conversion
  serial-to-parallel, 282. *See also* Demultiplexing
Convolution, 59
  theorem, 60
Convolving functions, 393–398. *See also* Smoothing functions
  Gaussian, 395
  Gaussian-sinc, 396
  rectangular, 394
  spheroidal, 396
Coordinate
  conversion, 117
  systems, 64, 70, 86–91
Cornell University, 34
Correlator, 80*n*
  analog, 220
  comparison, lag, and FX, 293–297
  complex, 174–175, 188
  digital, 283, 289–296
  FX, 290–293
  hybrid, 297
  lag (XF), 289–290
  multiplexing in, 297
  output in the complex plane, 177, 189
  recirculating, 290
  simple (single-multiplier), 174–175, 198
  system, 80*n*
  voltage offset in, 241, 278, 413, 414
Cosmic Background Explorer, 11
Cosmic Background Imager, 160
Cosmic microwave background (CMB), 159, 404, 419, 521
  anisotropy of, 159, 404, 521
Costas loop, 376
Covariance matrix, 497
Cross, *see* Mills cross
Cross-correlation, 77–78, 80
  coefficient, 257
Cross power spectrum, 77, 284–287, 361, 495. *See also* Spectral line
Cross-talk (cross-coupling), 161, 245, 413
Cryogenic cooling, 181, 212
Crystal-controlled (quartz) oscillator, 231, 332, 342, 345, 348
Culgoora array (Australia), 155
Cycle time, 479, 550
Cyclotron
  frequency, 558
  radiation, 97

dBi, 614
Declination, 9
  coordinate conversion, 117
  measurement of, 467–468, 470–473, 499
Deconvolution, 426–438
  comparison of CLEAN and MEM, 434–438
Delay
  adjustment, 55, 238–239
  analog, 220
  circuits, digital, 282
  compensating, *see* Delay, instrumental
  errors, 238–239
  fractional sample correction, 295
  geometric, 50, 68, 171, 310, 357, 514, 643
  group, 308, 314, 366, 473, 485, 486, 555
  instrumental, 53, 91, 171, 238–239, 357
  measurement error, 366–367
  delay pattern, *see* Bandwidth, pattern
  reference, 173, 239
  subsystem, 220, 282–283
  tracking, 173
Delay resolution function, 369. *See also* Bandwidth synthesis
Delay-setting tolerances, 176, 238–239
Delta function
  CLEAN components, 427–428
  Fourier transform of series of, 144–145
  LO frequency, 352
  point source, 92, 140, 206, 469
  Price's theorem, 266, 271
  Shah function, 126–127, 392–393
  visibility sampling, 60, 190
Demultiplexing, 297
Depolarization, 578–579
Detector
  power-linear (square law), 20, 614, 627
  synchronous, 20, 222, 241, 523
Diameter, stellar, 13–16, 647
Dielectric constant, 524
  of plasma, 557
Diffraction at an aperture, 597
Diffraction pattern
  lunar occultation 632
  scintillation, 567, 576
Digital processing, 254–301
  multiplication, 284
  sampling, 256–282
    accuracy, 278–282
  spectral measurements, 284–297

Diode, 168–169, 222, 231, 350, 352
Direct detection, optical, 40, 644–645
Directional coupler, 222–223
Direction cosine, 64, 71, 601
Dirty beam, *see* Beam, synthesized
Dirty map, 427
Discrete Fourier transform, 128–129, 392–394
Dispersion, classical theory, 524–528
Dispersion measure, 576
Dispersion in optical fiber, 219–220, 249
Diurnal aberration, *see* Aberration, diurnal
Doppler effect, 51, 138, 289, 346, 349, 351, 485
  analysis and formulas, 417–421
  reference frames, 418–419
Double sideband system, 175–183, 196–198
Dynamic range, 422, 445–446, 469, 623

Earth, *see also* Geodetic measurements; Polar motion
  atmosphere, 507–513
  equatorial bulge, 481
  ionosphere, 554–564
  magnetic field, 558
  radius vector, 315, 516
  tectonic plate motion, 2, 485
  tides, 485
Earth rotation, *see also* Universal Time
  scanning, 17
  synthesis, 30
East-west array, 74, 142–147, 388–389
East-west baseline, 50, 206
Ecliptic, 481
Editing of data, 295, 383
  for interference, 615
Efficiency
  aperture, 10, 125, 636
  quantization, 188, 272, 276, 365
Electron density
  galaxy, 577
  interplanetary medium, 572
  interstellar medium, 577
  ionosphere, 556
Electronics
  historical development, 212
  subsystems, 212–214, 217–221
Elevation, 88, 117
Entropy, 432–433
Equatorial mount, 94–96
Equinox, 468, 481
Ergodic waveform, 3, 82
Error function (erf), 193, 267, 274
Errors
  additive, 412–413, 623
  clock (VLBI), 310, 314
  in maps, 412–413
  (l, m) origin, 414
  multiplicative, 413, 623
  phase, 233, 445
  pointing, 383, 413
Evolution of synthesis techniques, 12
Excess path length, 509
  interplanetary medium, 574
  ionosphere, 555, 559–560
  troposphere, 516–518, 541–543
Extended (broad) sources
  deconvolution, 428, 431, 434–435
  mosaicking, 446–453
  response, 412
  signal-to-noise ratio, 191

Fan beam, 24, 145
Faraday depth, 578
Faraday dispersion function, 578
Faraday rotation, 3, 97, 116, 558
  dispersion function, 578–579
  interstellar, 576–579
  ionospheric, 555, 558–559
Far-field assumption, 50, 68, 601
Fast Fourier transform (FFT), 128–129, 392–393
Fast Hartley transform, 128
Feeds, bandwidth, 117
Fiber optics, *see* Optical fiber
Fiedler events, 582
Field
  far, requirement, 601
  near, observations in, 601
Field of view
  bandwidth effect, 199–205
  fringe-frequency mapping, 488
  restrictions, 200–204, 601
  visibility averaging effect, 205–208
Filters, 79, 169
  baseband, 218
  Butterworth, 365
  digital, 297
  effect on signal-to-noise ratio, 235
  narrow-band, 233
  number of poles, 232–233
  phase-locked oscillator, use of, 233
  phase stability, 232–233
  Q-factor, 233, 350–351
  spectral-line, 290–297
Fleurs, Australia, 24, 27
Flux density, 6
Fort Davis, Texas, 35, 156
Fourier transform
  analog hardware, 422
  derivative property, 333, 403, 634
  direct, 387–388
  discrete, 128–129, 392–394

Fourier transform (*Continued*)
fast, 128, 290–291
integral theorem, 185
projection-slice theorem, 65–66
relationships, mapping, 134
shift theorem, 408
sign of exponent, 69*n*
similarity theorem, 201
three-dimensional, 76, 455–457
Fourth-order moment relation, 184, 258, 627
Fractional bit shift loss, *see* VLBI, discrete delay step loss
Fractional frequency deviation, 332
Fraunhofer diffraction, 595, 597. *See also* Field, far, requirement
Frequency
channels, 284
conversion, 168–169
multiple, 173, 178
optical, 646–647
demultiplexing, 298
multiplication, 231, 352
regulation, 625
response, 233–238
optimum, 233–234
tolerances, 235–238
Frequency standards, 332–346
cesium beam, 347
crystal oscillator, 231, 332, 342, 345, 348
hydrogen maser, 348–351
phase noise processes, 337–340
rubidium vapor, 346–347
Fresnel zone, 566, 632
Fried length, 536, 642
Fringe
envelope, 52, 55
fitting, 195
global, 326–331
function (pattern), 17, 52, 59
rotation, digital, 294, 358–361
rotation (stopping), 173–174, 180–181, 217, 246–247
search, *see* Signal search
visibility, 13
washing function, 55
white light, 57, 307, 644
Fringe frequency, 91, 172
in astrometry, 472–474
averaging, 616
baseline solution, 472
beam, effective, 490
definition, 91–92
effect of tracking, 91–92, 138–139
interference suppression effect, 616–620
ionospheric effect on, 560
mapping with, 488–490
measurement accuracy, 494
natural, 172
spectrum, 325, 489, 494
in VLBI, 319–320
Fringe rate, *see* Fringe frequency
Fringes, first radio record, 18
Front end, 213. *See also* Receiver
Frozen screen approximation, 537, 548
FX spectral line processor, 290

Gain
calibration, 248
errors, 235–238, 280
factor, 172
Gamma function, 573
Gaussian convolving function, 395
Gaussian random noise, 3
Gaussian random variable, 3, 184, 316–319, 340. *See also* Bivariate Gaussian probability distribution
Gaussian-sinc function, 396
Gaussian taper, 137, 141, 390–391, 428
Geodetic measurements, 35, 467*n*, 485
Geometric delay, 50, 68, 171, 310, 357, 514, 643
Gibbs phenomenon, 288, 405
Global fringe fitting, 326–329
GMRT, 32, 155
Goldstone, California, 35
GPS (Global Positioning System), 2, 353, 483, 485, 560, 562
Granlund system, 228–229
Grating array, 26–27, 421
Gravitational deflection, *see* Relativity
Green Bank, W. Virginia, 31, 34, 430, 540
Greenwich meridian, 86, 316, 473, 482, 484
Gridding (convolutional), 392–394. *See also* Cell averaging
Group delay, 308, 314, 366, 473, 485, 486, 555
Group velocity, 557, 576
Gyro frequency, 558

Hadamard matrices, 243
HALCA Satellite, 12, 38, 158, 375–376
Half-order derivative, 146
Hamming weighting (smoothing), 286
Hankel transform, 533
Hanning weighting (smoothing), 286–287
Hartley transform, fast, 128
Hat Creek Observatory, California, 33, 35, 540
Haystack Observatory, Westford, Massachusetts, 34, 324, 640
Heterodyne conversion, *see* Frequency conversion
Hilbert transform, 82, 84, 174, 188*n*, 283

Hinge point, 542
Hipparcos satellite, 2
  star catalog, 470
Historical development, 12–39
  analog Fourier transformation, 422
  mapping from one-dimensional profiles, 421–422
  receivers, 181, 212
  VLBI, 33–37, 304–306
Holes in spatial frequency coverage, 141
Holography, *see* Antenna measurements, holographic
Hour angle, 86–88, 117
Hubble constant, 420
Hybrid correlator, 297
Hybrid mapping, 35, 438–439
Hydrogen line, 5, 28, 348–349
Hydrostatic equilibrium, 510

IAU
  polarization standard, 100, 101
  radio-source nomenclature, 9
ICRF, 9, 35, 469
ICRS, 469
IEEE
  committee on frequency stability, 332
  polarization standard, 100
  power flux density, 6
IF, *see* Intermediate frequency
Illumination, aperture, *see* Antenna, aperture illumination
Image, 8*n*
Image defects, *see also* Phase noise
  correlator offset, 413
  distortion, 413
  errors in visibility data, 412–413
Incoherence assumption (spatial), 69, 600, 628
Incoherent averaging, 323–326, 331–332
Incoherent source, response to, 603–606
Index of refraction, *see* Refraction, index of
Inertial reference frame, 469
Infrared interferometry
  detection of planets, 644
  heterodyne detection, 646
Instrumental (compensating) delay, *see* Delay, instrumental
Instrumental polarization, 105–109, 112–116
  degrees of freedom, 114
Intensity, 8, 411–412
  derivation, 387–399
  interpretation, 411
  scale, 411, 439
Intensity interferometer, 22, 627–631
  optical, 647
  sensitivity of, 326, 631
Interference, radio, 613–626
  connected-element arrays, 615–621
  decorrelation effect, 620–621
  fringe-frequency averaging, 616–619
  harmful thresholds, 616
  ITU, 625
  satellites, 624–625
  single antenna, 615
  solar, 413
  (u, v) plane distribution, 618–620
  VLBI, 621–624
Interferometer
  adding (simple), 16, 18, 20
  basic components, 78–80
  compound, 27
  correlator, 20
  infrared, *see* Infrared interferometer
  intensity, *see* Intensity interferometer
  Michelson, 13–16
  optical (modern Michelson), 641–648
  sea, 18
  spectral-line, 28
Intermediate frequency (IF), 169
  amplifier, 218
  subsystem, 218
International Astronomical Union, *see* IAU
International atomic time (IAT), 483
International Celestial Reference Frame, 9, 35, 469
Interplanetary medium, 571–574
  electron density, 572
  excess path length, 574
  refraction, 571–574
  scintillation, 574–576
Interpolation, 127, 392–394. *See also* Gridding
Interstellar medium, 576–583
  dispersion measure, 576
  electron density, 577
  Faraday rotation, 576–578
  pulsar signals, effects on, 577
  scattering
    diffractive, 579–580
    Fiedler, 582
    refractive, 580–583
Invisible distribution, 427
Ionosphere
  absorption, 555–562
  acoustic-gravity waves, 563
  effects of irregularities, 562–564
  electron density distribution, 556
  Faraday rotation, 555, 558–559
  Gaussian screen model, 564–569
  index of refraction, 557–559
  phase stability, 555
  power-law model, 569–571

Ionosphere (*Continued*)
propagation delay, 559–569
refraction, 559–560
scintillation, 562–564
total electron content, 555, 560, 563
traveling ionospheric disturbances (TIDs), 563
Isoplanatic
angle, neutral atmosphere, 642
patch
ionosphere, 401, 460, 555
neutral atmosphere, 642
ITU, 625

Jansky (unit), 6
Jodrell Bank Observatory, England, 22*n*, 23, 31, 155, 304
Jones matrix, 109
J^2 synthesis (J-squared synthesis), 155
Julian year, 481
Jupiter, 34, 304

Kolmogorov turbulence, 531, 533–538, 569–571, 579
Kramers–Kronig relation, 524, 527

Leakage (polarization), 106, 117–120
Leakage (sampling), 127
Leap second, 483
Least-mean-squares analysis, 490–502
accuracy, 498
CLEAN, 429
correlated measurements, 497
covariance matrix, 497
design matrix, 498
error ellipse, 497–498, 502
estimation of delay, 495
estimation of fringe frequency, 494
large data base reduction, 499
likelihood function, 490
matrix formulation, 496–497
nonlinear case, 499
normal equation matrix, 497–498, 500
partial derivative matrix, 497
precision, 498
self-calibration, 441
sinusoid fitting, 195
source position errors, 500–502
variance matrix, 497
weighted, 491
Lensclean, 461
Light, speed (velocity) of, 50, 467
Likelihood function, 490
Line of nodes, 481. *See also* Equinox
Linear arrays, 142–147
Lines, radio, *see* Spectral line
Linked-element array, 35*n*
Lloyd's mirror, 18
LO, *see* Local oscillator
Local oscillator, 168–169, 217. *See also* Frequency standards
independent, 34. *See also* VLBI
laser, 644
multiplication, stability, 351–352
nonsynchronized, 631
phase switching of, 247
signed-sum of frequencies, 173
synchronization of, 221–232
Local standard of rest, 418–419
Long baseline interferometer, 23, 304
Long wavelength arrays, 163
Loran, 353, 483
Lorentz factor, 420
Lorentzian profile, 525, 527, 531, 533
Low frequency mapping, 459–461
Low-noise input stage, 181, 212–214
Lunar occultation
optical, 631, 636
radio, 21, 468, 576, 631–636

Magellanic Cloud, small, 450
Magnetic fields
in frequency standards, 348, 349, 351
interstellar, 577
terrestrial, 558
Magnetic tape recording, 34, 353–356
Magnitude of visibility, 62*n*
Map, 8*n*
Mapping
synthesis, definition, xx*n*
two-dimensional, 64–65
wide field, 74–77, 204–205, 446–450, 454–458
visibility amplitude only, 444
Maryland Point Observatory, Maryland, 35, 324
Maser frequency standard, 348–351
Maser radio sources, 6, 34, 304, 306, 324, 580
mapping procedures, 485–490
spatial coherence, 611
Master oscillator, 217
Mauna Kea, Hawaii, 33, 155, 540, 545–549
Mauritius Radio Telescope, 155
Maximum entropy method (MEM), 432–434
Maximum likelihood method, 367, 490, 501
Maxwell's relation, 525
Meridian, 86
Greenwich, 86, 316, 473, 482, 484
local, 87, 88, 468, 484
plane, 86, 471
transit (crossing), 484
MERLIN, 24, 155, 444, 616, 620
Meter, definition of, 467

Michelson interferometer, 12, 13–16
Microwave link, *see* Radio link
Millibar, 509
Millimeter wavelength arrays, 33, 163–164, 181, 451
Mills cross, 24, 26, 137, 141
Minimum redundancy, *see* Arrays, minimum redundancy; Bandwidth, synthesis
Mirror-image reception pattern, 59
Mixer, 168–169. *See also* Frequency conversion
  sideband separating (image rejecting), 248–249
MKSA units, 524*n*
Model
  adaptive calibration, 441
  circular disk, 15, 432
  Cygnus A, 22-23
  delta function (CLEAN), 428
  fitting, 401–404
  Gaussian, 15, 29, 402
  moments of, 403
  rectangular, 15
  stellar envelope, 402
  without phase, 401
Modern Michelson interferometer, 643–645
Modulated reflector, 222
Molonglo, Australia, 26
Moon, *see* Lunar occultation; Precession
Moon as a calibration source, 414–416
Mosaicking (mosaic mapping), 446–453
  arrays for, 451–453
  linear, 449
  nonlinear, 449–450
Mueller matrix, 112
Mullard Radio Astronomy Observatory, *see* Cambridge (England)
Multifrequency synthesis, 452–454
Multiplier (voltage), 20, 170. *See also* Correlator
Mutual coherence function, 595–597

Nançay, France, 27
Narrabri, Australia, 647
National Aeronautics and Space Administration (NASA), 2, 35, 356
National Geodetic Survey (NGS), 2
National Radio Astronomy Observatory (NRAO), 31, 34, 148, 305, 335, 356, 443. *See also* ALMA; Green Bank; Very Large Array (VLA); Very Long Baseline Array (VLBA)
Natural weighting, 191, 388, 392
Naval Observatory, U.S. (USNO), 484
Naval Research Laboratory (NRL), 2, 324
NAVSTAR, *see* GPS
Near field observations, 601
Negative frequencies, 55, 61, 83, 84
Network Users Group (U.S.), 34–35
Neutral atmosphere
  opacity, 543–547
  phase stability, 547–550
Nobel Lecture, Ryle, 32
Nobeyama Radio Observatory (NRO), Japan, 33, 540
Noise, *see also* Signal-to-noise ratio
  amplitude and phase, 192–193
  in complex visibility, 188–189, 196
  equivalent power (NEP), 645–646
  Gaussian, 3
  in map, 189–192
  in oscillators
    flicker-frequency, 338–340
    flicker-phase, 338–340
    random-walk-of-frequency, 338–340
    white-frequency, 338–340
    white-phase, 338–340
  photon shot noise, 346, 351, 646
  power, 10, 213
  quantum effect, 39–40
  response to, 183–189
  temperature measurement, 214–217
  in VLBI, 316–319
Non-coplanar baselines, 76–77, 454–458
  polyhedron mapping, 457
  snapshot combination, 458
  3D Fourier transform, 456-457
  variable point-source response, 458
Non-negative least squares, 435
North Liberty, Iowa, 35
NRAO, *see* National Radio Astronomy Observatory
Nuffield Radio Astronomy Laboratories, *see* Jodrell Bank Observatory, England
Nutation, 2, 9, 481–482
Nyquist rate (frequency), 256–257. *See also* Sampling theorem

Observation, planning and reduction, 413–414
Occultation observations, *see* Lunar occultation
Opacity, 518–521
  measurement of, 521–523
Optical depth, *see* Opacity
Optical fiber, 218–220, 229-230
  dispersion, 220, 249–250
  high stability, 230
Optical interferometry, 40, 641–648
  direct and heterodyne detection, 644–647
Orbiting VLBI, *see* OVLBI
Oscillator coherence time, 340–342
Oscillator strength, 525
Outer product, 110, 243
Oversampling, 257, 259–260, 263–264, 270, 272, 277

OVLBI, 37–39, 158, 373–377
data link, 375–376
round-trip phase, 374
timing link, 375–376
Owens Valley Radio Observatory, California, 28, 29, 33, 35, 554

Parabolic-cylinder reflector, 122
Paraboloid reflector, 123
Parallactic angle, 88, 97, 104, 114, 116
Parallax, 482
Parametric amplifier, degenerate, 181
Parseval's theorem, 192, 298, 324, 391, 534, 619, 623
Partial coherence, 604
Passband
Gaussian, 55, 204
rectangular, 55, 202, 234
tolerances, 235–238
Pencil beam, 24, 145
Permittivity, 524*n*
Phase
errors, effect on sensitivity, 233
noise
effects on maps, 445, 487
in frequency multipliers, 351–352
in frequency standards, 324, 332–340
ionospheric, 562–564
neutral atmospheric, 530–539, 631
in receivers, 192–193, 316–319
Phase closure, 22-23, 35, 306, 387, 399–401
Phased array, 129–132, 155, 296*n*, 369–373
correlator array, comparison with, 129–132
randomly phased, 370
as VLBI element, 369–373
Phase data
imaging without, 444
uncalibrated, 438–444
Phase-locked oscillator, 224, 230–232, 342
loop natural frequency, 230–231
Phase reference
feature, 486
position, 57, 68, 86
Phase referencing in VLBI, 476–480
atmospheric effects, 550–551
for masers, 486
Phase stability
analysis of, 332–342
of filters, 232–233
of frequency standards, 342–351
of local oscillators, 351–352
in reference distribution, 221–230
Phase switching, 240–248, 278, 280
in early arrays, 26
interaction with fringe rotation and delay, 246
in Mills Cross, 24
in simple interferometer, 18–21
Phase tracking center, 68. *See also* Phase reference position
Pico de Veleta, Spain, 638
Planar arrays, 159–161
Planck formula, 8, 10, 215–217, 645
Planetary nebula, 4, 386, 402
Planets, 432, 468, 481. *See also* Burst radiation, Jupiter
as calibration sources, 414–416
Plasma, *see also* Interplanetary medium; Interstellar medium; Ionosphere
absorption in, 562
frequency, 557
index of refraction, 558–559
oscillations, 97
propagation in, 555–583
RF discharge, 346, 348
turbulence, 569–571
Plateau de Bure, France, 33, 540
Pointing correction, 383
Point-source response, 56, 133, 140, 427, 531. *See also* Beam, synthesized (dirty)
Point-spread function, 648. *See also* Point-source response
Poisson distribution, 39
Polarimetry, 97–120
Polarization
calibration, 112–116
circular, 98, 100, 104–105, 115, 116
complex degree of, 578–579
degree of, 98
design considerations, 115–117
ellipse, 99–100
emission processes, 3, 97
instrumental, 105–109, 240
linear, 98, 103–104, 116
matrix formulation, 109–112
mismatch tolerance, 240
parallactic angle effect, 104
position angle, 98, 99. *See also* Faraday rotation
Polar motion, 2, 482
measurement of, 484–485
Position measurements
early, 21
methods, *see* Astrometry
Power combiner, 130, 132
Power (density) spectrum, 54, 77
atmospheric phase, 535–537
correlator output, 186
interplanetary scintillation, 575–576
phase and frequency fluctuations, 332–342
Power flux density, 6

Power-law antenna spacing, 149–150
Power-law turbulence relations, 538
Power reception pattern, *see* Antenna, reception pattern
Poynting vector, 8
Precession, 2, 9, 481–482
Price's theorem, 266, 271
Principal response, 392
Principal solution, *see* Principal response
Probability
  of error, 319–323
  of misidentification, 322–323
Probability distribution
  of angle of arrival, 563
  bivariate Gaussian, 255–256, 497, 568
  of delay-setting error, 238–239
  Gaussian, 124, 184, 255–256, 273, 316, 318, 490, 531
  Rayleigh, 193, 317, 319, 322
  Rice, 193, 317
Projection-Slice theorem, 65–66
Prolate spheroidal wave functions, 397
Propagation
  constant, 315, 508
  interplanetary, 571–576
  interstellar, 576–583
  ionospheric, 554–564
  neutral atmospheric, 508–543
Proper motion, 9, 482
Pulsar, 368–369
  astrometry, 469
  correlator gating, 296
  determination of vernal equinox, 469
  dispersion measure, 572, 576–577
  proper motions, 579
  scintillation, 580
  spatial coherence, 611
  timing accuracy, 345, 353
Pulsars, *see* Radio source
Pulse calibration (VLBI), 352

Q-factor of
  cavity, 350–351
  filter, 233
QPSK modulation, 376
Quadrature
  network, 174, 278, 286
  phase shift ($\pi/2$), 182, 188, 246, 283
Quadruple moment theorem, *see* Fourth-order moment relation
Quadrupod, 125, 639
Quantization
  comparison of schemes, 277–278
  correction, 276–277, 295, 300–301
  efficiency factor, 188, 272, 276, 357, 365
  eight (or more) levels, 273–276
  four-level, 264–271
  indecision regions, 280–282
  noise, 254, 273–276
  repeated (requantization), 298, 373
  three-level, 271–272
  thresholds, 264–265, 271, 273
  two-level, 261–264
  in VLBI systems, 357, 365
Quantum noise, 39–40, 646
Quantum paradox, 39
Quasar, 3, 35–36, 306, 442. *See also* Radio source

Radamacher functions, 242, 244–246
Radial smearing, *see* Bandwidth, effect in maps
Radiative transfer, equation of, 521
Radio interference
  airborne and space transmitters, 624–625
  decorrelation, 620–621
  fringe-frequency averaging, 616–619
  threshold pfd and spfd
    short and intermediate baselines, 615–621
    total power systems, 614–615
    VLBI, 621–624
Radio lines, *see* Spectral lines
Radio link, 22, 23, 24, 35*n*, 218, 375
Radiosonde data, 523*n*, 541, 545
Radio source
  0134+329, 9
  0748+240, 539
  1548+115, 442
  1622+633, 158
  1638+398, 478
  1641+399, 478
  Cassiopeia A, 17, 21, 22, 31
  Centaurus A, *see* NGC5128
  Crab Nebula, 22, 572
  Cygnus A, 9, 18, 21, 22, 562
    central component (VLBI), 37
    fringe pattern, 17, 19
    map or image, 23, 31, 32, 33, 443
  J1745-283, 476
  Jupiter, 34, 304
  M87, *see* NGC4486
  NGC4258, 35, 37
  NGC4486, 22, 435
  NGC5128, 22
  NGC7027, 4, 9, 386
  Orion Nebula, 5, 7
  Orion water-line maser, 638
  P-Cygni, 402
  PSR 1237+25, 610
  PSR B2021+51, 480
  Sagittarius A* (Sgr A*), 35, 475, 580
  sun, 18, 26–27, 155, 421–422. *See also* Sun

Radio source (*Continued*)
Taurus A, *see* Crab Nebula
3C33.1, 29
3C48, 3, 4, 9, 386
3C138, 115
3C147, 386
3C224.1, 430
3C273, 36, 468, 480, 575
3C279, 480, 610
3C286, 115, 386
3C295, 386
Venus, 254
Virgo A, *see* NGC4486
W3(OH), 324
W49, 486
Radio source nomenclature, 9
Radio spectrum, regulation of, 625
Raised cosine weighting, *see* Hanning weighting
Rayleigh distribution, 193, 317
Rayleigh–Jeans approximation, 8, 10, 213, 215, 216. *See also* Planck formula
Receiver
electronics, 212–221
phase switching, 18–21, 24, 26, 240–248. *See also* Phase switching
temperature, 10, 214–217
for cascaded components, 214
Reception pattern, *see* Antenna, reception pattern; Voltage reception pattern
Recording systems (VLBI), 353–356
Redundancy measure, 143
Reference frames, *see* ICRF; ICRS
Reflections
in cable, 224, 227, 235
in optical fiber, 219
Reflector, modulated, 222
Refraction
anomalous, 539–540
interplanetary, 571–574
ionospheric, 557, 559–560
in neutral atmosphere, 508, 513–518
origin of, 524–528
in plane parallel atmosphere, 513–515
optical, 507, 528
spherically symmetric, 515, 571–574
Refractivity, 509. *See also* Refraction, index of
optical, 529
Smith–Weintraub equation, 528–529
Relative sensitivity of systems, 193–199
Relativistic effects
general relativistic bending, 573–574
gravity, 420
Lorentz factor, 420
time transfer effects, 353
Resolution
atmospheric limitation of, 530–540
MEM, 434
Restoration from samples, *see* Sampling theorem
Retarded baseline, 315–316
Reuleaux triangle, 152–153, 155
Reynolds number, 535
Rice distribution, 193, 317
Right ascension, 9
measurement of, 468, 470–472, 499
zero of, 468–469
Ringlobes, 144–147
RMS bandwidth, 366, 474, 495
Robust weighting, 392
Rotation measure, 577, 579
Round-trip phase, 221–228, 375, 384
Ruze formula, 125

Sampling, 256–278. *See also* Quantization
of bandpass spectrum, 256–257
digital, accuracy of, 278–282
Sampling theorem, 126–127, 144, 146, 256, 448, 637
Satellite
data link, 34, 374–375
interference from, 622, 624
signals, Faraday rotation, 560
time transfer, 353
Scattering, 576, 579–583, 610. *See also* Scintillation
Schwarzschild radius, 420
Scintillation
angular spectrum of, 565, 570
correlation bandwidth, 567
correlation length, 566
critical source size, 567
Gaussian screen model, 564–569
index, 575
interplanetary, 574–576, 610
interstellar, 158, 579–583
ionospheric, 562–564
neutral atmosphere, 534–539
power-law model, 569–571
scattering angle, 565, 570
thin screen, 564–569
Sea interferometer, 18–19
Second, definition of, 347, 467. *See also* Time
Seeing, 507. *See also* Scintillation
cell, 642
disk, 648
Self-absorption, 4
Self-calibration, 440–444
Sequency, 243
Serial-to-parallel conversion, 282
Serpukhov, Russian Federation, 26

Shadowing, 384
Shah function, 127
  two-dimensional, 293
Shift-and-add algorithm, 648
Short-spacing data, 146, 164, 451–452
Shot noise, photon, 346, 351, 646
Sideband(s), 169
  double, 175–183, 197–198
  fringe-frequency dependence, 91–92
  partial rejection of, 208–210
  relative advantages of single, double, 181
  separation, 181–183
  sideband-separating (image-rejection) mixer, 248–249
  single (upper, lower), 169, 171, 172
  unequal responses, 208–210
Sidelobe, *see also* Ringlobes; Synthesized beam (dirty beam)
  bandwidth smearing of, 202
  envelope model, 613–614
Sidereal rate (earth rotation), 91
Signals
  cosmic, 3–9
  ergodic, 3, 82
  spurious, 240–241, 246–247. *See also* Errors
Signal search (VLBI), 319–326
Signal-to-noise ratio, *see also* Noise
  aliasing effect, 398–399
  basic analysis, 183–199
  coherent averaging, 323–326
    of frequency standard, 343–346
  frequency response, effect of, 233–235
  fringe-frequency mapping, 488
  incoherent averaging, 323–326
  intensity interferometer, 326, 631
  in interference calculations, 613–624
  loss factors, VLBI, 357–366
  in lunar occultations, 634–635
  in maps, 189–192
  optical, 645–647
  in phased arrays, 369–373
  quantization, effect of, 260–278
  quantum effect, 646
  receiving system, 11
  systems, relative, 193–199
Signal transmission, 218–220
Simeiz, Russian Federation, 34
Sinc function, 52
Single sideband mixer, *see* Sideband separating mixer
Site testing
  opacity, 543–547
  phase stability, 547–550
SKA (Square Kilometer Array), 32
SMA (Submillimeter Array), 12, 155, 639
Smearing
  circumferential, *see* Visibility, averaging
  radial, *see* Bandwidth, effect in maps
Smithsonian Astrophysical Observatory (SAO), 12, 155
Smith–Weintraub equation, 528–529
Smoothing functions, 286
Snapshot, 151, 154
Snell's law, 514–515
  spherical coordinates, 515, 572
Soil temperature, 221
Solar mapping, 26–27, 421–422
Solar system studies, 26–27, 574, 601
Solar wind, 573–574
Source, *see* Radio source
  calibration, 308, 383–387
  coherence, 603–607
  completely coherent, 606–607
  extended, *see* Extended sources
  far-field condition, 50, 68, 601
  incoherence requirement, 69, 596–597, 600
  model, *see* Model
  radio, *see* Radio source
  subtraction, 414. *See also* CLEAN algorithm
South Pole, 545, 546, 548
Space Interferometry Mission (NASA), 2, 644
Space VLBI, *see* OVLBI
Spatial frequency, 58, 61, 132–135, 387
  coverage, 132–135, 426–427
  filter, 133
Spatial incoherence, *see* Source, incoherence requirement
Spatially coherent source, 606, 607
Spatial sensitivity
  of aperture antenna, 451–452
  of correlator array, 132–137
  support of, 133
Spatial transfer function, *see* Transfer function
Specific intensity, *see* Intensity
Speckle imaging, 648–650
  shift-and-add, 648
  phase information, 650
Spectral
  correlators, 283–298
  flux density, 6
  power flux density, 6
Spectral line(s)
  absorption, 28
  accuracy, 409–410
  adaptive calibration, 459
  analog correlator, 220
  atmospheric absorption, 518–523, 544
  bandpass calibration, 404–406
  bandpass ripple, 284–288, 405–406
  baseline ripple, 288

Spectral line(s) (*Continued*)
calibration procedures, 404–409
chromatic aberration, 409
CLEAN procedures, 459
digital correlators, 283–298
Doppler shifts, 417–421
reference frames, 418–419
double sideband observation, 181
examples of
CO, 9, 545
hydrogen, 4, 5, 28, 348
H_2O, 6, 35, 324, 485, 489, 553–554
H_2CO, 28
OH, 6, 304, 485
SiO, 6
presentation, 410–411
radiation, *see* Maser radio sources; Radio lines
systems, 28, 220, 283–297
table of important, 5–6
velocity reference frames, 419
VLBI procedures, 35–37, 314, 406–409, 485–490
Spheroidal wave functions, 396–397
Square Kilometer Array (SKA), 32
Stanford, California, 27
Stars
observation of, 13–16, 468, 631, 636, 644, 647
proper motion, 482
visibility model, 402
Step recovery diode, 352
Stokes parameters, 97–98
Stokes visibilities, 102–105
Strehl ratio, 125
Structure function
phase (spatial), 530, 535–538, 569–570
phase (temporal), 341, 537
refractive index (spatial), 535
Structure function measurements, 540
Submillimeter Array (SMA), 12, 155, 639
Sun
coronal refraction, 571–573
gravitational effects, 481
interference from, 413
ionosphere, 555
observation of, 18, 26–27, 155, 421–422
relativistic deflection, 573–574
solar time, 482
solar wind, 571–573
Superluminal motion, 35, 36
Support of a function, 133
Survey interferometers, 24–26
Swarup and Yang system, 222–223
Symmetry, n-fold, 148
Synchronous detector, 20, 222, 241, 523
Synchrotron radiation, 3, 97, 306, 578
Synthesis mapping, xxi*n*
evolution of techniques, 12
Synthesized beam, *see* Beam, synthesized
System equivalent flux density (SEFD), 11, 387, 408
System temperature, 10, 185–188, 199, 384
correction for atmospheric absorption, 522
measurement of, 248

Tangent plane, 72, 74, 76, 456
Taper, *see* Gaussian taper; Weighting
Target source, 385
T-array, 25, 137, 148
TDRSS experiment, 12, 38
Tectonic plates, 2, 485
Telephone signal transmission, 307
Temperature
antenna, 10
receiver, 10, 214–217
system, *see* System temperature
Temperature coefficient of length, 221
Thomson scattering (incoherent backscatter), 560, 572
3C Sources, *see* Radio source
Time
averaging of visibility, 205–208
definition of second, 347
demultiplexing, 297
International Atomic (IAT), 347, 483
multiplexing, 231
solar, 482
time synchronization, 353
transfer methods, 353
universal time, 353, 482–484
Timing accuracy, 94, 353
Tipping-scan method, 522
Tolerances in
bandpass (frequency response), 235–239
delay-setting, 238–239
polarization, 240
three-level sampling, 279–282
Tomography, 422
Total electron content, *see* Ionosphere, total electron content
Transfer function, 132–135, 138–140, 387, 426–427. *See also* Spatial sensitivity
OVLBI, 158–159
VLA, 151
VLBA, 157
Transmission lines, *see* Cables; Local oscillator, synchronization; Optical fibers; Waveguide
Traveling ionospheric disturbances (TIDs), 563-564
Triple product, 330–331

Tripod, 125
Troposphere, *see* Atmosphere, neutral
Truncated function, 84–85
Turbulence
  Allan variance of, 537
  inner and outer scales of, 535, 570–571, 581
  Kolmogorov, 534–539
  in neutral atmosphere, 534–539
  power-law relations, 538
  spectrum of phase fluctuations, 537
  structure function of phase, 535–538
Two-dimensional array, 64–65, 147–158
Two-dimensional synthesis, 64

(u, v) plane (spatial frequency plane), 64, 70–71
  in CLEAN algorithm, 432
  coordinates, 64, 70
  coverage, *see* Spatial frequency coverage
  holes in coverage, 141
  in interference susceptibility, 617–619
  interpolation in, 127, 129, 393–398
  loci, 88–91, 151, 154, 157–159
(u', v') plane, 74–76, 90–91
(u, v, w) components, 70–71, 73, 455–458
  in fringe-frequency averaging, 617–618
  in visibility (time) averaging, 206–208
Uncertainty principle, 39, 346, 646
Undersampling, 257, 260
Uniform weighting, 391–392
Unit rectangle function, 236, 394
Universal time, 482–485
Usuda, Japan, 158
UTR-2, 154

van Cittert–Zernike theorem, 73, 594–602
  assumptions, 600–602
  derivation, 594–597
Van Vleck relationship, 262, 267
Van Vleck–Weisskopf profile, 520
Varactor diode, 231, 350
Variance matrix, 497
Velocity standard, *see* Spectral lines
Vermilion River Observatory, Illinois, 35
Very Large Array (VLA), 9, 32, 33, 218, 238, 460, 476, 552, 610
  antenna configuration, 148–150
  atmospheric phase noise, 479, 539, 540
  delay increments, 239
  dynamic range, 446
  images from, 33, 442, 443
  interference thresholds, 616, 620
  opacity at site, 543
  phased-array mode (VLBI), 370
  phase switching, 247
  self-calibration, 442–443
  (u, v) spacing loci, 151
Very Long Baseline Array (VLBA), 35, 156–158, 356
  phase referencing, 478–479
Very long baseline interferometry, *see* VLBI
Visibility
  averaging, 205–208
  complex, 27, 61
    defined, 69
  frequencies, 92–93
  fringe (Michelson), 13
  lensclean, 462
  at low spatial frequencies, 446–449, 451–453
  model fitting, 401–404
  most likely value, 319
  reduction due to phase noise, 233, 530–534, 568–569
  Taylor expansion of, 403
Visibility-intensity relationship, 68–71, 594–602
VLA, *see* Very Large Array
VLBA, *see* Very Long Baseline Array
VLBI
  antenna polarization (parallactic) angle, 116
  antennas
    nonidentical, 96–97
    in space, 158–159, 373–377
  arrays, 34–35, 155–158
  astrometry, 472–480, 499–500
  atmospheric limitations, 475, 534
  bandwidth synthesis, 366–368
  burst mode, 368–369
  calibration sources, 308, 387. *See also* Phase referencing
  clock errors, 310–314
  closure phase, 35, 328–329, 438–444
  coherence time, 305, 340–342
  coherent and incoherent averaging, 323–326, 330–332
  data encoding, 353–355
  development of, 33–37, 304–306, 575
  discrete delay step loss, 363–365
  double sideband system, 183, 196
  fractional bit shift loss, *see* discrete delay step loss
  frequency standards, precise, 342–351
  fringe detection, 329
  fringe fitting
    two-element, 319–326
    global (multielement), 326–332
  fringe rotation, 357–361, 366
  fringe rotation loss, 358–361
  fringe sideband rejection loss, 361–362
  in geodesy, 35, 485
  group delay, 314, 366

VLBI (*Continued*)
hybrid mapping, 438–440
interference sensitivity, 621–624
K-4 system, 356
local oscillator stability, 351–352
Mark I system, 305, 324, 355–356
Mark II, III, and IV systems, 355–356
masers, mapping, 485–490
networks, 34–35
noise in, 316–319
orbiting, *see* OVLBI
phase calibration system, 352–353
phased-array elements, 369–373
phase noise, 192–193. *See also* atmospheric limitations
phase referencing, 476–480
phase stability, oscillators, 332–342
polar motion observations, 484–485
probability distributions, 316–323
pulse calibration system, 352–353
quantization loss, 305, 365
recording systems, 353–356
relativistic bending measurements, 573–574
retarded baseline, 315–316
satellite link, 34
sideband separation, 183, 196
signal-to-noise ratio, 305–306, 325–326, 358–366
spectral line, 314, 364, 406–408
S2 system, 356
TID, observation of, 563–564
time synchronization, 353
triple product, 331
water vapor radiometry, 541–543, 553–554
Voltage reception (response) pattern, 78–79, 134, 599
measurement of, 636

Walsh functions, 242–246
natural order, 244
orthogonality, period of, 241
sequency, 243
Water vapor
22-GHz line, 526–527
absorption, 518–523
compressibility factor, 528–529
effect on phase, 530–534
maser, 324, 485–490
refractivity, 508–510, 528–529
resonance model, 524–528
turbulence, 534–539
worldwide distribution, 511
Water vapor radiometry, 541–543, 552–554
Waveguide, 218
w component, 70–71, 73, 91, 455–458, 468, 620
Weighting
antenna excitation, 137, 451–452
function
atmospheric, 532
spectral, 286–288, 406
natural, 191, 388, 392
of visibility, 190–191, 388–392
Westerbork Synthesis Radio Telescope, 32, 104, 144, 370, 396
Westford, Massachusetts, *see* Haystack Observatory
White light fringe, 57, 307, 644
Wide field mapping, 74–77, 204–205, 446–453, 454–458
Wiener–Khinchin relation, 54, 77, 184, 256, 284, 621

X-ray interferometry, 645

Young's two-slit interferometer, 39
Y-shaped array, 148–150

Zeeman effect, 97, 115, 348, 351,
Zenith opacity, 520, 543–547
Zero padding, 293
Zero spacing problem, *see* Short-spacing data