

IET Electromagnetic Waves Series 50

Channels, Propagation and Antennas for Mobile Communication

Rodney Vaughan and
Jørgen Bach Andersen

Geotab Exhibit 1118
Geotab v. Fractus
IPR2025-01026

IET ELECTROMAGNETIC WAVES SERIES 50

Series Editors: Professor P.J.B. Clarricoats
Professor E.V. Jull

Channels, Propagation and Antennas for Mobile Communications

Other volumes in this series:

Volume 1	**Geometrical theory of diffraction for electromagnetic waves, 3rd edition** G.L. James
Volume 10	**Aperture antennas and diffraction theory** E.V. Jull
Volume 11	**Adaptive array principles** J.E. Hudson
Volume 12	**Microstrip antenna theory and design** J.R. James, P.S. Hall and C. Wood
Volume 15	**The handbook of antenna design, volume 1** A.W. Rudge, K. Milne, A.D. Oliver and P. Knight (Editors)
Volume 16	**The handbook of antenna design, volume 2** A.W. Rudge, K. Milne, A.D. Oliver and P. Knight (Editors)
Volume 18	**Corrugated horns for microwave antennas** P.J.B. Clarricoats and A.D. Oliver
Volume 19	**Microwave antenna theory and design** S. Silver (Editor)
Volume 21	**Waveguide handbook** N. Marcuvitz
Volume 23	**Ferrites at microwave frequencies** A.J. Baden Fuller
Volume 24	**Propagation of short radio waves** D.E. Kerr (Editor)
Volume 25	**Principles of microwave circuits** C.G. Montgomery, R.H. Dicke and E.M. Purcell (Editors)
Volume 26	**Spherical near-field antenna measurements** J.E. Hansen (Editor)
Volume 28	**Handbook of microstrip antennas, 2 volumes** J.R. James and P.S. Hall (Editors)
Volume 31	**Ionospheric radio** K. Davies
Volume 32	**Electromagnetic waveguides: theory and application** S.F. Mahmoud
Volume 33	**Radio direction finding and superresolution, 2nd edition** P.J.D. Gething
Volume 34	**Electrodynamic theory of superconductors** S.A. Zhou
Volume 35	**VHF and UHF antennas** R.A. Burberry
Volume 36	**Propagation, scattering and diffraction of electromagnetic waves** A.S. Ilyinski, G. Ya.Slepyan and A. Ya.Slepyan
Volume 37	**Geometrical theory of diffraction** V.A. Borovikov and B.Ye. Kinber
Volume 38	**Analysis of metallic antenna and scatterers** B.D. Popovic and B.M. Kolundzija
Volume 39	**Microwave horns and feeds** A.D. Olver, P.J.B. Clarricoats, A.A. Kishk and L. Shafai
Volume 41	**Approximate boundary conditions in electromagnetics** T.B.A. Senior and J.L. Volakis
Volume 42	**Spectral theory and excitation of open structures** V.P. Shestopalov and Y. Shestopalov
Volume 43	**Open electromagnetic waveguides** T. Rozzi and M. Mongiardo
Volume 44	**Theory of nonuniform waveguides: the cross-section method** B.Z. Katsenelenbaum, L. Mercader Del Rio, M. Pereyaslavets, M. Sorella Ayza and M.K.A. Thumm
Volume 45	**Parabolic equation methods for electromagnetic wave propagation** M. Levy
Volume 46	**Advanced electromagnetic analysis of passive and active planar structures** T. Rozzi and M. Farinai
Volume 47	**Electromagnetic mixing formulae and applications** A. Sihvola
Volume 48	**Theory and design of microwave filters** I.C. Hunter
Volume 49	**Handbook of ridge waveguides and passive components** J. Helszajn
Volume 50	**Channels, propagation and antennas for mobile communications** R. Vaughan and J. Bach-Anderson
Volume 51	**Asymptotic and hybrid methods in electromagnetics** F. Molinet, I. Andronov and D. Bouche
Volume 52	**Thermal microwave radiation: applications for remote sensing** C. Matzler (Editor)
Volume 502	**Propagation of radiowaves, 2nd edition** L.W. Barclay (Editor)

Channels, Propagation and Antennas for Mobile Communications

Rodney Vaughan and
Jørgen Bach Andersen

The Institution of Engineering and Technology

Published by The Institution of Engineering and Technology, London, United Kingdom

First edition © 2003 The Institution of Electrical Engineers
Reprint with new cover © 2006 The Institution of Engineering and Technology

First published 2003
Reprinted 2006

This publication is copyright under the Berne Convention and the Universal Copyright Convention. All rights reserved. Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may be reproduced, stored or transmitted, in any form or by any means, only with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Inquiries concerning reproduction outside those terms should be sent to the publishers at the undermentioned address:

The Institution of Engineering and Technology
Michael Faraday House
Six Hills Way, Stevenage
Herts, SG1 2AY, United Kingdom

www.theiet.org

While the author and the publishers believe that the information and guidance given in this work are correct, all parties must rely upon their own skill and judgement when making use of them. Neither the author nor the publishers assume any liability to anyone for any loss or damage caused by any error or omission in the work, whether such error or omission is the result of negligence or any other cause. Any and all such liability is disclaimed.

The moral rights of the author to be identified as author of this work have been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

British Library Cataloguing in Publication Data
Vaughan, R.
Channels, propagation and antennas for mobile communications.
– (IEE electromagnetic waves series; no. 50)
1. Antennas (Electronics) 2. Radio wave propagation
3. Mobile communications systems
I. Title II. Bach Andersen, J. III. Institution of Electrical Engineers
621.3'845

ISBN (10 digit) 0 85296 084 0
ISBN (13 digit) 978-0-85296-084-4

Typeset in India by Newgen Imaging Systems (P) Ltd, Chennai
Printed in the UK by MPG Books Ltd, Bodmin, Cornwall
Reprinted in the UK by Lightning Source UK Ltd, Milton Keynes

To Helen and Dorte

Contents

Preface **xxvii**

1 Background and introduction to mobile communications **1**
1.1 Three elements in mobile communications 1
1.2 Original mobile antennas: firecrackers, clotheslines and chicken wire 2
1.3 Mobile and personal communications systems 9
1.3.1 Ongoing visions for systems 9
1.3.1.1 Operating environments 10
1.3.2 Traffic estimate example 10
1.3.2.1 Spectral implications 11
1.3.2.2 Power flux density 12
1.3.3 Radio data networks 13
1.3.4 Biological effects 13
1.4 Modern multi-user system requirements 14
1.4.1 System spectral efficiency 14
1.4.2 Cells for frequency re-use 14
1.4.2.1 Co-channel interference 16
1.4.2.2 Near-far effect 18
1.4.2.3 Other factors 21
1.4.3 Power variation in wideband channels 21
1.4.4 Multiple access techniques 21
1.4.5 Duplex techniques for two-way channels 23
1.5 Usage and restrictions of the radio spectrum 24
1.5.1 The rule-makers 24
1.5.2 Issues in mobile communications 25
1.5.3 Radio-link issues for research 25
1.6 Mobile channel terminology 25
1.6.1 Nonlinearities 28

1.6.2 Aspects of capacity, channel efficiency, modulation and coding 29
1.6.2.1 Energy per bit 32
1.6.3 A note on reciprocity 33
1.7 Effect of multipath on the digital channel 33
1.7.1 Short-term multipath dominating the mobile channel behaviour 33
1.7.2 Fast fading of the analogue channel 34
1.7.2.1 Narrowband and wideband mobile channels 35
1.7.3 The effect of fading on the digital channel: irreducible bit-error ratio 35
1.7.3.1 Timing errors from random frequency modulation 35
1.7.3.2 Intersymbol interference from multiple time delays of the propagation channel 36
1.8 Signal processing for mitigation of the multipath effects 37
1.8.1 Diversity basics for fading channels 38
1.8.1.1 Delay time and 'clock' time 38
1.8.1.2 Impulse response of narrowband and wideband channels 38
1.8.1.3 Broadband and wideband 39
1.8.2 Basic diversity 40
1.8.2.1 Antenna diversity for flat-fading channels 40
1.8.2.2 Adapted antenna patterns in optimum combining 41
1.8.2.3 Time diversity (rake receiver) for frequency-selective channels 43
1.8.2.4 Equalisation for ISI in frequency-selective channels 44
1.8.2.5 Linear combination of antenna branches and time-dispersed signals 45
1.8.2.6 Frequency diversity 46
1.8.2.7 Transmit diversity 46
1.8.2.8 Transmit and receive diversity 47
1.8.2.9 Convergence of multiple access, modulation, and multipath mitigation schemes 47
1.8.2.10 Slow fading and macro diversity 47
1.8.2.11 Mean path loss limits diversity action 48
1.9 Example analysis of simple digital wireless link 48
1.9.1 Free space path loss 50
1.9.1.1 Example for basic SNR calculation 51
1.9.1.2 Modulation and BER 51
1.9.1.3 Path loss and BER 52
1.9.1.4 Interference in a multi-user scheme 53

1.9.2 Non-line-of-sight 54
1.9.2.1 Non-line-of-sight and interference 56
1.9.3 Measurement-based indoor/outdoor path loss 56
1.9.4 Link example 57
1.9.4.1 Discussion 58
1.10 Summary 59
1.11 References 59
1.12 Other reading 61

2 Multipath propagation in mobile communications 63
2.1 Multipath reception and transfer function model 63
2.1.1 Polarisation of multipath fields 63
2.1.2 Fields summed by an antenna 64
2.1.2.1 Scalar model using discrete effective scatterers 65
2.1.2.2 Constant point scatterers for localised modelling 65
2.1.3 Moving receiver 65
2.1.3.1 Phase linearisation in localised model 66
2.1.4 Baseband equivalent transfer function from discrete scatterers 67
2.1.5 Fourier model using continuous scattering medium 68
2.1.5.1 Summary 68
2.1.6 Resolvability of scatterers 69
2.1.7 Time domain representation 70
2.1.8 Polarisation and antenna pattern effect 71
2.1.8.1 Static receiver with 3D effective scattering distribution 71
2.1.8.2 Moving receiver with 2D effective scattering distribution 73
2.1.8.3 Summary 74
2.2 Statistical basis of transfer functions: correlation and spectra 74
2.2.1 Channel correlation functions 75
2.2.1.1 Wide-sense stationarity and uncorrelated scattering 76
2.2.1.2 Averaging 80
2.2.1.3 Summary 80
2.2.2 Power profiles 80
2.2.2.1 Fixed position: impulse response correlation 81
2.2.2.2 Fixed position: transfer function correlation 81
2.2.2.3 Fixed direction 82
2.2.2.4 Spreads 82
2.2.3 Averaged transfer function of channel: channel gain 83

2.2.4 Envelope, complex signal, and power signal correlations 83
2.2.4.1 Narrowband Gaussian process and power correlation 84
2.2.4.2 Similarity of envelope and power correlations 86
2.2.4.3 Estimated correlation coefficient 87
2.3 References 87

3 Basic multipath mechanisms 89
3.1 Reflection from a smooth, planar surface 89
3.1.1 Basic configuration 89
3.1.2 Sum of vertical and horizontal fields at the receiver 91
3.1.3 Norton surface wave 91
3.1.4 Sum of direct and reflected space waves at grazing incidence 92
3.1.5 Fresnel coefficients 93
3.1.6 Example of Brewster angle effect with a moving receiver 94
3.1.7 Phase of the reflection at normal incidence 95
3.1.8 Circular polarisation 96
3.1.9 Image model 98
3.1.10 Grazing incidence 99
3.1.11 Propagation effects at grazing incidence 99
3.1.12 Distance dependence at grazing incidence 100
3.1.12.1 Receiver height dependence at grazing incidence 101
3.1.13 Frequency dependence at grazing incidence 101
3.1.14 Example of distance dependence: outdoor cell 103
3.1.15 Fade lengths with distance 104
3.2 Propagation along a rough surface 105
3.2.1 Kirchhoff's solution for scattering from a rough surface 105
3.2.1.1 Phase-only effects 107
3.2.1.2 Numerical illustrations 109
3.2.1.3 Effect of surface roughness on the specularly reflected field 109
3.2.1.4 Specular scattering with small roughness factor 109
3.2.1.5 Specular scattering from a surface with statistically described height fluctuations 111
3.2.1.6 Gaussian distributed surface height 111
3.2.1.7 Numerical examples of specular scattering from a rough surface 112
3.2.1.8 Limitations of the Kirchhoff model 113

3.2.2 Perturbation theory for the field scattered from a rough dielectric medium 115
3.2.2.1 Statistics of horizontal properties of a rough surface 115
3.2.2.2 Effective surface parameters for coherent field 115
3.2.2.3 Horizontal polarisation: asymptotic cases, isotropic surface 116
3.2.2.4 Vertical polarisation: asymptotic cases, isotropic surface 117
3.2.3 Effect of surface roughness on diffuse field 118
3.2.3.1 Polarisation aspects of rough surface scattering 121
3.3 Diffraction 123
3.3.1 UTD diffraction by a wedge 125
3.3.2 Transition zone boundary 132
3.3.3 UTD transition zone diffraction 133
3.3.3.1 An antenna and one screen 136
3.3.3.2 Diffraction over two screens 138
3.3.3.3 Diffraction over *N* screens by recursion 140
3.4 Ray tracing and radiosity 142
3.4.1 Ray tracing 143
3.4.2 Ray launching 144
3.4.3 Radiosity or diffuse scattering 144
3.4.4 An example of a street junction 145
3.5 References 146

4 Propagation modelling 149
4.1 Propagation in urban environments above rooftops 149
4.1.1 Okumura–Hata case for urban environments 149
4.1.2 2D diffraction models 151
4.1.3 Bridged knife edges 154
4.1.4 COST231 model 154
4.1.5 A comparison of methods 157
4.2 Propagation in rural and hilly terrain 157
4.2.1 Open undulating terrain 158
4.2.2 An integral equation formulation 158
4.2.3 Combining terrain and building effects, the parabolic method 163
4.3 3D effects 164
4.4 Indoor modelling 165
4.4.1 Parametric models 165
4.4.1.1 Model 1 165
4.4.1.2 Model 2 166
4.4.1.3 Model 3 166

4.5 Time domain features 167
4.5.1 Rural environments 167
4.5.2 Urban environments 170
4.5.3 Indoor environments 171
4.6 Angular domain features 173
4.6.1 Rural environments 173
4.6.2 Urban environments 174
4.6.3 Indoor environments 178
4.6.3.1 Experimental examples of angular power for indoors environment 178
4.7 References 178

5 Short-term channel behaviour from the two-path model 181
5.1 Introduction 181
5.2 Static terminal in a static two-source scenario 182
5.2.1 Impulse response 182
5.2.2 Delay time moments 183
5.2.3 Transfer function 184
5.2.3.1 Magnitude 184
5.2.3.2 Phase 185
5.2.4 Group delay 186
5.2.5 Mean group delay 188
5.2.6 Non-minimum phase case 188
5.2.7 Effect of finite pulse width on delay spread 188
5.2.7.1 Impulse response 188
5.2.7.2 Discussion 190
5.2.8 Probability functions 193
5.2.8.1 Summary 195
5.2.9 Effect of finite bandwidth on the fading channel 195
5.2.9.1 Statistical quantities 196
5.2.9.2 An effective value of a_2 197
5.2.9.3 Summary 197
5.2.10 Interpretation of negative group delay 198
5.2.10.1 Example 198
5.2.10.2 Mechanism for negative group delay 202
5.2.10.3 Summary 203
5.2.11 Coherence bandwidth 203
5.2.12 Coherence bandwidth–delay spread product 205
5.2.12.1 Summary 208
5.3 Dispersion metrics from time–frequency theory for the static frequency-selective channel 209
5.3.1 Terminology 209
5.3.2 Time–frequency distribution discussion 209
5.3.2.1 Introduction 209
5.3.2.2 Correlation function 212

5.3.2.3 Generalised form of time–frequency distribution 212
5.3.2.4 Kernel choice for defining distribution properties 213
5.3.2.5 Local variances: instantaneous bandwidth and local delay spread 214
5.3.2.6 Global time variance: delay spread 216
5.3.2.7 Interpretation of local delay spread 216
5.3.2.8 Example 1: $k = 1$ 216
5.3.2.9 Example 2: $k'(0) = 0$ and $k''(0) = 1/4$ 217
5.3.2.10 Example 3: $k'(0) = -1/2$ and $k''(0) = 1/4$ 217
5.3.2.11 Summary 218
5.3.3 Local approximation of transfer function 218
5.3.3.1 Single term expansion 218
5.3.3.2 Rihaczek bandwidth 219
5.3.3.3 Summary 220
5.3.4 Mean delay for a signal in a channel 221
5.3.5 Dispersion in the two-path model 222
5.3.5.1 Examples 226
5.3.6 Summary 228
5.4 Moving terminal in a static two-source scenario 229
5.4.1 Two-dimensional transfer function 229
5.4.2 Effect of finite bandwidth on the two-dimensional power transfer function 232
5.4.3 Doppler frequency moments 233
5.5 Statistics of the 'few-path' model 234
5.5.1 Three-path model 234
5.5.1.1 Application 234
5.5.2 The few-path model 235
5.5.2.1 Equal amplitudes 235
5.5.2.2 Power of in-phase and quadrature components 236
5.5.2.3 Envelope 239
5.5.2.4 Closed forms 240
5.5.2.5 Different amplitudes case 242
5.6 References 242

6 Short-term behaviour of many-path models and scenarios 245
6.1 Many-path model 245
6.1.1 Moving terminal in a static many-path scenario: linear systems model 246
6.1.2 Bello functions in mobile communications 247
6.1.3 Finite bandwidth effects on the many-path impulse response 248
6.1.3.1 Doppler spectrum 250

6.1.4 Finite length effect: Fourier angular resolution from a mobile trajectory 251
6.1.4.1 Resolution in the directions of the trajectory 251
6.1.4.2 Resolution at broadside to trajectory 251
6.2 Derivatives of the transfer function 252
6.2.1 Phase derivatives: group delay and random FM 252
6.2.1.1 Use of Gaussian conditioning theorem 254
6.2.1.2 Student-T distribution for phase derivatives 255
6.2.2 Group delay distribution 256
6.2.3 Group delay excursions with mobile position 256
6.2.3.1 Doppler frequency excursions 258
6.2.3.2 Fade depth in the 1D and 2D models 258
6.2.4 Spatial gradient of power 259
6.2.5 Dynamic two-path model 261
6.3 Doppler moments 263
6.3.1 Basic definition 264
6.3.1.1 Spatial periodicity of transfer function 267
6.3.2 Angle of arrival from a cluster of scatterers 268
6.3.2.1 Conditional probabilities 271
6.4 Correlation spacings 271
6.4.1 Correlation distance–Doppler spread product 271
6.4.1.1 Correlation distance 272
6.4.1.2 Uncertainty principle 272
6.4.1.3 A pattern for minimum spatial correlation function spread 272
6.4.1.4 Spatial periodicity 273
6.4.1.5 Summary 273
6.4.2 Correlation distance for directional antennas or scenarios 274
6.4.3 Frequency correlation 280
6.4.4 Combined space and frequency correlation 280
6.4.4.1 Summary 282
6.5 Integrating the transfer function over finite bandwidths and distances 282
6.5.1 Effect of integrating power over finite bandwidth 282
6.5.1.1 Power signal formulation 283
6.5.1.2 Spread about mean 285
6.5.2 Effect of integrating power over a finite distance 286
6.5.2.1 Integrated power 286
6.5.2.2 Spread about mean 287
6.5.2.3 Sampled power 288
6.5.3 Continuous form: space and frequency integration 293
6.6 Examples of modelled scatter distributions 293
6.6.1 Single source 294

6.6.2 Distributed line source distribution with constant amplitude and phase 294
6.6.3 Distributed line source distribution with random phase 295
6.6.4 Distributed circular line source with constant amplitude, phase and delay 296
6.6.5 Discrete uniform scenario, uniform delay (Clarke scenario) 297
6.6.6 Clarke scenario with single phase wave delay 297
6.6.7 Continuous uniform scenario with uniform phase and exponential delay profile 299
6.6.8 Discrete uniform scenario with uniform phase and exponential delay profile 299
6.7 Averaged scenario models 300
6.7.1 Basic formulation 300
6.7.2 Cross-polar discrimination (XPD) 302
6.7.2.1 Ray model from general scattering 303
6.7.2.2 Reflections from smooth, planar surfaces 306
6.7.3 Uncorrelated scatterers 308
6.7.3.1 Fourier resolution 309
6.7.3.2 Measurement by time series correlation 310
6.7.3.3 Correlation over a bandwidth 311
6.7.3.4 Correlation over time and bandwidth 311
6.7.4 Gaussian scenario model 312
6.7.5 Laplacian scenario model 312
6.7.6 Vehicular mobile 313
6.7.7 Personal terminals 314
6.7.8 Base stations 315
6.7.9 Mobile satellite 317
6.7.9.1 Geostationary 317
6.7.9.2 Non-geostationary 318
6.8 A diffuse model for outdoor environments 320
6.8.1 Parameters of model 320
6.8.2 Directionality in the simple ellipse model 322
6.9 References 325

7 Aspects of simulation and measurement 327
7.1 Short-term fading simulation 327
7.1.1 Model for narrowband signals 327
7.1.2 A simple Matlab simulation 331
7.1.3 Discretisation options for the uniform scenario 333
7.1.4 Signal repetition distance (cyclostationarity from simulations) 335
7.1.5 Delay spread example with path-loss effect 338
7.2 Reference parameters for simulations 340
7.2.1 Basic pdf for incident power 340

7.2.2 Basic received signal 340
7.2.3 Averaging 341
7.2.4 Statistical quantities for signals from a uniform scenario 342
7.2.5 Two signals 343
7.2.6 Moments 346
7.2.7 Joint pdfs of two signals 347
7.2.8 Joint envelope and joint phase pdfs 348
7.2.9 Envelope correlation coefficients 350
7.2.10 Phase correlation coefficient 351
7.2.11 Signal dynamics for moving receiver 352
7.2.12 Level crossing problems 353
7.2.13 Level crossings of diversity combined signals 354
7.2.14 Fade duration 355
7.2.15 Random FM spectrum and rate 357
7.2.16 Level crossing rate of random FM 358
7.3 Direct generation of channel functions 359
7.3.1 Gaussian independent samples 359
7.3.2 Gaussian correlated samples 360
7.3.3 General correlation matrix case 362
7.3.4 Moving average technique for generating correlated signals 364
7.3.5 Generation of transfer functions in the time (space) domain 365
7.3.6 Generation of transfer functions in the frequency domain 366
7.3.7 Continuous delay profile 369
7.3.8 Discrete delay profile 369
7.4 Envelope–phase relations for fading channels 370
7.4.1 Frequency-selective channel (network theory) 371
7.4.2 Pole-zero description of mobile signals 372
7.4.3 Zero description of fading signal 376
7.4.4 Application to single variable modulation correction in narrowband systems 377
7.4.4.1 Magnitude correction 377
7.4.4.2 Phase correction 378
7.4.5 Channel characterisation using the Hilbert relations 378
7.4.6 Delay spread limits from magnitude-only frequency transfer function 379
7.4.7 Angular distribution from spatial transfer function 381
7.4.8 Suppressing the effect of the phase ambiguity in Hilbert relations 383
7.4.9 Summary 384

7.5 Testing the multipath model – the prediction of short-term fading 384
7.5.1 Introduction 384
7.5.2 Using the propagation model for signal prediction 385
7.5.3 Noise-free case: deterministic method 387
7.5.3.1 Partitioning of u for equal Doppler contributions 387
7.5.4 Additive noise case including subspace/super-resolution methods 391
7.5.4.1 Point source model 391
7.5.4.2 Vector notation 391
7.5.4.3 Gaussian continuous source model 392
7.5.4.4 Basic linear estimator for the continuous model 392
7.5.4.5 Estimating a correlation matrix from samples 393
7.5.4.6 Estimating the sources 394
7.5.4.7 MUSIC 395
7.5.4.8 Minimum norm algorithm 395
7.5.4.9 Principal Components Linear Prediction (PCLP) 396
7.5.4.10 ESPRIT 396
7.5.4.11 Simulation results 396
7.5.5 Real-world experimental results 398
7.6 Correlation analysis for measured narrowband signals 400
7.6.1 Estimating the correlation coefficient of Gaussian signals from finite samples 400
7.6.1.1 Independent samples 401
7.6.2 Envelope correlations 403
7.6.2.1 Clipped envelopes 407
7.6.2.2 Formulation for clipping 408
7.6.2.3 Soft clipping 409
7.6.2.4 Hard clipping 410
7.6.2.5 Summary 414
7.6.3 Demeaning of envelopes 414
7.6.3.1 Additive mean 416
7.6.3.2 Multiplicative mean 417
7.6.3.3 Summary 419
7.7 Wideband channels: channel sounding 419
7.7.1 Full channel sounding 420
7.7.2 Partial channel sounding by rectangular pulse 422
7.7.2.1 Spectral occupation efficiency 423
7.7.2.2 Energy efficiency 424
7.7.2.3 Product of efficiencies 424

7.7.3 Super-resolution by deconvolution 424
7.7.3.1 Deconvolution 426
7.7.3.2 Subtractive deconvolution 426
7.7.3.3 Inverse filter 426
7.7.3.4 Instantaneous delay spread 427
7.8 References 428

8 Antenna principles 435
8.1 Basic antenna parameters and elements 435
8.1.1 Directivity and gain from patterns 435
8.1.1.1 Far field and directivity relationship 436
8.1.1.2 Distributed directivity 437
8.1.1.3 Directive gain 438
8.1.1.4 Pattern models 439
8.1.1.5 Gaussian beam 439
8.1.1.6 Pattern models from circular functions 440
8.1.1.7 Effect of beam misalignment using the distributed directivity 441
8.1.2 Antenna efficiency factors 443
8.1.2.1 Polarisation efficiency 444
8.1.2.2 Load match parameters 444
8.1.2.3 Radiation efficiency (antenna efficiency) 446
8.1.2.4 Antenna radiation efficiency 447
8.1.3 Radiation resistance from the Poynting vector 448
8.1.4 Relation between directivity, effective area, and complex height 449
8.1.4.1 Effective gain of scatterer (radar cross section) 452
8.1.5 Antenna temperature 452
8.1.5.1 Sources of noise 454
8.1.5.2 Gain-to-temperature and modelled temperature distribution for mobile satellite antennas 455
8.1.6 Noise figure contributions from connections to a receiver 455
8.1.6.1 Antenna with transmission line 455
8.1.6.2 System temperature 456
8.1.6.3 Noise figure with circuits 457
8.1.6.4 Cascade circuits 458
8.1.7 Groundplane principles for antennas 460
8.1.7.1 Source elements 460
8.1.7.2 Antenna impedances 460
8.1.8 Generic elements: dipoles and monopoles 461
8.1.8.1 Folded dipoles 462
8.1.8.2 Folded N-pole 463
8.1.8.3 Approximations for the currents on wire dipoles 465

8.1.9 Fields of a current distribution 466
8.1.9.1 Basic relations for radiation (far fields) from moving charges 466
8.1.9.2 Complete fields in terms of current distribution 467
8.1.9.3 Far-field approximation 468
8.1.9.4 Rectangular, spherical and cylindrical source coordinates 468
8.1.9.5 Practical distance to far field 469
8.1.9.6 Use of auxiliary potential functions 470
8.1.9.7 Fields for electric dipole 471
8.1.9.8 'Infinitesimal', uniform dipole 471
8.1.9.9 Sinusoidal dipole current 472
8.1.9.10 Uniform distribution dipole source 473
8.1.9.11 Magnetic sources 473
8.1.9.12 Slot feeding 475
8.1.9.13 Booker's relation 475
8.1.9.14 Cavity-backed slot 477
8.1.9.15 Self-complementary antennas 477
8.1.9.16 Combination of electric and magnetic sources for field calculations 478
8.1.10 Radiation and ohmic resistance for dipoles 480
8.1.10.1 Method of moments 480
8.1.10.2 Integration of far fields 480
8.1.10.3 Induced emf method for dipole impedance 483
8.1.10.4 Wave structure approach 483
8.1.10.5 Effect on monopole impedance of finite circular groundplane 486
8.1.11 Ohmic loss for wires 488
8.1.12 Radiation and ohmic resistance for small loops 492
8.1.12.1 Proximity effect on ohmic resistance and radiation efficiency 493
8.1.12.2 Ferrite loaded loops 494
8.1.13 Summary of field quantities and relations 497
8.1.13.1 Table of field quantities and units 497
8.1.13.2 Vector differential operators 498
8.1.13.3 Gradient (Grad) 498
8.1.13.4 Divergence (Div) 498
8.1.13.5 Divergence theorem 499
8.1.13.6 Curl 499
8.1.13.7 Stoke's theorem 500
8.1.13.8 Laplacian 500
8.1.13.9 Maxwell's equations 501
8.2 Compact elements 501
8.2.1 Introduction to antenna Q from network considerations 501
8.2.1.1 Basic definitions 501

8.2.1.2 Relation to reactance 502
8.2.1.3 Q at resonance 503
8.2.1.4 Circuit model near single resonance 503
8.2.1.5 Bandwidth near single resonance 503
8.2.1.6 Measured Q 504
8.2.1.7 Comparisons of small antennas 505
8.2.1.8 Groundplane effects 505
8.2.2 Basic limitations of small antennas in isolation 506
8.2.2.1 Combined TE and TM modes 508
8.2.2.2 Combined mode antenna synthesis 508
8.2.3 Bandwidth enhancement of small antennas 509
8.2.3.1 Discrete component basic matching circuit 512
8.2.3.2 Transmission line transformers 513
8.2.3.3 Multiple quarterwave sections 513
8.2.3.4 Tapered section 514
8.2.4 Goubau antenna 515
8.2.5 Patch antennas 519
8.2.5.1 PIFA 520
8.2.5.2 Cavity model for radiation 520
8.2.5.3 Feed impedance 522
8.2.5.4 Square patch 526
8.2.5.5 Array of square patches 528
8.2.5.6 Multiple, stacked patches 529
8.2.5.7 Low-profile dual layer patch 531
8.2.6 Antennas on a handset (free space) 532
8.2.7 Antennas on a handset near a person 533
8.2.8 Handset antennas in a random environment 535
8.3 Antennas with circular polarisation 538
8.3.1 Introduction 538
8.3.1.1 Polarisation of reflections from a smooth surface 539
8.3.1.2 Mobile satellite links 539
8.3.2 Circular polarisation parameters and relations 540
8.3.3 Patterns from space loss function for satellite links 544
8.3.4 Circularly polarised patch elements 546
8.3.5 Distributed directivity for mobile-to-geostationary satellites 548
8.3.6 Circularly polarised, scanning-mode helix 550
8.3.6.1 Radiation mechanism of scanning-mode helix 552
8.3.7 Fan patterns for vehicular mobile satellite 554
8.4 Diversity antennas 554
8.4.1 Introduction 554
8.4.2 Signal combining in diversity 556
8.4.2.1 Diversity gain 556

8.4.2.2 Diversity algorithm performance from distributions of the branch signals 557
8.4.2.3 Basic combining algorithms 558
8.4.2.4 Negatively correlated signals 560
8.4.2.5 Selection and switched combining techniques 561
8.4.2.6 Predetection and postdetection in non-switched combining: equal gain 563
8.4.3 Signal statistics of combined signals 564
8.4.3.1 Maximum ratio 564
8.4.3.2 Uncorrelated branches with different mean powers 564
8.4.3.3 Uncorrelated branches with common mean branch powers 565
8.4.3.4 Selection combining 566
8.4.3.5 Diversity gain relation between selection and maximum ratio 567
8.4.3.6 Equal gain combining 568
8.4.3.7 Special cases of combining with correlated branches: two branch 568
8.4.3.8 Equivalent equal branch powers for two correlated branches 569
8.4.3.9 Symmetric three-branch case 569
8.4.3.10 Switched diversity 571
8.4.4 Effect of correlated noise (interference) 571
8.4.5 Incident fields and antenna conditions for diversity 573
8.4.5.1 Incident fields and diversity antenna polarisation matrices 575
8.4.5.2 Correlation from the incident power distribution and antenna patterns 575
8.4.5.3 Mean power from the incident power distribution and antenna pattern 576
8.4.6 Antenna patterns for diversity 577
8.4.7 Space diversity 578
8.4.8 Angle diversity 579
8.4.8.1 Rule of thumb for Gaussian beams in circular coordinate system 581
8.4.8.2 Correlation function for angularly displaced beams: Gaussian beams for high directivity 581
8.4.8.3 Correlation function for angularly displaced beams: Circular function beams for low directivity 583
8.4.8.4 Realistic patterns in an ideal uniform scenario 585
8.4.9 Polarisation and field components 585

8.4.10 Energy density in the multipath field 585
8.4.11 Gain reduction caused by closely spaced elements 588
8.4.11.1 Directivity of linear array factor 588
8.4.11.2 Gain reduction from mutual coupling in a diversity array 590
8.4.11.3 Mutual impedance 590
8.4.11.4 Calculation of mutual impedances 591
8.4.11.5 Mutual impedance of monopoles 592
8.4.11.6 Mutual impedance from minimum scattering antennas patterns 593
8.4.11.7 Network model for diversity antennas 594
8.4.11.8 Loaded circuit and open circuit correlations 595
8.4.11.9 Received power for resistive loads 595
8.4.11.10 Optimum diagonal load resistance 597
8.4.11.11 Two-element array example 598
8.4.11.12 Received power using two-branch optimum load resistance 599
8.4.12 Examples with measurement techniques 600
8.4.13 Example: monopoles on a groundplane 600
8.4.13.1 Four in-line monopoles 604
8.4.13.2 Experimental estimation of diversity gain by equivalent number of ideal branches 604
8.4.14 Example: sloping monopoles on a groundplane 607
8.4.14.1 Element patterns 607
8.4.14.2 Distributed directivity 609
8.4.14.3 Radiation resistance 610
8.4.14.4 Correlation coefficient 612
8.4.14.5 Diversity gain 613
8.4.15 Example: patch antennas on a groundplane 615
8.4.16 Example: switched diversity by switched parasitic elements 617
8.4.17 Example: polarisation diversity 620
8.4.17.1 Mobile terminals 620
8.4.17.2 Base stations 622
8.5 References 623

9 Array antennas in a multipath environment 629
9.1 Introduction 629
9.2 Adaptive antennas in cellular networks 631
9.2.1 Interference rejection 631
9.2.1.1 Spatially uncorrelated signals 632
9.2.1.2 Spatially correlated signals 634
9.2.1.3 Direction of arrival, DOA 636
9.2.1.4 User discrimination: uplink case 637
9.2.1.5 Downlink case with unknown channel 639

9.2.2 Gain enhancement 639
9.2.2.1 Uncorrelated case 640
9.2.2.2 Correlated case 640
9.2.3 Discussion 641
9.3 Multi-element arrays: MIMO systems 642
9.3.1 Singular value decomposition, SVD 642
9.3.2 Maximum gain and transmit–receive diversity for a known channel 645
9.3.3 Maximum gain and transmit–receive diversity for an unknown channel 647
9.3.4 Summary of gain and diversity for multi-arrays 648
9.3.4.1 Gain and capacity impact on data rates, range and frequency 649
9.3.5 Spectral efficiency of parallel channels 650
9.3.5.1 Capacity for M transmit antennas and one receive antenna 653
9.3.5.2 Capacity of the (M, N) array 653
9.3.6 Effect of correlations and pinholes 654
9.3.6.1 Effect on gain and diversity 655
9.3.6.2 Examples of correlations and pinholes 657
9.4 Outage for optimally combined receiving arrays with many interferers 659
9.4.1 Introduction: analogue and digital outage 659
9.4.1.1 Performance of optimum combining 659
9.4.2 Statistical model of receiving array with interferers 660
9.4.2.1 Transfer function and its covariance 661
9.4.2.2 Array weights and output power from the sources 661
9.4.2.3 Noise 662
9.4.2.4 Sample covariance matrix 662
9.4.2.5 Optimum weights for maximum SINR and minimum MSE 662
9.4.3 SINR for a larger number of interferers than array elements 663
9.4.4 Cumulative distribution function for SINR 665
9.4.4.1 Case for fixed interferer powers 665
9.4.4.2 Common, fixed interferer powers and no noise 665
9.4.4.3 Case for variable interferer powers 665
9.4.4.4 Example 668
9.5 Capacity outage for transmit and receive arrays 669
9.5.1 Introduction 669
9.5.1.1 Basic MIMO capacity formula for unknown channel at the transmitter 671

9.5.2 Statistical model including correlated, Rician channels with different SNRs 672
9.5.3 Method for computing the approximate cdf of capacity 672
9.5.4 Asymptotic expansion for cdf 674
9.5.5 Asymptotic capacity and variance for Rayleigh channels 675
9.5.5.1 Rayleigh uncorrelated case 675
9.5.5.2 Rayleigh, correlated case 675
9.5.6 Coefficients for capacity cdf with percentile outage for Rayleigh, uncorrelated channels 676
9.5.6.1 Example 677
9.5.6.2 Mean capacity expression 678
9.6 References 680

Appendix A Field strength and path loss 683

Appendix B Basic statistics for mobile communications 685
B.1 Probability and statistical independence 685
B.2 Probability density function: fundamental theorem and transformation 685
B.3 Expectation and covariance 686
B.4 Correlation (autocorrelation) and ergodicity 687
B.5 Correlation functions and a deterministic channel 688
B.6 Wide-sense stationarity (WSS) 689
B.7 Power spectrum 689
B.8 Cross-correlations 690
B.9 Covariance spectrum 691
B.10 Uncorrelated noise example 691
B.11 Variance of mean estimate 692
B.12 Maximum likelihood and the Cramér–Rao bound for the variance of an estimate 692
B.13 Chi-square test for pdf 695
B.14 Kolmogorov–Smirnoff test for cumulative density function 696
B.15 References 696

Appendix C Gaussian-derived distributions in mobile communications 697
C.1 Gaussian 697
C.2 Rayleigh 700
C.3 Rice 704
C.4 Rice envelope 706
C.5 Calculating the Marcum Q function 710
C.6 Rice phase 712
C.7 Rice generalisations: Hoyt and Beckmann distributions 712
C.8 Random phasor plus Rayleigh 715

C.9 Lognormal 715
C.9.1 Example: lognormal plus Rayleigh 716
C.9.2 Relation between first moments 716
C.9.3 Decibel units 716
C.10 Suzuki 719
C.11 Rice with lognormal mean (mixture distribution) 720
C.12 Nakagami 720
C.13 Gamma (chi-squared) 721
C.14 Generalised gamma 723
C.15 Additive mixtures with non-Gaussianity 726
C.16 Middleton's Class A impulsive noise distribution 727
C.17 Diversity distributions 727
C.18 Multivariate Gaussian 728
C.19 Chi-squared 728
C.20 Maximum ratio combination of Rayleigh envelopes 733
C.21 References 733

Appendix D Fresnel zones 735

Appendix E Group delay equivalence in the time and frequency domains 737
E.1 Reference 738

Index 739

Preface

Non-line-of-sight wireless services

The development of mobile communications over the past couple of decades is popularly referred to as the wireless revolution. This is because mobile communications have led to the widespread usage of wireless personal terminals for both voice and data services, and because the wireless terminals have replaced, for many people, the wired telephone as the first choice for voice, and even internet, connection. In this sense, mobile communications systems have revolutionised connectivity between people, and have also allowed new options for communicating with remote equipment and sensors. The promise of the future generations of such systems is further increases in the digital communications capacity with new types of services.

Systems which can provide an increasing capacity for an increasing number of users are being made possible through new scientific results and technologies. These demand-driven technologies strive to use minimal radio spectrum per user because the spectrum is a shared, finite resource.

The growing number of users means that most wireless connectivity must be via non-line-of-sight paths. Such wireless channels experience multipath propagation with the associated degradation from fading gains and changing phases. So for mobile and multiuser communications, classical link parameters such as path loss and antenna gain take on a statistical nature, and a statistical approach to their description and measurement is required.

The increasing importance of antennas with signal processing

Increasing link capacity within fixed spectral constraints requires more sophistication in the signal processing and the electronics at the link terminals. This extra signal processing is to make more efficient use of the transmission medium. Especially with wireless, multipath propagation is a fundamental and capacity-limiting problem for spectrally efficient information transmission. The multipath propagation mechanisms therefore need to be thoroughly understood in order to design the new generations of efficient wireless systems.

From a receiver's viewpoint, the effect of the antenna pattern is difficult to separate from the effect of the distribution of the multipath components arriving at the antennas. This means that antenna pattern parameters crop up within the radiowave propagation topics. From an analysis viewpoint, the antennas and the propagation effects combine to define the wireless channel. The resulting channel behaviour and limitations in its modelling also require a thorough understanding in order to design effectively for digital communications.

Finally, the antennas themselves are taking an increasing role in telecommunications systems. This is because the antennas, in particular multiport antennas with signal processing, offer a profound increase in the spectral efficiency of wireless links. So designers must also understand the principles of antenna elements, diversity array design, and array performance. This returns us to the subject of radiowave propagation, because design and deployment of high gain and adaptive antennas for mobile, multiuser communications requires an understanding of the multipath propagation.

A book of three subjects

The goal of the book is to provide the reader with a basis for innovative design within the physical layer of mobile communications and other fading channel links. From the above discussion, this book therefore targets the three basic subjects in physical layer wireless communications: channels, propagation and antennas. Strictly speaking, this requires three books! However, for mobile communications, the antennas, propagation and the radio channel are so inter-related that much insight would be lost by separating them completely. The overlapping nature of each subject is the cornerstone of engineering non-line-of-sight systems with high spectral efficiency.

The approach of this book is to emphasise principles which are to be understood, rather than presenting a handbook of quickfire recipes. However, the tools required, such as the mathematics – in particular the statistics – tend to be treated in more of a handbook style. The authors have tried to avoid the details and derivations of established mathematical results except where considerable insight is on offer.

Coverage of the various topics draws on a wide variety of communication engineering, physics and statistical sources. These disciplines all tend to have their own historic terminology and notation, and where possible, the different sections follow the convention of the primary discipline. Although this results in some unavoidable variety in the notation across the book, the terminology should be nevertheless clear within each section.

A reading of the appropriate sections will deliver an understanding of the physics and effects of multipath, the use of multipath channels for communications, and the use of both elemental and array antennas to launch and receive the radio waves efficiently. Specific parts of the text are suitable for specific electrical engineering courses including postgraduate courses. There is sufficient depth for experienced design engineers to benefit from the new material, but system designers and engineering managers will also find this text useful for gaining a sound understanding of

the physical channel limitations and the opportunities for multiple antenna systems in wireless communications.

Contents

Chapter 1 introduces the concepts with a brief look at a century of developments which, from the start, have featured non-line-of-sight paths. This goes back to Rutherford's observation that radio waves travelled through or around brick walls, and Marconi's experiments that demonstrated that radio waves could propagate effectively around the world instead of travelling along a straight line into space. The basic effects of the multipath propagation on digital communications links are also introduced in Chapter 1. Chapters 2 to 4 cover propagation and scattering mechanisms, in particular the multipath statistical relations for the short-term multipath; the scattering mechanisms of reflection, rough surface scattering and diffraction; and their experimental analysis and effects. The propagation modelling includes both the short-term (standing wave) effects and the long-term (shadow and space loss) effects.

Chapters 5 to 7 cover short-term multipath channel behaviour, including odd phenomena that occur such as negative group delay. These chapters draw on simple propagation models with signal processing concepts. The multipath propagation models comprise distributions of point source scatterers for multipath scenarios. Topics include communications aspects such as the effect of the changing channel on synchronisation, antenna pattern effects related to diversity, simulation techniques and reference results which are useful for checking aspects of simulations, and aspects of channel sounding.

Chapters 8 and 9 concern the principles and design of antennas, a deceptively large set of topics. We start with the basics of element theory, antenna and polarisation parameters, and ground plane effects. The elements comprise linear wires, loops, and slots (patch antennas). The analysis and design of diversity antennas follow, including a variety of examples. This leads to antenna arrays for multipath environments, including transmit and receive diversity, and MIMO (multiple input, multiple output) systems. The large improvement that such systems can bring to the communications spectral efficiency is also treated, including SINR outage and 'potential capacity' outage for MIMO systems. Such calculations relate to a potential (ergodic) capacity, rather than a realisable capacity. They assume that there is a perfectly coded signalling system in place, that the communications signal processing is working ideally and finally, that the physical channels, with interference and noise, follow well-defined statistical models. The non-line-of-sight modelling includes concepts such as propagation through a keyhole, the transmission pipeline equivalent of propagation between sets of local scatterers via a common isolated scatterer.

With single port antennas, the multipath-degraded channel must be accepted, and the engineer must either try to effectively over-design the link budget – for example with a narrowband system, having the signal-to-noise ratio excessive for most of the time in order to allow for the fades – or else fix a badly behaving channel by extensive coding. The multiple antenna approach is, in principle, more sound: the

antenna signals can be combined to repair the multipath-degraded channel as seen by the transmitter and receiver, and thereby make available a much-improved channel which can then be coded appropriately. In practice, the signal coding can be used for some forms of signal combination. The realisable capacity of economically feasible multiport systems is an ongoing research and development topic in communications.

Acknowledgements

This book was originally invited by the late Professor James R.Wait. Like many projects, it would probably never have been started if we knew how long it would take, and the patience of those who have been involved is appreciated. Rodney Vaughan acknowledges Industrial Research Limited, New Zealand, and Jørgen Bach Andersen acknowledges the Center for PersonKommunikation at Aalborg University, Denmark. RGV thanks several of his former and present IRL colleagues, in particular, Neil L. Scott, for their ongoing support, discussions, and assistance with research and computing. RGV also thanks Miles Leonard-Taylor for assistance with some of the diagrams. He also acknowledges the New Zealand Foundation for Science, Research and Technology for supporting his research programs at IRL in communications technologies and communications signal processing. JBA thanks many of his present and former students and colleagues at CPK for their cooperation, especially Patrick Eggers, Preben Mogensen, and Gert Frølund Pedersen to mention a few. He also acknowledges the fruitful international cooperation through the various European COST programs, 207-231-259-273.

Finally, we owe much to our wives, Helen and Dorte, for their love, tolerance and patience during our prolonged work on this project. The time missed in being together with our families during the research and writing for this book can never be replaced. Thank you, Helen and Dorte.

Chapter 1

Background and introduction to mobile communications

This background and introduction sets the scene for an appreciation of the issues of mobile communications. The chapter opens with a look at the basic enabling technology for modern, multi-user mobile communications, and then gives a glimpse of the extraordinary history of the antennas and how infrastructures have developed. The electromagnetic propagation behaviour and the antenna possibilities emerge as the critical, interconnected issues, and are common to a large variety of mobile and personal communications systems.

1.1 Three elements in mobile communications

To the user, a mobile or personal communications system is simple: it is a terminal, such as a telephone, that uses a radio link instead of a wire link. The visible impact is that the terminal is compact for portability, and that it has an antenna, although for personal communications the antenna is often no longer an explicit feature.

To the communications engineer, however, the user's terminal in a multi-user system is one component in a vast, complex, circuit. These circuits are the conduits for voice, data and information access for the users. Three elements are central to realising the circuit of a modern mobile and personal communications system: *microelectronic componentry*, especially for the portable terminals; the allocation and *use of the radio spectrum*; and an *infrastructure* to allow coordination within the communications system.

Microelectronics technologies have allowed sophisticated communications signal processing circuits to become viably miniature and efficient. The functional possibilities and quick design and availability of microelectronic designs have now served to focus on the need for the infrastructure to support the visible service. It is aspects of this infrastructure that now contribute to the bottleneck for the expansion of mobile and personal communications systems. Other principal contributors to this bottleneck are the costs and power consumption. The cost of a terminal will ultimately fall

victim to mass production techniques, but the power consumption is more of a fundamental technology issue. Continuing increases in the efficiency and sophistication of the digital signal processing, RF processing, and other technology such as increased energy density for batteries, will continue to foster an increase in the usage of personal communications systems. If the expansion of wireless device usage continues at its current rate of growth, interference between users competing for spectral space will become the main focus of affordable, global coverage. In some mobile and personal communications, interference is already a limiting factor to growth.

The infrastructure of a modern, portable, communications system includes not only the source and destination identification and signal routing of traditional telephony, but also dynamic channel allocation and power control for the portable terminal, connections to other services such as the Public Switched Telephone Network (PSTN), carefully planned radio systems (base station sites, structures and antennas), electromagnetic compatibility, etc., and typically a more complex marketing strategy for the 'new' product. The infrastructure is contained both in the turn-key communications systems that service providers purchase, and in the day-to-day running that the service provider undertakes. The 'wired' parts of infrastructures do not act to limit the service *per se* by capacity limitation, but limit the number of base stations in less populous (i.e. less wired) areas. The 'unwired' aspects of the infrastructure, such as the issues of radio coverage, spectrum usage, system standardisation and roaming agreements, comprise a large effort behind 'seamless', global coverage.

While the technical feasibility of implementing mass wireless communications has thus focused on the infrastructure, much of this focus is being increasingly directed on to the diminishing resource of the radio spectrum, and the growing need for its efficient utilisation. Coding and modulation advances over the past 10 years have allowed the realised, static channel capacity to come reasonably close (within fractions of a dB in the signal-to-noise ratio) to the Shannon limit. However, the dynamic channels often found in mobile communications require more technological progress to reach a similar efficiency. But the issue of system efficiency and spectrum usage also depends on the required radio frequency power levels, which in turn depend on the system configuration (cell size, re-use pattern), local geography, etc.

Finally, efficient spectral usage comes down to basic limitations of controlling radiowave behaviour in complex environments, which is the radiowave propagation problem; and how efficiently the radiowaves can be gathered (and launched), which is the antenna design problem. The antenna design is not just concerned with electrical efficiency – it is a strong influence in the evolution of most communications systems, and this is especially so with personal communications. Compactness and efficient use of multiple ports are important aspects of the antenna technology.

1.2 Original mobile antennas: firecrackers, clotheslines and chicken wire

The history of technologies will often provide a clue for the future. Mobile antennas, which have provided the most visible evidence of the mobile communications systems, offer an irresistible glimpse of the past.

From 1820 to 1870 the basics of electric and magnetic fields were developed in Europe and Britain with the landmark progress linked to Ørsted, Ampere, Faraday and Maxwell. The so-called Maxwellians – Fitzgerald, Poynting, Heaviside, Thomson and Larmar – supported Maxwell's controversial theories on electromagnetic waves, but experimental evidence was hard to come by and this fuelled the controversy. Around 1886, some 22 years after Maxwell's theories were published, Hertz succeeded in generating electromagnetic waves. His transmitter used spark discharge, and the detection used similar apparatus comprising a condenser made of metal rods and a metal loop.

Hertz's experiments confirmed Maxwell's theories for the existence and velocity (being the same as the speed of light) of the mysterious, new waves. Although Hertz apparently saw little future for his otherwise exciting experimental phenomena, others did. By 1894, in New Zealand, Rutherford found that he could detect the newly discovered 'Hertzian waves', even after they had passed through a brick wall. This work led Rutherford's experimental physics career to work with Thomson in Britain and, at least for Rutherford, to more promising directions.

It was Marconi, similarly inspired by the Hertzian waves, who saw the potential for wireless signalling, and his historical work began in Italy. Without a relatively prolonged education, Marconi, by 1895, had managed to get a transmitter and receiver going. Although Hertz observed directional gain for his 80 MHz waves by using parabolic reflectors, it was Marconi who effectively invented the radiowave antenna ('. . . a wire, or capacity aerea [sic], placed at a height above the ground and also in connecting at the receiver end . . .') (Marconi, 1909). In fact, the first mobile antenna installation is attributed to Marconi, who in 1897 fitted an antenna to a vehicle in Chelmsford, England. (Some two years before this, Marconi demonstrated radiowave communications using a ship-borne antenna for one end of the link, which could also be considered a mobile installation!) The vehicle was a steam bus and the arrangement is shown in Figure 1.2.1, where the cylindrical wire frame antenna resembles a large chimney, and has been referred to as the 'firecracker' antenna, after its resemblance to a huge firecracker. The large size is partly due to the low frequencies used by Marconi's AM radio system. From a radio communication viewpoint, Marconi used a spark transmitter, which is a wideband incoherent source that would provide serious interference problems in today's environment. Later, the Poulsen[1] arc transmitter allowed coherent narrowband transmissions and was used for (relatively!) spectrally efficient radio telephony from 1907 until the 1930s. This first mobile installation demonstrates a bother which remains, albeit to a lesser extent, today – the antenna is a nuisance! Much of the following is from standard historical sources as well as from Jakes, 1974; Rowan and Altgelt, 1987; and Altgelt, Duffy and Rowan, 1988.

The obvious need for a lower-profile antenna resulted in the 'clothesline' antenna, depicted in Figure 1.2.2, from which the origin of the name is obvious. These installations (i.e. the antennas) were from around 1914 in America. In 1923, a

[1] After Valdemar Poulsen (1869–1942) who patented a radio frequency adaptation of William Duddel's continuous audio frequency 'singing arc' (1899).

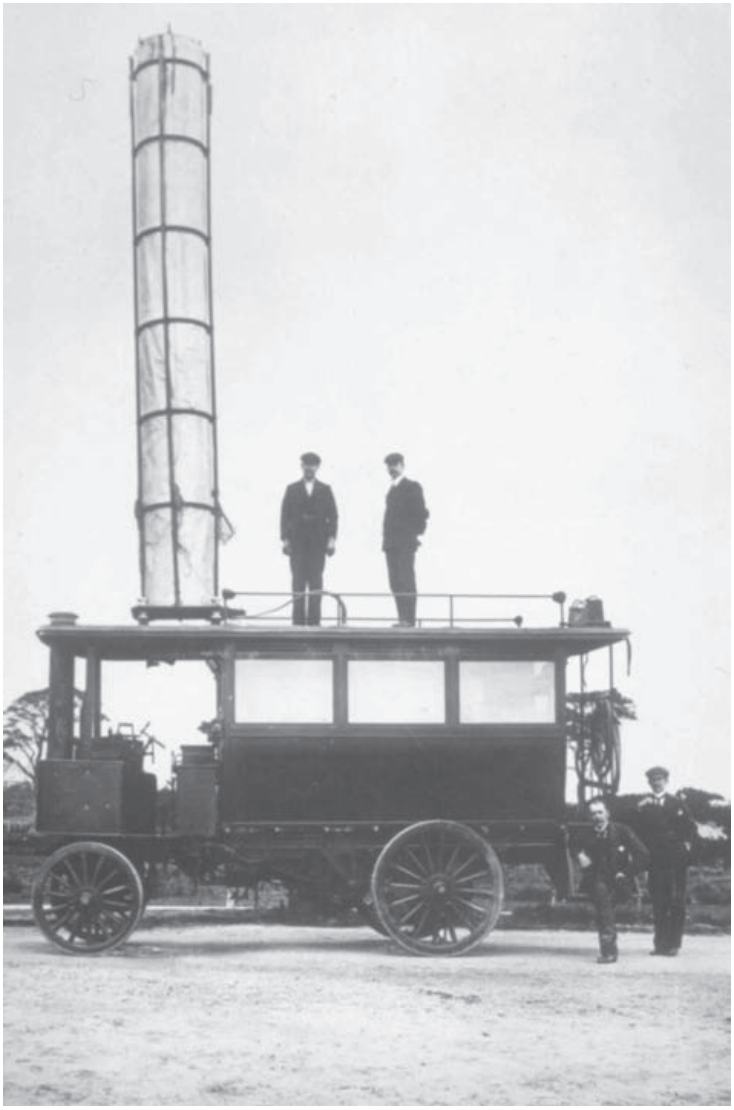


Figure 1.2.1 The first mobile antenna installation, the Thornycroft steam bus at Chelmsford, England, with the 'firecracker' antenna, and its inventor, Marconi, at the right

Californian bus company installed radios in its buses, again using the clothesline antenna.

By 1925, a smaller antenna, the 'frame' antenna comprising several tens of wire turns on an approximately quarter square-metre wooden frame, was being mounted on commercial buses in Chelmsford (Figure 1.2.3) and on some private cars. These installations were for reception of commercial radio.

Private (dispatch type) radio began to emerge at the same time and also used the clothesline antennas. Tests in America by the Detroit Police Department in 1921 led to permanent mobile operation by 1928. These systems worked at about 2 MHz.

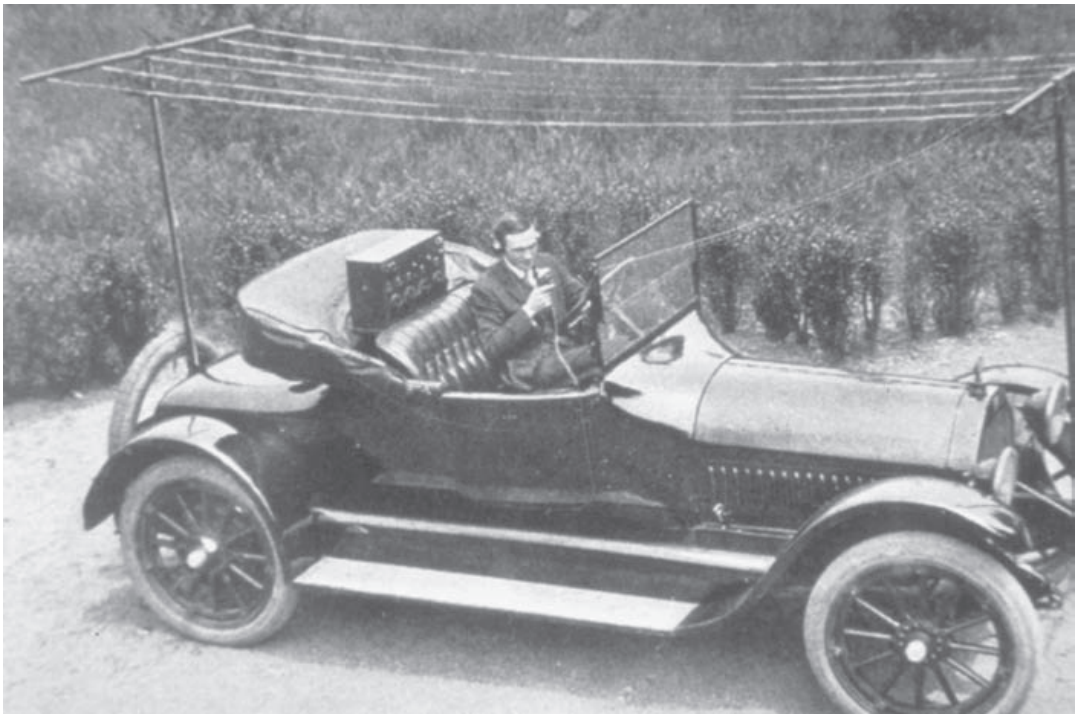


Figure 1.2.2 The 'clothesline' antenna, from around 1914 in America, provided a lower profile for the low-frequency reception

By 1930, the clothesline antenna had evolved towards a wire-netting ('chicken wire') insert in the fabric roofs of private cars. These installations were called chicken wire antennas, as depicted in Figure 1.2.4. They were probably the first serious attempt to conceal antennas for commercial reasons.

The popularity of both broadcast radio and the car in America fostered more innovation. Further attempts to conceal the antennas resulted in wire antennas mounted under vehicles, including under the running boards. Installation difficulties and a susceptibility to road conditions presented mortal problems for these antennas. In 1933, an attempt was made by Ford to use the externally attached spare wheel as an antenna, although this also turned out to be impractical and was short-lived. At about the same time, two-way radio was used by the New Jersey Police.

As antenna evolution continued, the inventions became, alas, less colourful, but necessarily more efficient and compact. By 1936, inventors returned their attention to the roof of the car, with a conductor mounted along the length of the centre of the roof, supported by insulators on suction cups. The result was a reasonably low-profile design which was essentially a horizontal dipole. At about the same time, the rod antenna began to appear in various forms, and by the late 1930s it had become, at least in a generic sense, an industry standard. The family members included mechanical deployments such as a spring-loaded 'flip-up' arrangement (by Ford in 1937), and both manual and vacuum-operated telescopic versions (by Buick and Cadillac, around 1940). The customary popular names for these antennas included 'fishing pole' and 'buggy whip' antennas. Inclusion of fashion became more obvious with



Figure 1.2.3 A frame antenna, comprising several wire loops, mounted on a Chelmsford bus around 1925

designs such as illuminated rods (by Ford, in 1937) and the equally blatant 'Clipper' antenna (a more descriptive name was understandably elusive), by Galvin, in 1937, which is best described by diagram, in Figure 1.2.5.

Currently, the whip antenna, comprising either a short monopole element, or else a normal mode helix element (known as the 'rubber duckie' when wound on a flexible, rubber core), has become a standard antenna for vehicular and personal communications. However, renewed efforts to conceal the antenna have resulted in patch-type antennas and other low-profile designs being applied to mobile terminals. Also, the front and back windscreens have offered an alternative site for mounting wire antennas, the first appearing to have been by Pontiac in 1969. Other proposals for broadcast reception have included roof-mounted slots (Nagy, 1988) and planar structures embedded in fibreglass roofs (Figure 1.2.6), and for the typically higher

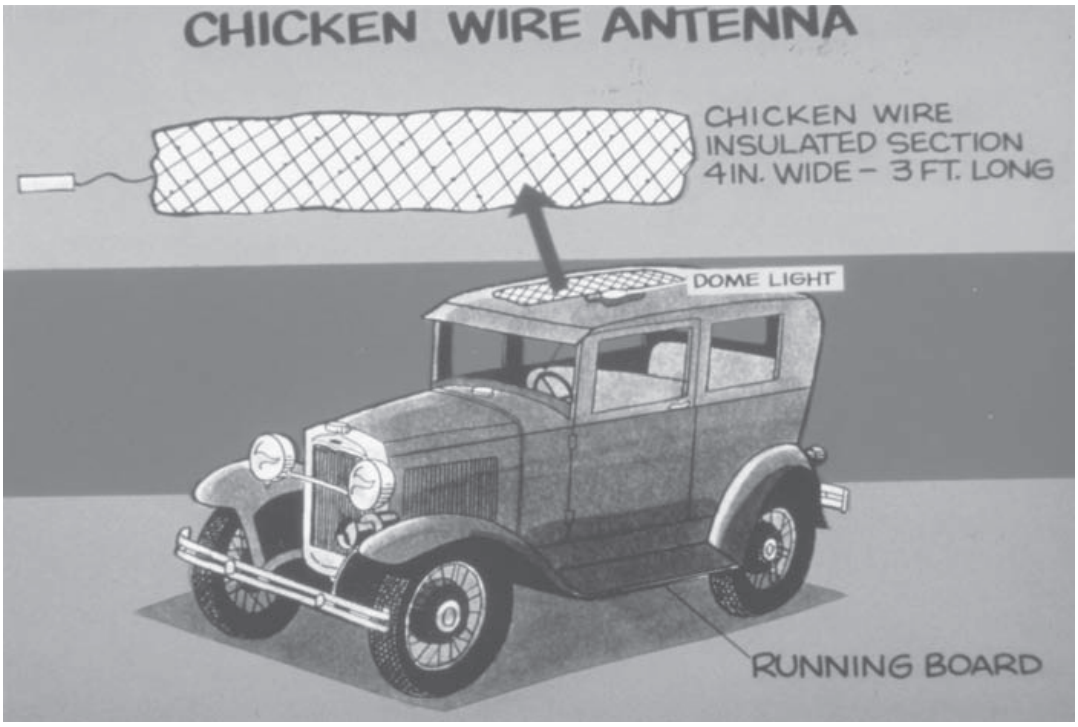


Figure 1.2.4 *The first concealed antennas? The wire-netting was fixed into the fabric top of private cars by 1930. A recent rendering of the concept is in Figure 1.2.6*

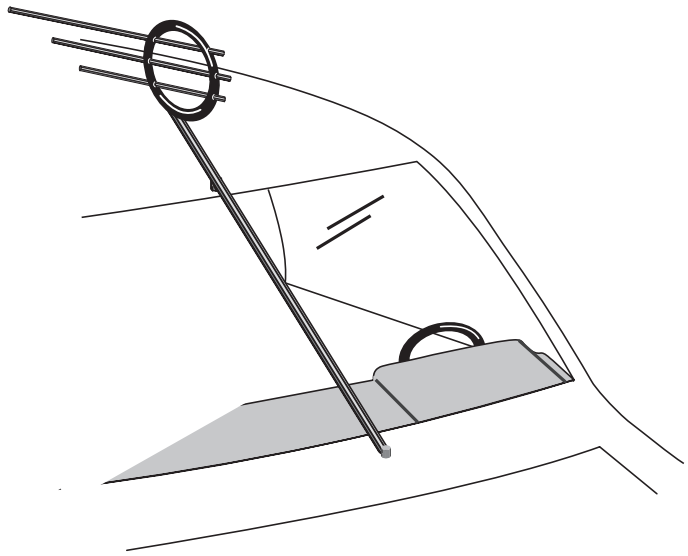


Figure 1.2.5 *The 'Clipper' antenna, from 1937, was an early attempt to make the antenna a fashion feature which matched the style of the cars of the day*

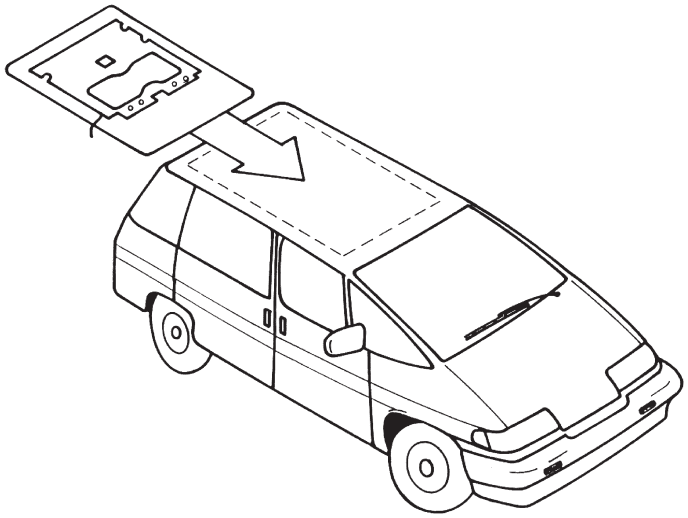


Figure 1.2.6 Planar structure in a fibreglass roof for coupling to external features. The system was used by some 1992 vehicles from General Motors (Machine Design, 1992)

frequencies of mobile communications, internally mounted antennas such as ceiling-mounted patches (Mushima and Taga, 1980), roof-mounted patches (Vaughan and Bach Andersen, 1984), Yagi antennas (Kagoshima, Takahashi and Ebine, 1991), and transparent, window-mounted, mesh patch antennas (Wu and Ho, 1991).

Future mobile communication systems will (indeed, as some current public systems already do) incorporate multiple port antennas, which revives the design problem of minimising (to marketable dimensions) the physical size and presence of the antenna.

The antennas for mobile communications have been a major problem for both the systems designers and marketers of mobile systems, as well as for the practical installers of vehicular terminals. While the early broadcast reception systems of the 1920s used frequencies around 2 MHz, smaller elemental antennas for mobile communications became possible in the 1930s with the development of the technology and infrastructure for the 30–40 MHz band. The trend continued in mobile communications with 150 MHz in the 1940s and 450 MHz in the 1950s until the present, with most modern mobile systems operating above 800 MHz.

In personal communications, the antenna is arguably a greater problem still, since it must be integrated within a light, compact, 'go-anywhere' terminal. Early personal antennas were restricted to monopole-like structures, usually mounted on back-pack transceivers for military use. More recently, microwave patch-type antennas for handhelds have appeared, and miniature pagers, operating at lower frequencies (typically 160 MHz), rely on the electromagnetic energy being coupled to and from the body to create more efficient reception.

Although broadcast transmission or base station antennas for mobile communications have not been discussed here (their history is relatively bland), multiple port antennas for base stations are included in later chapters.

1.3 Mobile and personal communications systems

The personal communications terminal is compact – or, more descriptively, 'pocketable'. The term *mobile communications* logically covers the concept of personal communications. However, mobile communications has evolved to imply a strong vehicular flavour, in particular land vehicles (cf. land–mobile communications implying cars, buses, trucks), but also covering aeronautical and marine vehicles using links via both land base stations and satellites. Mobile vehicular communications is not generally designed for operation within buildings, etc., whereas personal communications could be expected to have no such 'everyday location' limitations.

Mobile communications has also evolved to imply a two-way capability, with paging services an exception. Radio broadcasting (to vehicular mobiles or pedestrians) is sometimes excluded from the subject of mobile communications in its modern context. However, broadcasting to moving receivers involves most of the same problems for both antennas and propagation as for two-way communications. This is particularly obvious for the newer, direct broadcast services by satellite, for example. Various mobile systems include:

- cordless telephones and telephony;
- radio telephone;
- cellular radio-telephone;
- pagers;
- personal communication networks;
- aeronautical services;
- navigation services;
- broadcast radio;
- radio local area networks (RLANs);
- mobile satellite communications;
- mobile computing systems.

1.3.1 Ongoing visions for systems

In the longer term, high-capacity (tens of Mbits s^{-1}) terminals should become 'go-anywhere' personal ones. The 'video wristwatch terminal' is now a technical possibility, at least for short-range communication and short battery life, but the infrastructure allowing it to operate anywhere in the world, which also involves regulatory and political issues, represents a formidable hurdle. Realising such visions will require, in particular, a working understanding of the possibilities and limitations of the radiowave propagation at the frequencies of interest, and the principles of compact antenna techniques.

Various acronyms have been used for future systems. Such terms as UMTS (Universal Mobile Telecommunications Services), FPLMTS (Future Public Land Mobile Telecommunications Systems), and IMT-2000 (International Mobile Telecommunications) have been used to describe visions for system planners and regulators to focus on. Such visions have a role for both technologists and business society. The concepts aim to recognise the need for mobile terminals for both data

and voice to roam between networks in different countries, including use for mobiles, including ships and aircraft, moving between countries.

The continually updated general objectives of the ITU (International Telecommunications Union) include coverage for a wide range of user densities and coverage areas, application for fixed services, usage of a variety of terminals, networking issues, and compatibility and interconnection with other services. Technically, the primary objectives are to integrate the signalling and communication and to follow the open system interconnection (OSI) model for interface standards. Operationally, the objectives are for unique user identification and charging, and the ability for each terminal to both initiate and receive service as required. Requirements for regional and global compatibility include uniformity in the architecture, control procedures, modulation methods, data rates and channel radio frequencies. Quality of service (QOS) is also an issue. The quality of new speech services, for example, must be at least equivalent to that of the traditional Public Switched Telephone Network to gain user acceptance. These basic requirements assist in defining radio system parameters such as power budgets, allowable bit error rates for given modulation, and coding schemes and outage probabilities.

1.3.1.1 Operating environments

Operating environments have been typically classified along the lines of: residential; indoor office; vehicle with a mobile base station; mobile (i.e. vehicle); and portable (i.e. pedestrian mobile). Ongoing research aims to characterise and model these environments in terms of propagation path loss, penetration loss, multipath effects (shadow and Rayleigh-like fading, delays, and interference), traffic density and traffic asymmetry (ratio of initiated to received calls). Each of the environments has *radio infrastructures*, which are classified along the lines of *spot* (single cell), *island* (a collection of cells providing coverage from a neighbourhood to a major urban area) and *contiguous* (everywhere in an intended service area). On a geographical scale, *regional* implies multiple countries, and *global* and *national* are self-explanatory.

1.3.2 Traffic estimate example

Based on current understanding of the nature of personal communications, some examples of traffic estimates for voice/data can be formulated. The example of this section is early (CCIR, 1992b, pp. 16, 17) but it remains highly representative of the traffic-oriented approach to multi-user system capacity. Paging-type traffic is considered negligible. Voice traffic in cities is estimated by assuming a mean value of 400 (vehicular) mobiles per kilometre of street, half of which have the terminal operating[2] at 0.1 Erlang. This results in a traffic density of

[2] In this example from the CCIR (now called the International Telecommunications Union, Radio (ITU-R)), 1 Erlang is taken as a two-way phone connection of 1 hour's duration. However, the strict definition of the Erlang is the average number of calls per unit time multiplied by the average length of a call measured in the same units, so that the Erlang is a dimensionless quantity. Expressed differently, the

$(400 \times \frac{1}{2} \times 0.1) = 20$ Erlang km^{-1} of street length, or 300 Erlang km^{-2} for a typical urban street layout.

Assuming a comparable amount of traffic for the personal (pedestrian) mobiles gives a coarse estimate of 600 Erlang km^{-2} in dense city areas. Peak estimates for personal terminals follow from assuming 3000 pedestrians per kilometre of street length, 80 per cent of whom carry portables using 0.04 Erlang per portable terminal, which results in 1440 Erlang km^{-2}. The peak-to-mean ratio of traffic is thus in the order of 3. In multi-storey office buildings, the traffic may increase by a factor of 10. For example, one portable per 10 m^2 using 0.2 Erlang gives traffic per floor of 20 000 Erlang km^{-2}. From these assumptions and estimates, indoor densities are evidently the major challenge to cellular techniques.

For non-voice data, some protocols allow queuing, which can be used to maximise the spectral efficiency. The traffic requirements, which were until relatively recently (before internet-type access for mobile terminals) thought to be small, are now expected to dominate voice requirements. For example, in a low rate service such as mobile fax, an assumption of 3000 terminals km^{-2} of which 15 per cent are using a mean call time of 6 min h^{-1}, gives a traffic estimate of 45 Erlang km^{-2}. Similarly, for interactive data services, assumptions of a cumulative channel occupancy of 15 s h^{-1} for a portable (10 pages h^{-1} at a rate of 8 kbits for each page, at 4.8 kbits s^{-1}) and 30 s h^{-1} for a mobile (i.e. 4.5 mErlang and 9 mErlang, respectively), with 2000 portables and 3000 mobiles per square kilometre, gives a total traffic estimate of $9 + 27 = 37$ Erlang km^{-2}. Other data services take the total estimate for non-voice services up to 150 Erlangs km^{-2}, i.e. just 10 per cent of the voice traffic. For indoor non-voice services, the estimate is 5000 Erlang km^{-2}. (The sum of indoor voice and non-voice traffic estimates is thus 25 000 Erlang km^{-2}.) In the future, video and multimedia services are expected to dominate, with internet access being the central issue. Internet-type traffic is typically very asymmetric, with much more data on the downlink (from the base to the mobile) than the uplink, an issue which has an impact on the system design.

1.3.2.1 Spectral implications

Using 8 kbits s^{-1} for mobile speech and 16 kbits s^{-1} for personal terminals (these need to have lower d.c. power consumption and be inexpensive), the spectral requirements follow from the traffic estimates. Making assumptions about the system layout and operation, traffic characteristics and channel efficiencies (CCIR, 1992b, p. 18), the result is 167 MHz for the mobile and 60 MHz for the personal stations. This gives a current estimate for the spectral requirement of 230 MHz (see Section 1.5). This estimate is considered conservative since it does not account for factors such as the expanding requirement for data networking for road traffic management, system spectral inefficiencies arising from multiple operator involvement, and further quality-of-service issues.

Erlang is the traffic intensity measured in calls per mean length of call. It is named after Agner Krarup Erlang (1878–1927).

1.3.2.2 Power flux density

The traffic and spectral estimates allow, in turn, an estimate of the power flux density (power per area, or power per area per hertz) resulting from the service. This allows a coarse estimate of the potential interference from such a scenario. In this estimate, the effect of distances to the base and mobiles are not included, so the figures correspond to an artificially worst case.

Table 1.1 is an example of a power flux density estimate and is based on CCIR, 1992b, p. 20. The traffic density is estimated as above. For the vehicular traffic, base and mobile together give 11 W Erlang^{-1}, or 6402 W km^{-2}. For the personal terminals, the power density is 108 W km^{-2}. After converting to a power flux density using the bandwidth allocations (see Table 1.1), the potential received interference levels can be calculated. For example, a receiver with a short monopole antenna (with an effective cross-section of $3\lambda^2/16\pi$) and a 30 kHz radio bandwidth, operating at about 1 GHz, will record a level of about

$$38(10)^{-12}[\mathrm{W\,m^{-2}\,Hz^{-1}}] \times 3(0.3)^2/(16\pi[\mathrm{m^2}] \times 3(10)^4[\mathrm{Hz}] \approx -47\,\mathrm{dBm}$$

and (similarly) about -58 dBm for the vehicular and personal cases respectively. These values are over 60 dB above typical receiver sensitivities (~ -130 dBm) and illustrate the interference-limited nature of such potential high density multi-user systems. It is evident that interference mitigation techniques will need to continue to develop for realising such traffic rates.

The high levels of power flux density also have implications for sharing with other services and activities. A dramatic example is the potential impact on the radio astronomy service. One estimate for the power flux density in rural areas is -150 dB (W m^{-2} Hz^{-1}) (CCIR, 1992b, p. 23) which is many orders of magnitude above the

Table 1.1 Potential power flux densities for a system in an urban area. R1 and R2 refer to radio channel designations (after CCIR (1992), p. 20)

Stations	Base and mobile (R1)	Personal (R2)
e.i.r.p (power per 'simplex Erlang')	10 W (from base) 1 W (from mobile)	3 mW (indoor) 20 mW (outdoor)
Traffic density (worst case)	600 Erlang km^{-2}	25 000 Erlang km^{-2} (indoor)[a] 1440 Erlang km^{-2} (outdoor)
Assumed bandwidth	167 MHz	60 MHz
Estimated pfd (power/Erlang) ×(Erlang km^{-2})/bandwidth	38 μW km^{-2} Hz^{-1} -104 dB(W m^{-2} Hz^{-1})	2.5 μW km^{-2} Hz^{-1} -116 dB(W m^{-2} Hz^{-1}) (indoor only)

[a] This takes into account vertical frequency reuse in buildings

harmful interference levels for radio astronomy (-250 dB (W m^{-2} Hz^{-1}) for the 500–3000 MHz band).

1.3.3 Radio data networks

The RLANs and *ad hoc* networks mentioned above go further than person-to-person communications; they allow users to swap data. Meetings between people are increasingly including the transfer of data, typically using computer-based notebooks. The meetings can be at informal places – airports, etc. – as well as at formal venues. The convenience and security of swapping data without relying on an infrastructure has prompted the *ad hoc* network concept as an extension to traditional personal communications. The terminal is basically a battery-powered computer with radio communications. Radio networks have taken on a particularly large set of acronyms as the application ideas evolve. Some generic examples are wide area- or wideband- or wireless-local area networks (WANs, WLANs, etc.), personal- (PANs) and metropolitan-area networks (MANs), and so on.

The landmark starting point for the RLANs seems to have been in 1991, when the European Telecommunications Standards Institute (ETSI) formed an official group to consider the concept, and a draft standard (High Performance Radio LAN – HIPERLAN) became available in 1995. The basis has developed from a 'coffee table' network to a system which replaces wired networks such as Ethernet. This proposed an ambitious 10–20 Mbps capability over a range of 0.5–50 m. In Europe, CEPT (European Conference of Postal and Telegraph Administrations) allocated spectrum at 5.2 GHz and 17.2 GHz to a 23.5 Mbits s^{-1} HIPERLAN system (COST, 1999). The operating environment is typically indoors and the proposed systems are complicated by their need to be capable of uncoordinated spectral sharing. The 902–925 MHz ISM band which American RLAN-type products occupy (under the FCC Part-15 ruling) is not available in Europe. The next ISM bands are at 2.4 GHz, and 5.8 GHz. For example, the LAN IEEE 802-11b standard has a total of 300 MHz of bandwidth, at 5.1–5.3 GHz and 5.7–5.8 GHz, and uses orthogonal frequency division modulation (OFDM) to spread the signal over a wide spectrum for combating the multipath. Bluetooth systems are intended for much less ambitious rates of the order 1 Mbps, with small ranges of the order 30 m.

The spectral requirements are thus a major hurdle to implementing globally portable 'personal' radio data networks. The channel impairments, however, are the greatest hurdle for the technology to overcome. Not only do the terminals need to be low power, but issues in the multi-user protocol become important. For example, a loss of coverage to a terminal creates network connectivity violations, which have the potential of upsetting the multi-user access protocol (Barton and Wilkinson, 1994). In such ambitious systems, the choice of modulation, multi-access and other signal processing requirements are governed by the propagation problems and spectral issues.

1.3.4 Biological effects

With the fast growth of radio personal communications, the human body is being subjected to increasing amounts of microwave radiation. The perception amongst users

of the potential health risk has become a high-profile topic. Personal cellphones with antennas radiating adjacent to heads and residents living close to base stations are obvious targets for public concern. While no definite evidence can be produced that the extra radiation at the expected levels is harmful, it is similarly difficult to establish that it is truly harmless in the long term. It seems that this topic will become increasingly important and that pressure from public perception will have a more significant influence on systems and, in particular, the antenna design for handheld terminals.

1.4 Modern multi-user system requirements

In small, private networks with relatively low-traffic requirements (such as fire service systems), the priority requirement is usually coverage. This is ensuring that the system works as often as possible when the mobile terminals are anywhere in the region of operation. For large, high-traffic systems, such as public mobile telephone services, the system requirements revolve increasingly about system spectral efficiency.

1.4.1 System spectral efficiency

A basic figure of merit for a multi-user system is the measure of *how many users can be served per unit bandwidth per unit volume*. Volume is used since high-rise zones in cities concern volumes of users, rather than areas of users. Some studies use *area per floor level* for high-rise environments. Clearly for suburban type environments, area is an appropriate measure instead of volume. For speech, the measure is in Erlang $\mathrm{km}^{-2}\,\mathrm{Hz}^{-1}$. Consequently, the system spectral efficiency (cf. Lee, 1982, p. 259; after Henry, 1979; Macario, 1991, p. 51) is

$$\eta_{\mathrm{S}} = \frac{N_{\mathrm{u}} R_{\mathrm{b}}}{N_{\mathrm{c}} B} \tag{1.4.1}$$

where N_{u} is the number of users that can be served simultaneously by the system, B is the one-way channel bandwidth (including guardbands, etc.) in Hz, N_{c} is the total number of channels used by the system (including control channels, etc.), and R_{b} is the information throughput in bits s^{-1}. While the one-way single-channel efficiency R_{b}/B is obviously important, in large systems the principle gain in system efficiency is made by re-using the bandwidth at the same radio frequencies. This re-use involves geographically separating channels using the same frequency so that the number of users can be increased without increasing the total (two-way) bandwidth $2N_{\mathrm{c}}B$ allocated to the system.

1.4.2 Cells for frequency re-use

The multi-user mobile communications systems of today are characterised by frequency re-use schemes. A ‘cell’ comprises a coverage area where a set of frequencies is used for the control and message channels. Adjacent cells generally use a different set of frequencies so that there is no co-channel interference between neighbouring

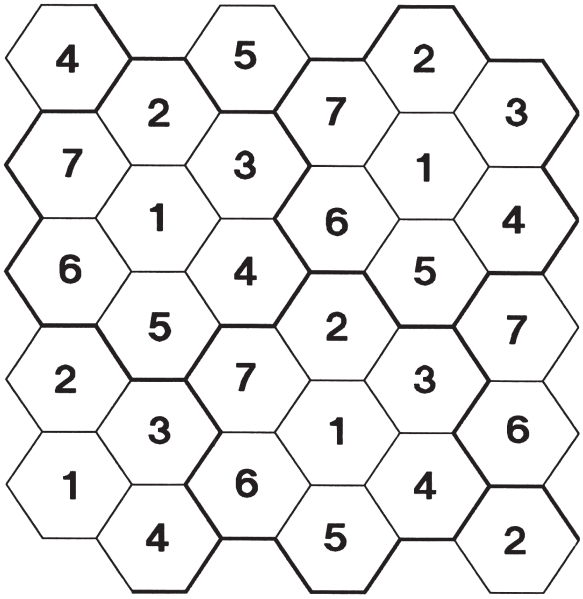


Figure 1.4.1 A functional cellular regular layout with a repeat pattern number of seven

cells. Functional schematics of cells are often laid out in regular hexagonal patterns although, in reality, the cells are irregular shapes and of different area, according to the geography and user density of a region, and the available base-station antenna sites. Figures 1.4.1 and 1.4.2 show some cellular layouts. Figure 1.4.1 is the regular layout with a repeat pattern of seven, such that there can be seven frequency allocations. This repeat pattern defines a distance between two cells that use the same set of frequencies. For such regular hexagonal cells, a fully tessellated pattern can use 3, 4, 7, 9, 12, . . . , frequency allocations. Most current systems are based on a hexagonal frequency re-use value of seven. For estimation of the system performance, only the six nearest co-channel sites are used for estimating the co-channel interference.

As the communications demand increases, the cells are split into smaller cells, so that no extra bandwidth is consumed while allowing more traffic (per volume or area) at the expense of installing new base stations. This is the single most spectacular aspect of the cellular concept: the system can continue increasing its numbers of users with fixed system bandwidth. Figure 1.4.2 (Macario, 1991) shows a typically functional cell pattern in which the cell size varies to cater for high-density traffic. A limit is reached when a cell is in the order of a single floor of a building, or a single room. This is a so-called pico-cell.

For a flat, homogeneous geography, almost regular cellular patterns can be realised. Typically, however, the patterns are most irregular, such as the example in Figure 1.4.3 from the hilly terrain of Wellington, New Zealand.

In fact, the actual cell boundaries are strictly defined as the places where hand-off occurs – the process of changing from one radio frequency set in the old cell to another set in the new cell, as the mobile moves between cells. This boundary can

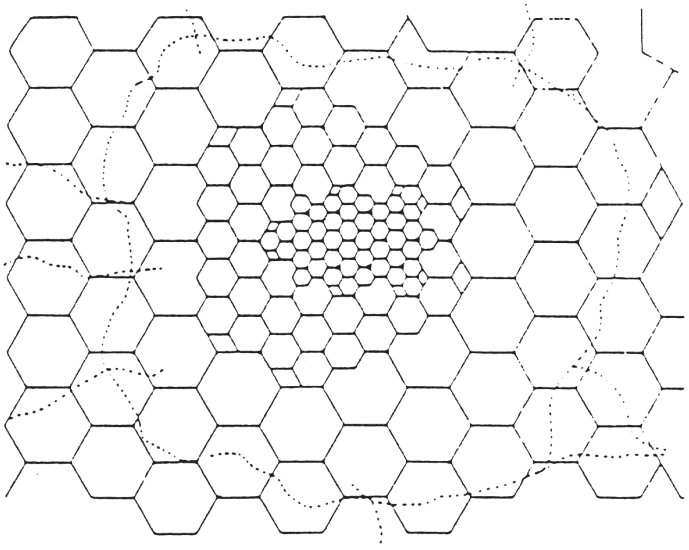


Figure 1.4.2 A functional schematic of a split-cell layout, where smaller cell sizes use lower power levels and shorter frequency re-use distances to gain high system spectral efficiency in busy areas (Macario, 1991)

change somewhat according to the mobile antenna pattern and its orientation (if it is directional). Also, longer-term effects such as changing vegetation cover can shift the hand-off boundary. Usually, there is some hysteresis effect in the sense that the cell boundary position depends on the direction of travel of the mobile between cells.

1.4.2.1 Co-channel interference

An underlying system problem in spectrally efficient, large multi-user systems is co-channel interference. As the signal power from a transmitting antenna decays with distance, often idealised with an inverse distance-to-a-power relation (e.g. inverse square law is exact for free space), there is normally some residual power reaching co-channel sites, as depicted in Figure 1.4.4. The noise floor is not explicit in the figure with the assumption that as the interference level approaches the noise floor, it becomes a noise contribution rather than a separate interferer. The obvious remedy for co-channel interference is to separate the co-channel cells more, i.e. have a greater repeat pattern number; but the need for system spectral efficiency pushes for a small re-use factor, which in turn reduces the cell separation.

In practice, the propagation is more complicated, and Figure 1.4.4 includes a typical real-world equivalent of the inverse distance relation. Here, there is a random component on the mean (over position – typically an area of a few metres square) of the signal strength and any deterministic decay law cannot be followed strictly. The mean co-channel interference, being the ratio of the two signals, can vary, typically by several tens of dB over distances of tens of metres. The most dramatic changes tend to be as a mobile moves into a region that is relatively shadowed from the strong

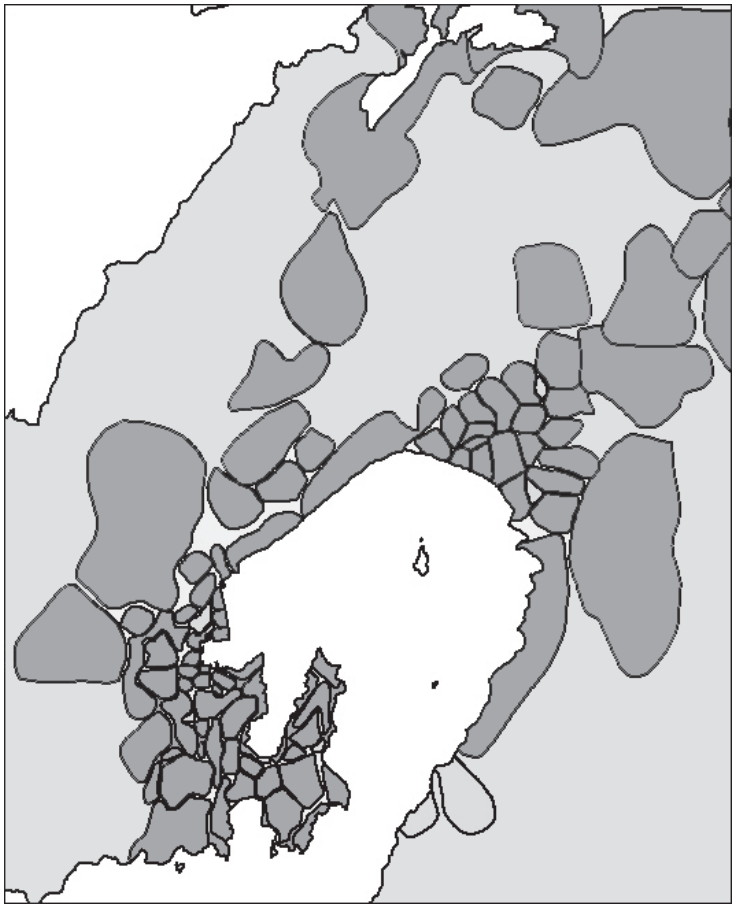


Figure 1.4.3 The irregular physical cell layout found in real-world systems. Local geography, the distribution of busy areas, and availability of base station sites constrain the layout. In this example, the white is water (Wellington harbour in the lower centre) and the business districts feature the smaller cells. The large cells are suburban areas and areas covering arterial roads (courtesy Telecom (NZ) Ltd)

flow of the electromagnetic energy; this often happens as a mobile moves past the corner of a building, for example.

If the short-term fading (discussed in Section 1.4.3) is included, the co-channel interference can be considered as an instantaneous quantity. Figure 1.4.5 depicts a short section of Rayleigh-like fading from two narrowband co-channels. The short-term fading can provide a variation by a further 40 dB, over a distance of a quarter wavelength (8 cm at 900 MHz), so that a short-term variation of 80 dB is possible between two fading signals.

The total variation over a few wavelengths can thus be up to 100 dB for narrowband channels, as illustrated in Figure 1.4.6. These statistical quantities give an idea of the huge 'safety margins' required in narrowband cellular system design. These margins are working against the drive for system spectral efficiency, and while direct help is available through antenna diversity, the multi-user access technique (see Section 1.4.4)

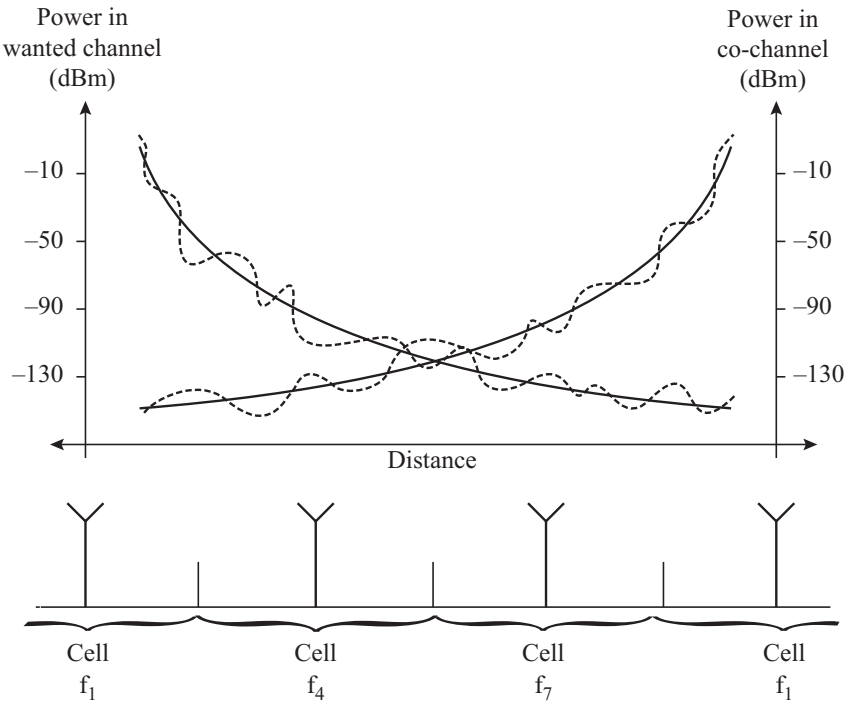


Figure 1.4.4 Schematic of mean co-channel interference between two cells using the same frequency. The mean co-channel interference varies monotonically between co-channel sites in an idealised model (solid lines) but, in practice, this may not be the case (dotted lines) and the wanted signal-to-co-channel signal ratio variation can be tens of dB in a range of several metres

and other diversity techniques (time, frequency) are sometimes applied to improve the efficiency.

Figures 1.4.5 and 1.4.6, also indicate a problem for the receivers – they must be capable of handling an extremely large dynamic range of signal strengths. Typical receivers have a dynamic range of 80 dB, which the Automatic Gain Control (AGC) stage can handle. Stronger signals can saturate the receiver (during which some receivers cease to function properly) and weaker signals will cause unworkable signal-to-noise ratios with a consequential outage in the link. By reciprocity, the base-station and mobile receivers require the same dynamic range although the absolute value of signal-to-noise ratio is not the same at each end of the link in general. In practice, the use of antenna diversity at the base station is more widespread than at the mobile, which relieves the dynamic range problem when the signal combining is pre-detection.

1.4.2.2 Near-far effect

The decrease in average signal strength as a function of distance creates another problem in multi-user mobile communications. A base station will receive a much

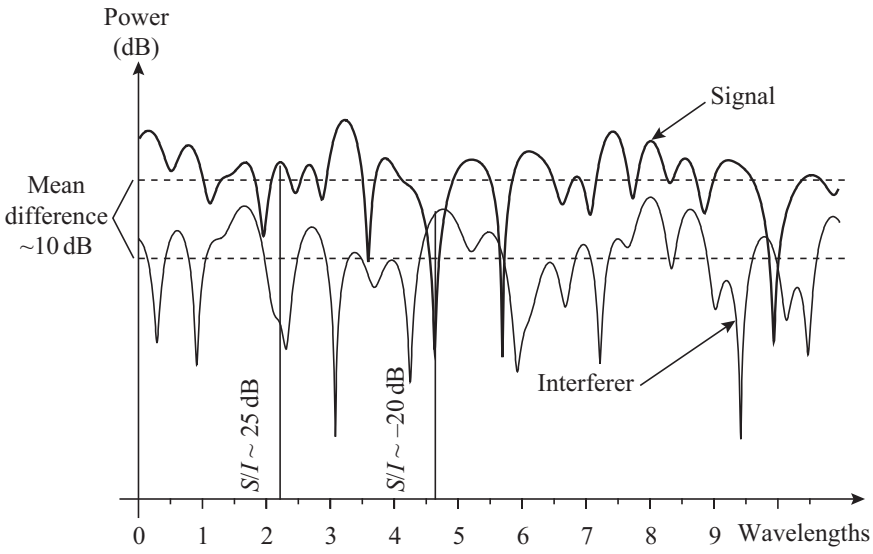


Figure 1.4.5 The short-term, Rayleigh-like fading component from two narrowband co-channels. The wanted signal-to-co-channel signal ratio can change by 80 dB within fractions of a wavelength

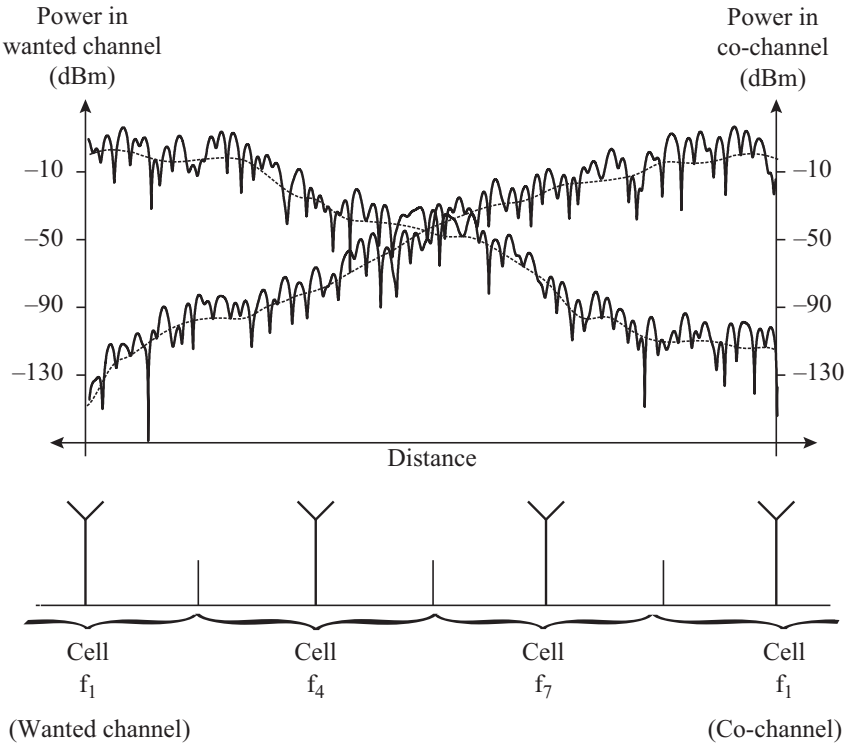


Figure 1.4.6 The Rayleigh-like envelope superimposed on local mean signal variation for co-channel narrowband signals. The wanted signal-to-co-channel signal power ratio can change by 100 dB over several wavelengths

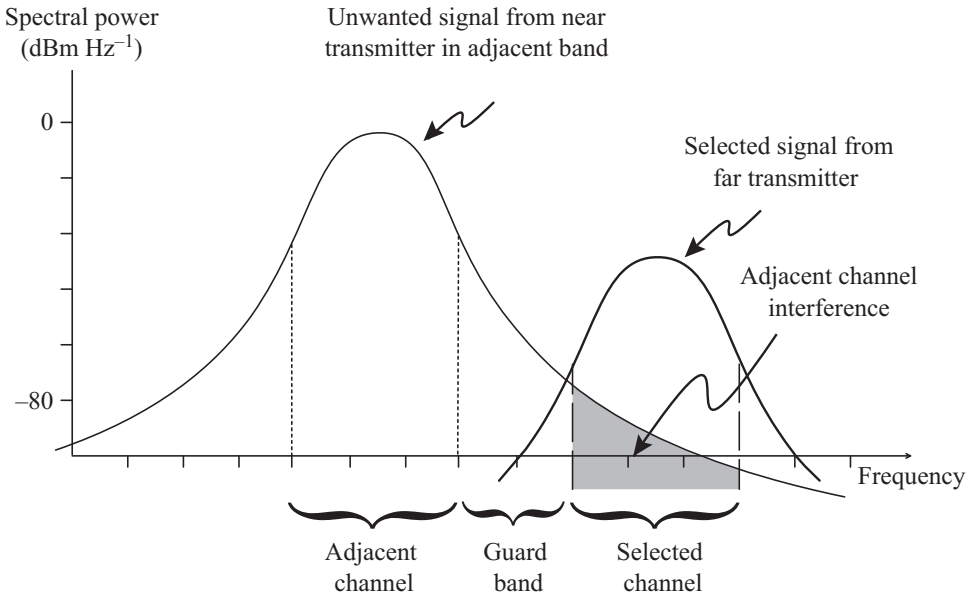


Figure 1.4.7 The near–far effect, in which the high power level from a nearby mobile can cause adjacent channel interference in the signal received from a distant mobile. The illustration here is for frequency-divided narrowband channels

stronger signal from a nearby mobile than from a distant mobile, for example one at the edge of a cell. This large mean power level difference becomes an interference problem if the spectrally unwanted transmissions from the near mobile become significant within the band of the transmission from the distant mobile. Usually, such interference occurs between adjacent bands (see Figure 1.4.7) since these are where most of the unwanted power is transmitted. It is too expensive to suppress completely such unwanted emissions, and the trade-off results in most systems being allowed adjacent channel transmissions, but at a level typically more than 80 dB below the power in the designated band. Nevertheless, the large dynamic range of mean signal strength in mobile communications still results in the near–far effect being a major constraint in system design. The above figures are idealised in their representation of spectral quantities. A portion of spectrum is occupied for a finite time, and the spectrum defined over finite time is different to the classical Fourier transform of infinitely long signals.

One technique to help suppress the near–far effect is to control, via the base station, the amount of power transmitted by the mobiles. A near mobile can be effectively instructed to transmit at a low power level while the distant mobile is instructed to transmit at a relatively high power level. Generally, such higher transmitted power levels at the cell boundaries act to increase the co-channel interference. Despite this, and the fact that power control increases the complexity of the mobiles and base station, the end result of power control is more efficient use of the spectrum in the communications system. In some schemes, the power control must be quite accurate.

For code division multiple access, for example, the power control should be good to about 1 dB, in order to avoid capacity degradation. In this case, however, it is co-channel, rather than adjacent channel interference that results in the need for power control.

1.4.2.3 Other factors

Many other inter-dependent service requirement factors also come into the system design stage such as: the tolerable outage rate (the relative amount of time in which the signal radio link is effectively blocked due to a lack of received signal strength); the tolerable signal-to-noise ratio or bit-error ratio; the maximum allowed transmitted power; the allowable co-channel interference levels; the type of receiver; and the height and gain of the antennas, etc.

In hexagonal layouts, the base-station antennas can be schematically considered to be omnidirectional and sited centrally in the cell or be directional with sectorial illumination. Again, such schematic niceties are compromised by real-world factors, and the choice of cell base-station antenna type is often dependent on these factors, rather than on an apparent advantage stemming from the ideal schematic layout. With mixed land and water paths, the propagation and radio signal coverage becomes particularly difficult to predict accurately, and usually experimental design must take place *in situ*.

1.4.3 Power variation in wideband channels

The above discussion used the narrowband channel model, where the short-term, Rayleigh-like fading is the same for all the frequencies in the band. This is called frequency-flat fading. For wider bandwidths, the fading becomes different for different frequencies across the band. The total received power, indicated with the RSSI (Received Signal Strength Indicator) label, is the averaged power over the bandwidth, so there is less (average) received power fading as the channel bandwidth is increased; see Figure 1.4.8. The averaged power over a band and over a distance are discussed further in Chapter 4.

Another view of frequency-selective channels is that the received signal has multiple time delays from the multipath propagation. The multiple delays change with time owing to the moving receiver and/or scatterers in the multiple paths. Everyday examples of multiple time delays in communications are echoes in telephone circuits, acoustic reverberation in rooms and ghosting in television images. The multiple time delays, or frequency-selective fading, degrades the quality and capacity of any communications link. As noted above, some recovery of the situation is possible through anti-multipath measures such as time diversity (for example, rake systems, discussed below).

1.4.4 Multiple access techniques

There are three basic techniques that allow multiple users to access the common radio channels. These are: channels divided by frequency (FDMA), by time (TDMA), or by

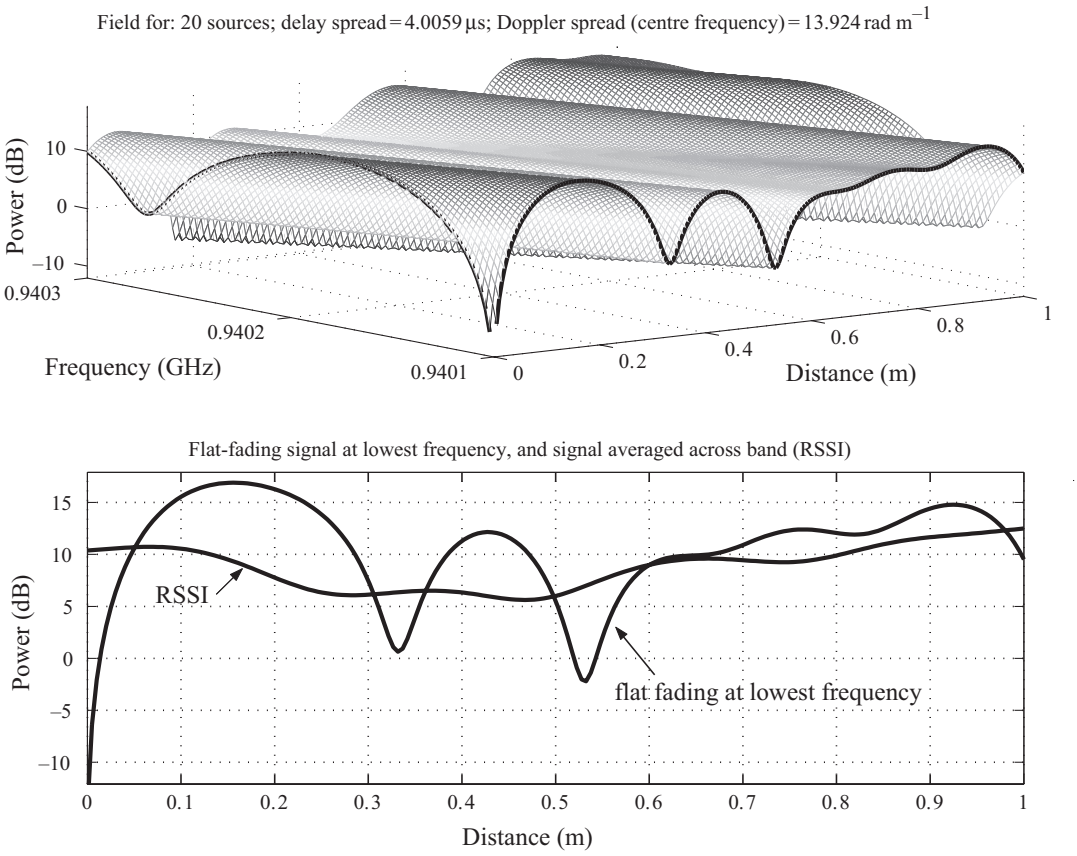


Figure 1.4.8 Wider band, or frequency-selective channels suffer less fading of the average (across the band) received power than narrowband, or flat-fading channels. In this example, the carrier frequency is from 940.1 MHz to 940.3 MHz and a large delay spread (see Chapter 2) of over 4 μs corresponds to an outdoor channel. (The delay spread is artificially large for the sake of demonstration.) The lower plot is a cut along the distance axis from the top plot and includes a graph of the power averaged over the bandwidth, a signal often called the received signal strength indicator (RSSI)

coding the channels which then overlap in both time and frequency (CDMA). These are depicted in Figure 1.4.9.

In FDMA, channels have a fixed frequency bin and that bin is occupied by the call for its duration, or until handover to another cell occurs. In TDMA, short packets of call information are time-compressed and occupy a slot of time within a packet of many slots. In CDMA, signalling codes for different channels are used which are chosen so that the interference between channels is minimal. This allows simultaneous use of time and frequency resource by many channels. A perceived advantage of CDMA is that the service degrades gracefully (a given channel gradually gets noisier) as the number of users increases. In FDMA and TDMA, the system capacity cuts off abruptly when all the frequency bins or time slots are used up.

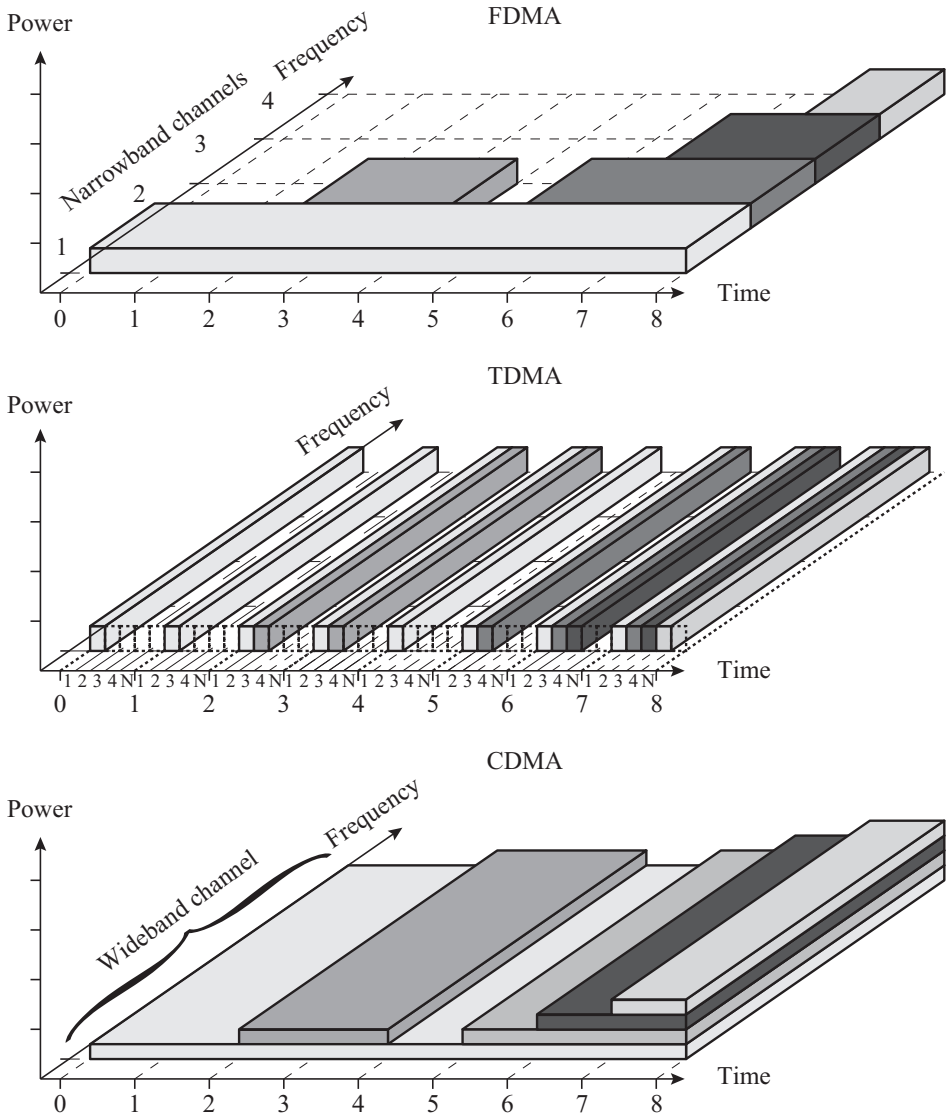


Figure 1.4.9 Multiple access techniques which share the spectrum amongst the users. In FDMA, the calls occupy their own (narrow) frequency band and are spaced with guard bands for minimal interference. In TDMA, the calls are time-sliced and compressed to occupy multiplexed time slots. In CDMA, the calls all occupy a common (wide) bandwidth but are coded so that there is minimal mutual interference

1.4.5 Duplex techniques for two-way channels

For two-way communications, a separate channel is required for information going in each direction. The possibilities are to use different frequency bins simultaneously or alternating time slots in a common frequency bin. These are called frequency division duplexing (FDD) and time division duplexing (TDD). Assigning directional

channels purely by coding is not currently considered a practical technique because of the duplexer implementation problems.

Frequency duplexing, which is the more traditional technique, requires filter sets and duplexers at each end to separate the channels, as well as separate local oscillator frequencies to undertake the (de)modulation. In time duplexing, no (frequency) duplexer is required, and only one filter set is required. However, extra signal processing is required in the form of the time-duplexing implementation. For comparison of these two techniques, consider the bandwidth of the single channel of the time-duplexing system to be the same as the total bandwidth as occupied by the two frequency-duplexed channels. The spectrum usage is thus similar. Some efficiency is lost in both systems because of the guard-time required between alternative directions in the time duplexing, and because of the guard-bands in the two frequency bins in the frequency duplexing. In time duplexing, there is an inherent time delay introduced.

For asymmetric data rates, the directional time assignments can be adjusted easily although end-to-end synchronisation is required. So time duplexing offers an advantage for asymmetric traffic rates and where some extra amount of delay can be tolerated. For systems using antenna diversity, the time duplexing offers the possibility of using the potential of the multiple antennas for transmit as well as receive. In the frequency duplexing, the multiple antennas are usually used only to their potential on receive. (It is assumed that the transmit and receive frequencies are well separated, which is usually the case.)

1.5 Usage and restrictions of the radio spectrum

The radio spectrum is a resource that requires cooperation and sharing between nations. In radio communications, frequency choice and power levels cannot be used indiscriminately and, in nearly all countries, users have laws to follow and a spectrum policing organisation to enforce the laws. It takes little thought to imagine the nightmare of interference, incompatibility and ensuing bad will, if the spectrum was not internationally regulated.

1.5.1 The rule-makers

The allowable usage of the spectrum is set by the International Telecommunication Union (ITU) and its regulatory arm, the International Radio Consultative Committee (CCIR), which recently reorganised and changed its acronym to the ITU-R. The ITU was founded in 1865 and is considered the oldest inter-governmental organisation. In 1947, it became a United Nations agency and in February 1992, the membership was 166 countries (ITU, 1992). The members have met every several years through the *World Administrative Radio Conference*, at which national interests are represented and compromises negotiated for international acceptance. When national interests are better served by deviating from the consensus reached, agreement of all the countries affected is required. The whole process is a daunting, but necessary, task to undertake,

and the evolutionary definition of spectral usage has had commercial interests and politics often winning out over logical, global methodology. The process is also an obvious user of research input, again in a diverse range of areas, including antennas, propagation, modulation, coding, equalisation, interference, networking, protocols, speech, space technology, grade-of-service, and technical specification issues.

The World Administrative Radio Conference concludes by members signing the consensus agreements which are then included, by way of revision, into the *Radio Regulations*, the bible of international radio spectral usage and related issues.

1.5.2 Issues in mobile communications

Mobile communications has become a predominant topic at the World Administrative Radio Conference. With the call for better compatibility between different mobile systems, and indeed the call for global personal communications, much standard setting and spectrum allocation is required. At the 1992 and 1999 Conferences, significant decisions were made on spectrum allocation for mobile communications, and this trend is expected to continue as demands for further mobile and personal services arise.

Any spectrum allocation covering many current and future personal and mobile communications services shows a hotchpotch of allocations. This is because of the historical development of various services and other spectrum users' activities, and also partly the result of technical considerations, such as propagation suitability for satellite communications.

1.5.3 Radio-link issues for research

With system configurations revolving around many issues, there is clearly a continuous need for research in mobile communications. Spectral usage is one issue, and this hinges on radio-link issues. The access and duplex techniques are important for mobile and personal communications and currently many architectures are being pursued with their performance being compared in various environments. This research will lead to guidelines for the planning and design of future multi-user systems. But such guidelines are still underpinned by the fundamental antenna and propagation issues. Understanding and characterising the various mobile channels – most of which are particularly hostile for efficient usage – and developing techniques for repairing and efficiently using the channels, are topics addressed in the following chapters.

1.6 Mobile channel terminology

The 'mobile channel' takes on several meanings in the scientific and technical literature. This book deals primarily with antennas and propagation as they relate to communications issues. There are several possible types of communications channels. Figure 1.6.1 illustrates the concept. Only half of the link is shown, since the other half can be viewed as a mirror image for this discussion. For the typical speech

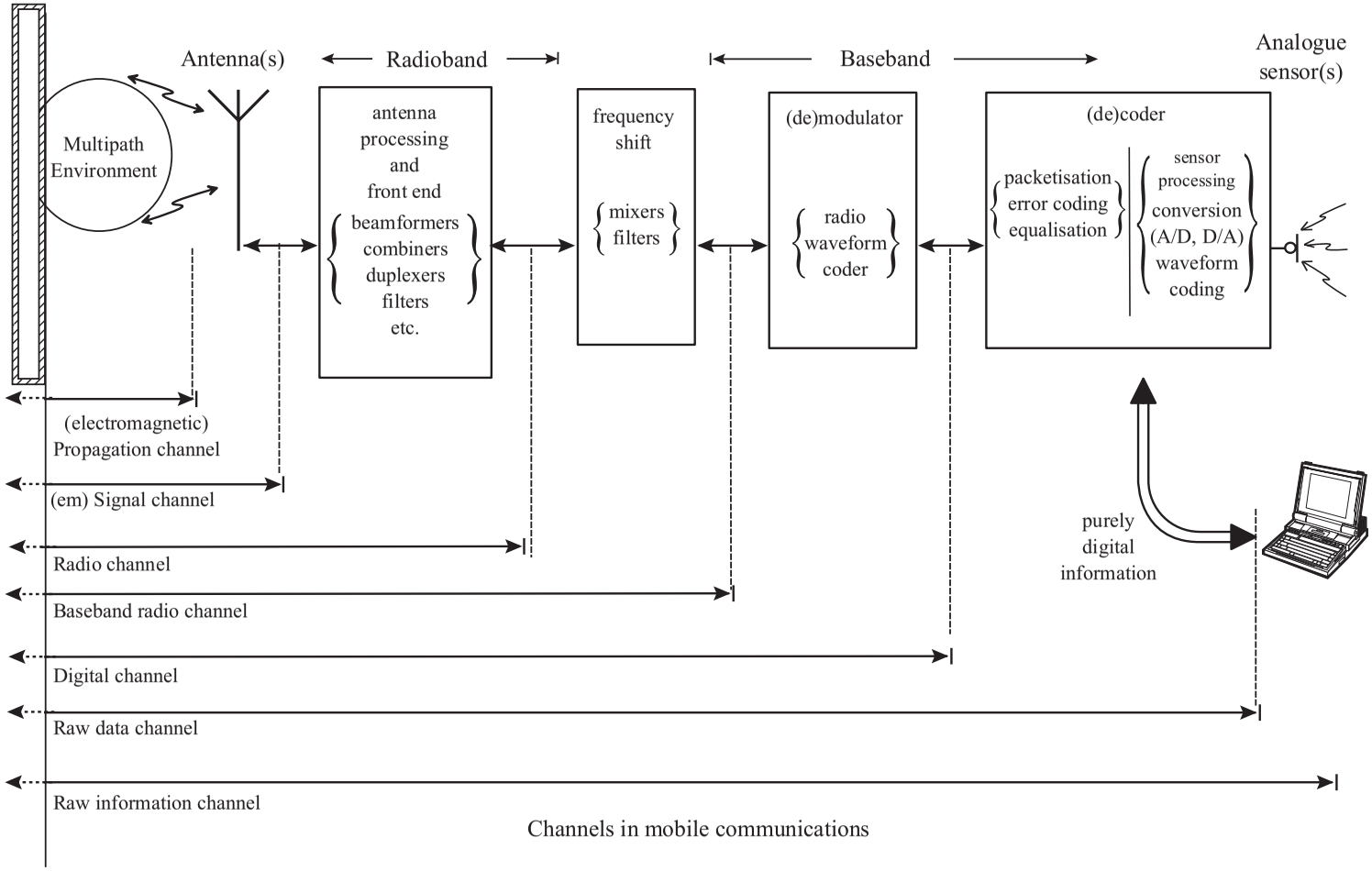


Figure 1.6.1 *A definition of channels in radio communications. The order and even presence of some of the block functions is somewhat arbitrary, with the antenna processing, for example, often implemented at baseband. The multi-user multiple access function is not explicit; it may be implemented in the modulator (e.g. CDMA), the frequency shift (cf. FDMA) or the front end (cf. TDMA)*

user, the immediate environment (including any effects of acoustic noise and room reverberation) forms part of the communications channel. So for speech (and also other analogue information sensors), there is an *information channel*. (Note that this is not information in the Shannon information-theoretic sense.) It includes the effects of multiple sensors, the environment and any intentional signal conditioning. The quality of the information channel is typically subjective, although analogue signal and network parameters can be applied for characterisation.

Once the analogue information is converted into electrical form by the sensors, the electrical signal is known as raw data so a *raw data channel* is defined. The quality of this channel can be quantified in terms of classical electrical network parameters. Once digitised, the signal undergoes data processing in preparation for transmission and reception. To the data user, or networker, the communications channel is purely digital with various rates (e.g. transmission rates, error rates, etc.), availability and latency (delay) being the quality criteria. This channel is called the *digital channel* (see, for example, Aulin, 1981) and is nonlinear.

Once the data are modulated (here modulation is considered only as the operation of coding the data into baseband analogue waveforms for radio transmission), the channel can be referred to as the *radio channel* (see, for example, Steele, 1992). The frequency shifting does not alter the radio channel in the sense of complex envelope signal description. The radio channel includes the effects of antenna processing, front ends, antennas and the electromagnetic propagation. In practice, antenna processing is often implemented digitally at baseband, but is nevertheless properly viewed as part of the radio channel of Figure 1.6.1 since it is an effective alteration of the antenna interaction with the free space waves.

In mobile communications, the antennas are often multiple port, with each port gathering different signals, and each signal arriving via multiple, free space waves. The definition of the channels here, in particular, requires clarification. The fields resulting from propagation, typically via a multipath environment, at a receiving antenna are considered as a transfer function, which is called the (electromagnetic) *propagation channel*. The fields can be converted to a voltage on an ideal, open-circuited antenna such that they are not perturbed. For example, the open-circuit voltage of an electrically small dipole is proportional to the unperturbed electric field in the polarisation defined by the orientation of the antenna.

The signal received by a loaded antenna element is referred to here as the output from the *electromagnetic signal channel*. This channel is often referred to in the communications signal processing literature as the propagation channel. This channel has a transfer function which comprises the propagation channel, the effects of the interaction of the antenna and the fields, and the antenna itself. The complex directive gain of an antenna, as well as its impedance and thence its matching to a receiver, are frequency dependent. The antenna thus contributes a filter in the propagation signal channel. Although the antenna's transfer function may be difficult to separate from the multipath propagation channel in a measured transfer function, it is possible of course to separate it in simulations and models. The intimate relationship between the antenna and multipath propagation is addressed in Chapter 2. Table 1.2 summarises the various channels.

Table 1.2 Summary of the channels of Figure 1.6.1

Channel	Input	Output	Comments
(em) Propagation	fields at transmitter(s) (e.g $\boldsymbol{E}_{\mathbf{T}}$ from antenna or $\boldsymbol{E}_{\mathbf{S}}$ from scatterers)	fields at receiver (e.g. $\boldsymbol{E}_{\mathbf{R}}$) or open circuit voltage of probe antenna	• linear • reciprocal • transmitted fields calculated • received fields calculated or probed
(em) Propagation signal	voltage source at transmitter	voltage across load in receiver	• linear • reciprocal • includes antenna loading effects at receiving end
Radio	(baseband) voltage from demodulator of transmitter	(baseband) voltage into demodulator of receiver	• linear if front end linear • reciprocal if front end and antenna processor are reciprocal • most systems nonlinear and nonreciprocal
Digital	digital data into the modulator	digital data from demodulator	• nonlinear • non reciprocal
Raw data	electronic signal from receiving electronic sensor	electronic signal at actuator	• ideally linear
Information	information source signals to sensor (e.g. sound waves)	(reproduced) information signals from actuator (e.g. sound waves)	

The many possible channel definitions here do not comprise a standard terminology. They are put forward here as a reminder to be aware that there can be many interpretations of the 'mobile channel' and that confusion will arise unless it is understood which form of the channel is under discussion.

1.6.1 Nonlinearities

It is normally assumed that the electromagnetic signal channel and the radio channel are linear. In practice, this is not always the case. In a communications link nonlinear

distortion can degrade the signals through amplitude-dependent phase changes, and vice versa. The information capacity is degraded by such distortion. Another serious ramification is that the signal spectrum becomes broadened, which affects other spectrum users.

For memoryless (resistor-type) nonlinearities, the transfer function can be found from a mathematical model of the nonlinearity frequency mapping $W_\mathrm{n} = K(\omega)$, where K can be multivalued. The transfer function is denoted $G_\mathrm{n}(K(\omega))$, so the impulse response of the nonlinearity, for simple $K(\omega)$, can be written

$$g_\mathrm{n}(t) = \frac{1}{2\pi} \int G_\mathrm{n}(K(\omega))\, \mathrm{e}^{\mathrm{j}K(\omega)t}\, \mathrm{d}\omega. \tag{1.6.1}$$

The main source of nonlinearities is active components, in particular the radio frequency amplifier. Oxidation on metal components can also cause nonlinearities in scattering fields. Even rusty fences near high-power broadcasting antennas have been known to create spectral violations! The choice of power amplification interacts in a fundamental way with the radio system design. The efficiency of the amplifier (in-band radio frequency output power over d.c. supply power) directly impacts on the battery duration (or size) for mobile terminals. For linear amplifiers, such as Class B, the efficiency is always less than 50 per cent. For tank circuit Class C type amplifiers, the efficiency can be over 80 per cent, but their nonlinear nature means that output stage filters are required for spectral compliance, and these can be expensive, and that constant envelope modulations must be used. Constraining the system to have a constant envelope modulation (such as GMSK, TFM, CPM, and other angle modulation schemes) has some current technology advantages, but also severely constrains the channel spectral efficiency. The tie-up with modulation brings coding issues into the picture as well, and these are reviewed very briefly below.

1.6.2 Aspects of capacity, channel efficiency, modulation and coding

The Shannon capacity theorem, in the form

$$C = B \log_2 \left(1 + \frac{S}{N}\right) \text{ bits s}^{-1}, \tag{1.6.2}$$

gives the fundamental limit of errorless capacity in bits s^{-1} of a channel of bandwidth B Hz and a signal-to- (additive, white, Gaussian) noise ratio of S/N. The formula is deceptively simple. Interpreting the equation in terms of bits s^{-1} is potentially confusing and is thus addressed further on. The formula shows that the noise is the only fundamental limiting factor to capacity. In practice, the mobile channel is limited by 'non-fundamental' factors arising from the multipath propagation, dispersion and, in multiple-user systems, interference. Shannon's formula represents a limiting capacity and would require limiting, or perfect signal coding and processing to achieve. To attain the Shannon capacity, for example, would require infinite delay in the coding. The theorem can be interpreted as follows: that C can be achieved without errors using 'sufficient coding'; and that attempting greater than C will always result in errors,

no matter what the coding is. In Chapter 9, the capacity of MIMO (multiple-input, multiple-output) systems, or transmit–receive diversity, uses the term 'capacity' for capacity efficiency, i.e. the 'capacity' is in bits $s^{-1}\,Hz^{-1}$.

The coding is required in order to force the transmitted signals to have a Gaussian nature (the symbols should be statistically independent, zero-mean Gaussian random variables), a condition which maximises the capacity for a given bandwidth. It falls into two categories: source coding and channel coding, discussed below. Other points regarding the capacity equation are that the noise is Gaussian, which represents a worst case, whereas in practice the noise may have a non-Gaussian character, and that successively transmitted bits or symbols are equally probable and mutually independent. In uncoded, digitised real-world information, successive symbols are often highly correlated and they comprise highly redundant information. However, this natural redundancy is seldom useful, in an information-theoretic sense, for protecting the information against transmission errors. *Source coding* is the generic term for coding the information signal so that it has minimal redundancy. For example, uncoded speech is typically 64 kbits s^{-1} but this can be reduced using current technology to 4.8 kbits s^{-1}, a compression ratio of 13.

Once the data have had as much redundancy removed – or concealed – as practicable, the Shannon limit can be approached more closely by using *channel coding*. This again falls into two aspects: forward error correction (FEC), and modulation. FEC is discussed first.

The FEC (re)introduces redundancy at the signal coding stage of the transmission and so the required data rate is increased. Typically, we use linear block coding schemes such as Bose–Chauhuri–Hocquenngham (BCH) and Reed–Solomon (Proakis, 1995), whose information rates (data rate after coding over the data rate before coding) are close to one. From Shannon's theorem, the resulting increased data rate will experience an increased error rate in the channel. However, a proportion of these errors can be corrected in the decoding stage at the receiver. The ensuing transmission error rate for the uncoded data is reduced. In the fading channel of mobile communications, *burst errors* (a relatively short sequence with a very high error ratio) occur, rather than random, uniformly distributed errors. The block codes are designed for random errors and so the uncoded data are often *interleaved*, i.e. reformatted so that concurrent symbols in the raw data become separated by many symbols in the transmitted data. At the receiver, burst errors in the interleaved, coded data become randomised after de-interleaving and the FEC decoding can now operate more optimally on the random error occurrences.

The other aspect of the channel coding is modulation. Traditionally, the modulation stage of a radio communications system has referred to the frequency shift from baseband to carrier frequency, but has also doubled for the alteration of the information signal to a form which was borne by the carrier for transmission (e.g. frequency modulation). In digital communications, the modulation stage also covers the arranging, at the transmitter, of the symbols onto constellations of points on the complex signal plane, and decoding the constellation points at the receiver. The concept of modulation returns the discussion to Shannon's theorem where, in a sense, it is misleading to interpret the equation as a transmission speed in bits s^{-1}. With

multilevel modulations (symbols comprising many bits), it is the symbol rate R_S, rather than the bit rate, R_b, that is relevant to the bandwidth because of the relation[3] $R_S \approx B$. The single-user *spectral efficiency* in bits s^{-1} Hz^{-1} is given by

$$\eta_C = \frac{R_b}{B} = \frac{N_S R_S}{B}, \tag{1.6.3}$$

where N_S is the number of bits per symbol. The higher the level of modulation (higher N_S) the higher the error rate for a given channel noise. However, higher-level modulations require a linearised transmission stage, including the amplifier.

The use of FEC can be independent of the modulation. This is when the data are (FEC) coded and then mapped onto the modulation symbols in an unrelated processing action. In *coded modulation*, the FEC and modulation processes are combined, and this more advanced technique offers better performance than using the FEC and modulation separately.

An example of using the channel efficiency as a modulation metric is as follows (Delogne and Baan, 1999). The basic multilevel modulation is quadrature amplitude modulation (N_S-QAM). The achievable channel efficiency (in a channel with no multipath or with the multipath effects mitigated) of the uncoded QAM is about

$$\eta_{QAM} = \log_2\left(1 + \frac{S}{N_O P_{QAM}}\right) \tag{1.6.4}$$

for an error rate of 10^{-7} where $P_{QAM} \approx 20$, i.e. 13 dB, which is the penalty against the Shannon limit. For examples of N_S-QAM, the channel efficiency and signal-to-noise ratio are given in Table 1.3. Incorporating basic FEC reduces the penalty to 9 dB and, using coded modulation, to about 6 dB. So a coding gain of about $13 - 6 = 7$ dB is possible using sophisticated coding techniques.

The more advanced, coded modulation techniques are not widely used for mobile communications, where practical constraints such as the extra signal processing complexity and battery size do not yet justify the ensuing improvement in spectral efficiency. Also, for single-user speech channels, long FEC codes for high error-correction performance may incur an unacceptable delay. Most signal processing in existing mobile systems is directed to the multiple access scheme (see below) and the mitigation of the multiple path effects. However, the choice of amplification/modulation issue can be put into the context of channel efficiency by using the above example. Many mobile systems use Class C amplifiers and constant envelope modulation in order to save battery power. The saving relative to a class A amplifier supporting multilevel modulation is about 2 dB (from $10 \log_{10}$ (80 per cent/50 per cent)). But this corresponds to a small relative reduction of efficiency, at least for reasonably high signal-to-noise ratios. However, the penalty in channel capacity is

[3] Nyquist's criteria (Nyquist, 1928) states that signals with a symbol rate R_S and no intersymbol interference require no spectrum outside of the range $(1+\alpha)R_S/2$ where $\alpha \geq 0$. For realisable transmission, $\alpha > 0$, and this condition is used also for controllable ISI, which can be removed at the receiver by equalisation (partial response signalling).

Table 1.3 Channel efficiencies of multilevel modulation (N_S-QAM) for an error rate of 10^{-7}

Channel efficiency, η_{QAM} (bits s^{-1} Hz^{-1})	Bits per symbol, N_S	Signal-to-noise ratio, dB
2	4	17
4	16	24
6	64	31
8	256	37

severe, being one half. This can be seen from the bandpass signal representation of a signal at centre frequency ω_C and bandwidth B,

$$\text{Real}\{r(t)\exp(\text{j}\omega_C t)\} = I(t)\cos\omega_C t - Q(t)\sin\omega_C t, \tag{1.6.5}$$

where the in-phase and quadrature components, $I(t)$ and $Q(t)$, respectively, are independent and bandlimited to $B/2$. Constraining the modulus of the envelope, $r(t)$, to be constant, means that $I(t)$ and $Q(t)$ become mutually dependent and can no longer convey independent data, and so the capacity is halved. So from an information-theoretic viewpoint, there is little justification for constant envelope modulations based on power savings of a Class C amplifier. However, the severe multipath of the channel means that much of the currently practicable signal processing power is devoted to mitigating the effects of dispersion and changing SNR. In the short-term future of mobile communications, the multiple access scheme, multipath effects and interference suppression appear to be set to continue dominating the signal processing. However, the ever increasing pressure for efficient spectral usage and available aperture for antennas may reach the stage where the fundamental Shannon limit on capacity becomes the biting constraint and the modulation and coding choices and processes will reflect this.

1.6.2.1 Energy per bit

Finally, it is instructive to view the capacity theorem in terms of the ratio of the energy per bit, ϵ_b, to the noise power density, N_0, in the channel. The average power per symbol is the constrained quantity

$$S = \frac{1}{T}\int_0^T E\{s^2(t)\}\,\text{d}t, \tag{1.6.6}$$

where $s(t)$ is the symbol and T its duration, and so the energy per bit in the channel can be written in terms of the capacity as

$$\epsilon_b = \frac{S}{C}. \tag{1.6.7}$$

Similarly, the noise in the channel can be written

$$N = BN_0 \tag{1.6.8}$$

and so the Shannon channel efficiency can be written

$$\frac{C}{B} = \log_2\left(1 + \frac{C}{B}\frac{\epsilon_b}{N_0}\right). \tag{1.6.9}$$

The formula shows explicitly that the spectral efficiency depends only on the received energy per bit, for fixed noise density. So with the demand for increased capacity (i.e. an increased bit rate), there is a need for an increased energy per bit, for a given modulation. The demand for decreased transmit powers, and the propagation loss, both act to reduce the received energy, but the antenna gain is a stage where the energy, and channel capacity, can be recovered.

1.6.3 A note on reciprocity

In this book, the principle of reciprocity is invoked tacitly. It is often more convenient or more illuminating to discuss specifically the transmit or receive case for the antennas and propagation issues. Unless explicitly noted, the transmit and receive cases will be interchangeable. This should not be confused with downlink (meaning base-station transmit and mobile reception) and uplink (mobile to base transmission) channels which can be different owing to the usage of different frequencies and subject to different noise and interference. Reciprocity is strictly related to an exchange of transmitter and receiver antennas, everything else (such as frequency) being kept the same.

1.7 Effect of multipath on the digital channel

1.7.1 Short-term multipath dominating the mobile channel behaviour

Multipath radiowave propagation is the dominant feature of the mobile channel. More often than not, the transmitted signal has no line-of-sight path to the receiver so that only indirect radiowave paths reach the receiving antenna. For microwave frequencies, the propagation mechanisms are a mix of: specular reflection from electrically smooth surfaces such as the ground; diffraction from edges of buildings, hills, etc.; scattering from electrically thin or small objects such as posts, cables, furniture, etc.; and also diffuse scattering from electrically rough surfaces such as some walls and grounds.

Some amount of multipath occurs in nearly all communications links. The basic phenomenon is that several replicas of the signal are received, instead of one clean version. This can be seen in the form of television ghosts. On transmission lines, reflections from mismatches on the line give the same effect, for example as echoes on telephone lines. On a long-distance point-to-point radio link, a direct line-of-sight wave, a single ground bounce, and atmospherically refracted waves can all contribute to the received signal.

Even when signal replicas are not sufficiently spaced in time to be discriminated (and processed as discrete contributions) by the receiver, the received signal may still suffer distortion. In the digital channel, the distortion contributes to intersymbol interference, which is when the signal of one symbol leaks into the signal of other symbols. The effect of the distortion can be to limit the capacity of the channel. The phenomenon is analogous to severe acoustic distortion recognised as the 'railway station effect', where an increasing power output (volume) does not increase the intelligibility of the message. In digital communications, the distortion caused by multipath creates a similar effect: an increase in transmitted power does not decrease the BER (bit-error ratio) as Shannon's theorem would otherwise imply. The amount and nature of the multipath propagation sets the level of power at which the BER becomes essentially independent of the SNR. The effect is often referred to as the 'irreducible BER', but the use of signal processing can reduce the BER incurred by the multipath. Early experimental examples of the irreducible BER in the digital channel are given below.

1.7.2 Fast fading of the analogue channel

The fading of the channel is treated in detail in the following chapters. Here, the basic phenomenon and its effect are reviewed. It is the interference, or phase mixing, of the multipath contributions that causes time- and frequency-dependent fading in the gain of the channel. The time dependency is normally from the changing position of the mobile, and so is also referred to as space- or position-dependent.

It follows that, at a given frequency, the power of the received signal, i.e. the gain of the mobile channel, changes with the position of the mobile terminal. This changing signal-to-noise ratio (SNR) is *signal fading*, and its associated phase change is often experienced as audio 'swooshing' or 'picket fencing' when an FM station is received by the antenna on a moving car. The fades can be as close as about half a wavelength, so for an FM station at a typical FM broadcast frequency of 100 MHz, and a vehicle travelling at 100 km h^{-1}, the swooshing can be at a rate of about 20 Hz.

If the mobile receiver is stationary, the signal may continue to experience some time-dependent fading, and this is caused by changes in the multipath environment, which may include moving vehicles, etc. In nearly all situations the changing mobile position dominates the time-variation of the mobile channel. Usually, the multipath environment is taken, or at least modelled, as unchanging. This is called static multipath. In this case, a static mobile experiences an unchanging channel. If now the radio frequency is swept, then the gain of the transfer function experiences fading in a similar way to changing the mobile position. This is because the electrical path distances of the multipath components are frequency dependent. For a single frequency, or continuous wave (CW) signal, the time- and frequency-dependent fades can be measured to some 40 dB below the mean power level, and up to 10 dB above the mean. This indicates the large dynamic range required of the receiver just to handle the multipath interference. This fading, illustrated in Section 1.4, is variously referred to in the literature as *fast fading*, or *short-term fading*, or *Rayleigh fading* after the Rayleigh distribution of the signal magnitude.

The maximum density of fading is a fade about every half wavelength on average and this occurs typically in urban outdoor environments and non-line-of-sight indoor environments. The fast fading dominates how the mobile channel can be used. For example, amplitude modulation at microwave frequencies is not normally a feasible option because, for a fast moving mobile, the fading interferes directly with the modulation.

1.7.2.1 Narrowband and wideband mobile channels

In a typical mobile microwave signal link, the *relative bandwidth* is small. This means that the spectral extent of the signal is less than a few per cent of the nominal carrier frequency. The fading within the frequency response of the transfer function is referred to as *frequency-selective fading*. If the bandwidth is sufficiently small so that all the frequency components in the band fade together, then this is called a *flat fading* channel.

In the mobile channel context, a *narrowband* channel has flat fading and a *wideband* channel has frequency-selective fading. The exact boundary between these cases is not well defined, and the cases are treated in later chapters since they relate to the antennas and propagation issues. Note that a narrowband channel can become a wideband channel, or vice versa, with a change of propagation environment. The use of a CW signal, or pilot tone, for channel characterisation, represents a limiting case of the narrowband channel. Historically, the fading channel behaviour has been the principal observed characteristic of the mobile channel. The fast fading represents another manifestation of the reception of several replica signals.

1.7.3 The effect of fading on the digital channel: irreducible bit-error ratio

1.7.3.1 Timing errors from random frequency modulation

The digital channel in Figure 1.6.1 is, in principle, the simplest channel to characterise experimentally since it concerns a BER measure. The fading in the mobile channel has a particular effect on the BER curves, including the 'irreducible BER', mentioned in Section 1.7.1. The example in Figure 1.7.1 (Miki and Hata, 1984) shows BER curves against carrier-to-signal ratio (CNR) – a parameter of the analogue radio channel – from simulations of the narrowband mobile channel with carrier frequency of 920 MHz. The static (no fading) curve shows the classical waterfall shape of the Gaussian channel. But the fading channel curves, shown with fading rates f_D, feature the irreducible BERs that occur at lower CNRs with increasing fading rate. The fading rate of 40 Hz corresponds to a mobile speed of about 40 km h^{-1} and a carrier frequency of 900 MHz, i.e. like using a cell phone in a moving car. The form of these curves holds essentially independently of the type of angle modulation used.

The mechanism for the bit errors is timing error caused by the random FM, imposed on the signal by the changing phase of the fading channel. It is addressed further in Chapter 5. In most digital modulation schemes, the digital channel quality is sensitive to even small errors in the symbol timing. Gagliardi (1988) specifically addresses the effect of timing errors on the BER. An example is given by Heald *et al*. (1998) in which the correlation receiver error mechanisms are demonstrated for

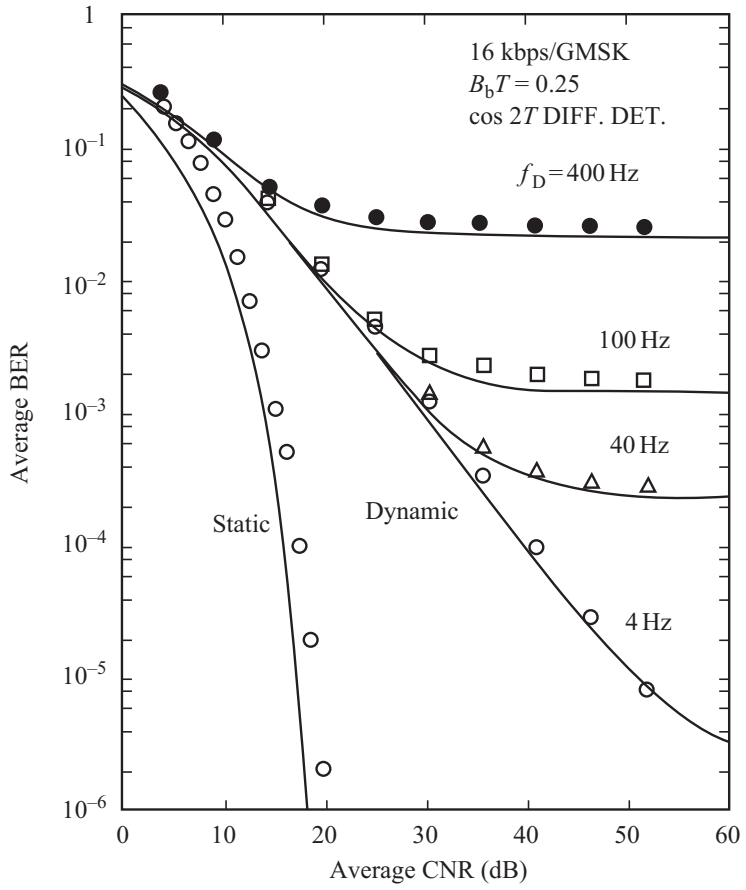


Figure 1.7.1 The irreducible BER effect for the digital mobile channel is when an increase of SNR does not improve the BER. The static (no fading) channel shows the classical waterfall shape of the Gaussian noise-limited channel but as the fading rate increases, the form of the curve alters (Miki and Hata, 1984)

incoherent MSK implemented with minimal word size. For example, a mean error of 4 per cent of the symbol time corresponds to an equivalent degradation in SNR of about 2 dB for an (undegraded) SNR of 14 dB. This corresponds to an increase in the BER of about two orders of magnitude.

1.7.3.2 Intersymbol interference from multiple time delays of the propagation channel

As the signalling rate increases and more channel bandwidth is required, the analogous irreducible BER effect occurs as a result of the several signal replicas arriving at different times. This spread of delays causes intersymbol interference when one dispersed symbol overlaps with other, similarly dispersed, symbols. In analogue parlance, this is called dispersive distortion. In the mobile channel the situation is complicated by the dispersion changing with time. The effect is depicted in the experimental example

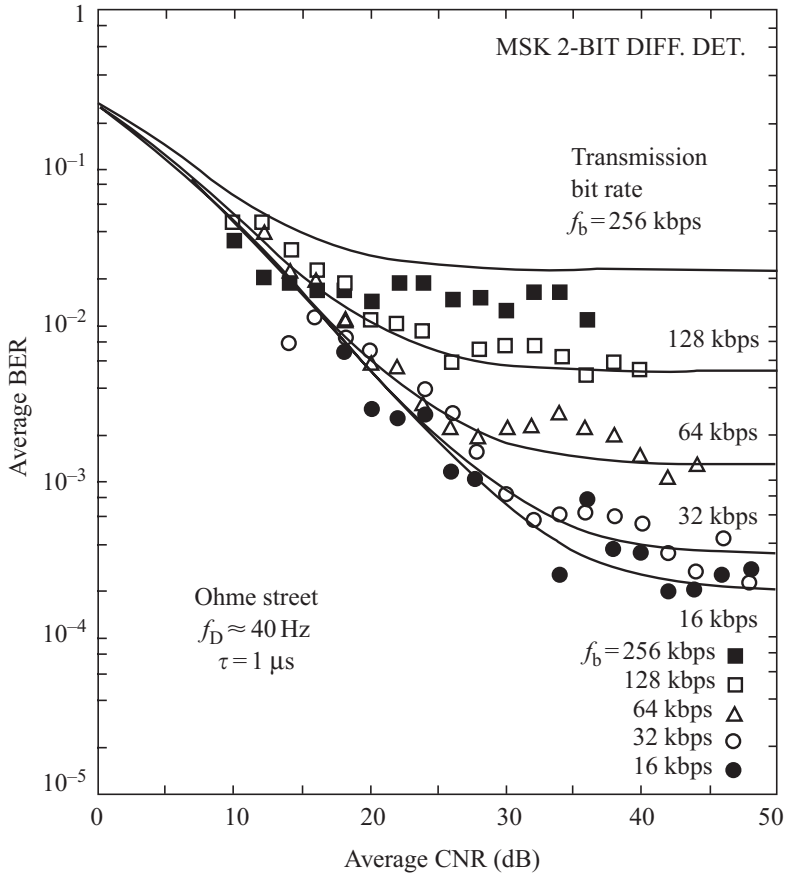


Figure 1.7.2 The irreducible BER effect caused by intersymbol interference. As the fading rate increases for a fixed spread of multipath propagation delay times, the irreducible BER increases (Sakoh et al., 1985)

of Figure 1.7.2 (Sakoh *et al*., 1985), where for a fixed fading rate of $f_D = 40$ Hz, the increasing digital transmission rate experiences an increasing irreducible BER. As in Figure 1.7.1, the effect is that the capacity of a given link cannot be increased by simply increasing the CNR, e.g. by increasing the transmitted power. Again, extra signal processing can improve the situation.

1.8 Signal processing for mitigation of the multipath effects

Several signal processing techniques can be applied to the mobile channel to reduce distortion and recover the capacity relative to the static channel. In flat-fading channels, the deep fades mean that the channel path loss is proportionally increased, which means a loss in the signal-to-noise ratio. The capacity of the channel decreases accordingly at these positions (or times). In frequency-selective channels, the SNR is averaged across the bandwidth of the channel, as shown in Figure 1.4.8 and so the variation in capacity with position is less than that of the flat fading channel.

In the fast-fading channel, *antenna diversity* uses multiple antenna elements to receive the same signal but with different multipath degradations, and combines the signals so that the resultant channel has better capacity than any of the channels from the individual antenna elements. The improvement in the SNR of the channel required for a given statistical performance measure is called the diversity gain. In the analogue channels, such as the radio channel, the measure is the percentage of time that the SNR is below a certain level. In the digital channel, the measure is similar for the BER statistic. Such measures are called *outages*. In the digital channel, for example, the outage is defined mathematically from the cumulative probability, written as $\text{Prob}(BER > BER_0) = 1 - \text{Prob}(BER \leq BER_0)$, or similar.

In the frequency-selective channel where the signal arrival times are resolvable, e*qualisation* and *rake* systems basically attempt to gather the delayed signal replicas and recombine them into a single signal which, ideally, becomes no longer distorted or faded. A combination of the equalisation, or rake, and antenna diversity, is often called *space-time processing*. All these techniques can be effective in improving the mobile channel. In fact, the use of antenna diversity offers very large potential capacities by effectively re-using the frequency in the same space. These diversity techniques are introduced below, with a short preview of the multipath channel characterisation. More recently, space-time coding has been used to describe narrowband MIMO coding.

1.8.1 Diversity basics for fading channels

1.8.1.1 Delay time and 'clock' time

In the study of propagation and multipath processing, it is important to understand the difference between the delay time and clock time. The delay time, denoted using τ, is a signal propagation parameter, and the time, denoted using t, refers to normal, or clock, time. These are treated as independent variables; however, they are sometimes equated during mathematical analysis.

1.8.1.2 Impulse response of narrowband and wideband channels

A narrowband channel is described as one in which the fading is the same across the bandwidth, and this is normally taken as a constant across the band. Strictly speaking, its phase is varying linearly across the bandwidth. In the usual electrical engineering notation, and using τ_0 for the time-of-flight delay, the transfer function and impulse response, for an ideal, static mobile channel, are expressed simply as

$$a\mathrm{e}^{-\mathrm{j}\omega\tau_0} = H(\omega) \quad \Leftrightarrow \quad h(\tau) = a\delta(\tau - \tau_0), \tag{1.8.1}$$

where the $\Leftrightarrow$ indicates the Fourier transform and ω is the radio-channel angular frequency. This model is developed more formally in Chapter 4. Here the channel gain is $|a|^2$, or equivalently, the path loss $1/|a|^2$, and the complex a term also contains the constant reference phase.

The delta function notation indicates, mathematically, an infinite bandwidth. For a finite bandwidth imposed on the flat fading, the impulse response at a given (clock)

time is replaced by a sinc-shaped function (the Fourier transform of the frequency window). As the mobile changes position, the value of the time-of-flight delay, τ_0, changes, and the channel gain also changes, i.e. the channel experiences flat fading. This is expressed mathematically by writing the parameters with position dependency, denoted by z, as in $a(z)$ and $\tau(z)$. The time dependency for static scatterers is then written in terms of the mobile velocity, V, as $a(Vt)$ and $\tau(Vt)$.

In a wideband channel, the alternative way to view the frequency-selective fading is to consider the dispersion, or spreading, of the channel impulse response. A transfer function whose magnitude is frequency selective has an impulse response that is spread out over a delay time range. (Recall that the flat-fading channel features a delta-like function for the impulse response.) The frequency-selective fading is position dependent and, equivalently, the dispersive impulse response, which is a function of delay time, is changing with clock time as the receiver position changes. The modelling equation for the impulse response is frequently written as a sum of delta functions, representing individual, or collections of, propagation paths, with each function having its own delay and complex amplitude, namely

$$\sum_i a_i \mathrm{e}^{-\mathrm{j}\omega\tau_i} = H(\omega) \quad \Leftrightarrow \quad h(\tau) = \sum_i a_i \delta(\tau - \tau_i). \tag{1.8.2}$$

Again, there is an implication of infinite bandwidth from the use of delta functions. The finite bandwidth complicates the situation with signals at different delays being forced to mix together. The channel bandwidth, normally defined by the communications system filters, thus has an impact on the dispersion seen by the signals. So the channel dispersion is governed by both the propagation channel and the communications system.

1.8.1.3 Broadband and wideband

The concept of wideband and narrowband in the propagation channel is independent of any signalling over the channel. However, in terms of communications signalling, 'wideband' is sometimes mistaken for 'broadband', which tends to be taken as meaning a large capacity catering for many multiplexed users. For example, a fibre optic or coax cable typically bears broadband channels, whose size is measured as a capacity in bits s^{-1}, and whose traffic normally comprises many different users.

Parameters of the propagation channel are, however, often normalised by communications parameters. For example, the dispersion of the propagation channel is often expressed in terms of the symbol rate of the communications.

In terms of antennas, 'wideband' and 'broadband' are often taken to mean a large relative bandwidth over which the antenna is matched or holds some other property such as pattern directivity. Typically, a wideband antenna is one which is matched over a relative frequency band of 10 per cent or more, with respect to the centre frequency. A narrowband antenna is matched over less than a couple of per cent of the relative bandwidth. Clearly a narrowband antenna can bear broadband communications channels if the centre frequency is sufficiently high to allow enough

capacity. Conversely, a wideband antenna may not support sufficient bandwidth to support a given broadband communications requirement.

1.8.2 Basic diversity

An established technique for combating the problems incurred by the fading is *diversity*. The idea is to employ two or more channels – each of which is a fading channel – and combine these such that a more reliable channel is realised. Each diversity channel is commonly referred to as a branch. The combiner can take several forms, from simply switching between the branches as required (i.e. switch when the current branch degrades to a threshold level of performance), to sophisticated optimum combining, which uses simultaneously all branches to minimise the outage with respect to some signal-to-interference-plus-noise ratio (SINR) or some average bit-error ratio, BER.

The diversity performance is measured by the diversity gain, which can be couched as a decibel value of equivalent signal-to-noise ratio. The diversity gain depends on the number of branches, the signal combination algorithm and the mutual statistics of the branch signals (i.e. the fading signal imposed on the signal-to-noise ratio of the branches).

For good diversity performance, the mean powers of the branches should be similar and the fading nature of the branches should be mutually uncorrelated. If the branch mean powers are too different then the more-powerful branch dominates and there is a corresponding decrease in the diversity gain. Similarly, as the branch mutual correlations increase, the diversity gain reduces relative to the uncorrelated case. There is little reduction in the diversity gain as long as the correlation is reasonably low, and 'uncorrelated branches' is often taken to be an envelope of correlation coefficient values of less than 0.7. The uncorrelated nature of the branches is normally defined by some spacing – a displacement of a correlation function – dependent on the type of diversity.

1.8.2.1 Antenna diversity for flat-fading channels

Antenna diversity refers to using different antenna patterns to receive information over different fading channels. Put simply, when the quality of one channel is low because of the multipath, the receiving system can rely (statistically) on the other channels to offer a high-quality channel. The concept is depicted in Figure 1.8.1. There are several degrees of freedom in realising antenna diversity. Uncorrelated fading between the polarisations can be used (polarisation diversity); similar patterns can be spaced apart so that they differ by a phase angle that depends on the direction of the incoming waves (space diversity); similar directive patterns can be rotated to receive from different directions (angle diversity); and finally two or all three of these techniques can be deployed together.

In space diversity, the spatial correlation function dictates the required antenna spacing distance. Similarly, the angular correlation function dictates the required angular spacing. However, there are only two orthogonal polarisations and so any polarisation spacing correlation function is periodic and there are only two uncorrelated branches available for pure polarisation diversity. The advantage of antenna

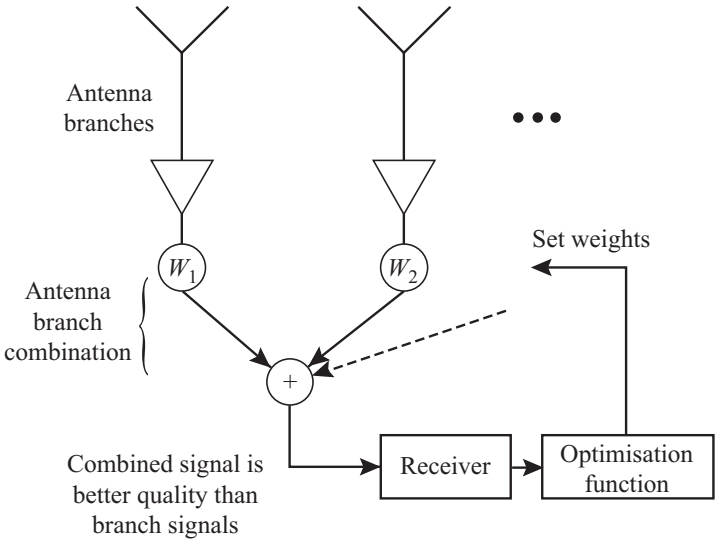


Figure 1.8.1 Antenna diversity combines the signals from different antennas whose different patterns receive mutually uncorrelated fading branches. The signal combination algorithms range from switching between branches as the signal quality degrades to setting complex weights to optimise the signal-to-interference-plus-noise ratio

diversity is that extra quality, or capacity, is being realised without the use of extra spectrum. The application of diversity is thus repairing the fading channel rather than accepting the fading channel and processing or coding to compensate for it. The disadvantage, of course, is the extra expense, inconvenience, and the physical presence of the required extra antennas.

As the capacity increases available from multiple antennas are becoming more widely understood, and as the design and accompanying processing techniques are becoming more feasible, the approach is gathering momentum in mobile and personal communications.

1.8.2.2 Adapted antenna patterns in optimum combining

The far-field pattern is a classical parameter of an antenna. In the more traditional usage of antenna signal combining for phased arrays, the action of the signal processing can be illustrated by the pattern adapting its shape to optimise the signal-to-interference ratio. Often such action is called 'null steering', because of the obvious creation or steering of nulls (or more correctly, minima) in the pattern towards interfering sources. In these cases, the wanted and interfering sources have discrete directions. The antenna is called a phased array because normally the weights are phase only. For an antenna with N_A elements, at least $N_A - 1$ different interferers can be suppressed.

It is convenient to view the capability in terms of *degrees of freedom* (DOFs). Each weight in the antenna equates to one DOF, and each DOF can be deployed by an optimum combiner to suppress an interferer, or contribute to the gain towards

the wanted signal. Normally, in a phased array, a constrained power minimisation algorithm effectively uses one DOF to maintain finite gain to the wanted signal and deploys the remaining DOFs to suppress the interferers. In optimum combining where there are $N_I < N_A$ interferers, then the interferers can be suppressed leaving $N_A - N_I$ DOFs to contribute to the gain of the antenna to the wanted signal. For example, a four-element array can at least suppress two interferers and provide twice the gain to the wanted signal relative to a single element. In many cases, more interferers can be suppressed depending on the spatial arrangement of interferers. For a phased array then, each DOF is purely real and allows a null to be steered in a (purely real) spatial direction.

In the mobile communications case, however, the wanted signal and the interferers each arrive from many directions and there are many more directions than there are elements in the antenna. In this situation, the action for suppressing interferers is one of setting complex weights (both amplitude and phase) to keep the interferers in a permanent fade and to keep the wanted signal out of a fade. Antenna diversity alone (with frequency-independent weights), is therefore not suitable for frequency-selective channels. The adapted pattern cannot be readily interpreted as in the angular case. The difference between the cases is illustrated in Figure 1.8.2. Instead of the

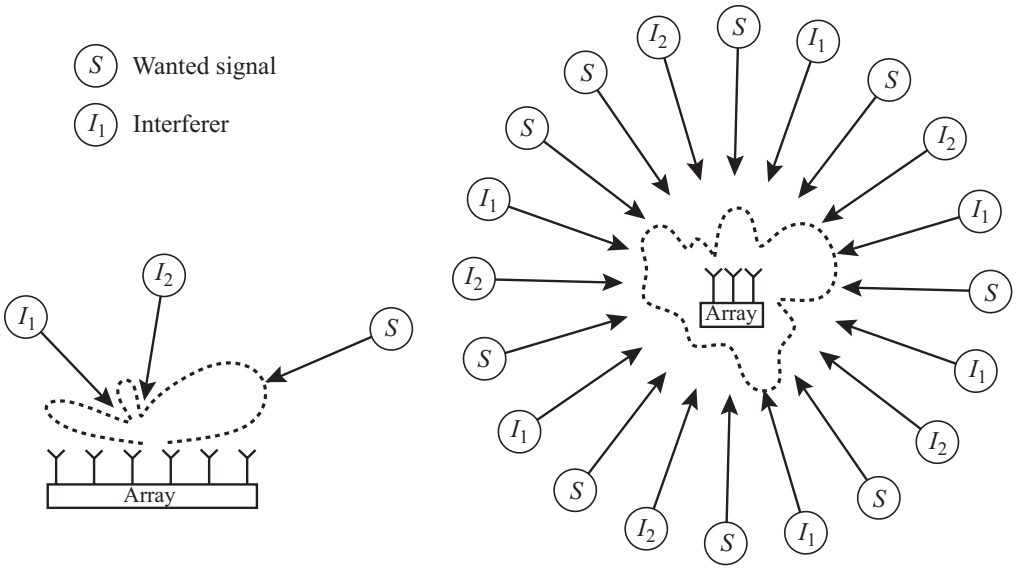


Figure 1.8.2 *The scenario for a traditional phased array (left), where a number of interferers in different directions can be suppressed and the action is one of minimising the received power from the interferers. The adapted pattern can be interpreted as 'null steering'. The mobile communications case (right) has many more interferer directions than elements and 'null steering' cannot work. Power is received from the interferers but the complex weights act to keep the total power from each interfering signal in a fade*

phased array pattern interpretation, complex weights are required for optimum combining and the adapted pattern is not normally meaningful. For N_A complex weights giving N_A complex DOFs, up to $N_I < N_A$ interfering signals (which can be distributed in many directions) can be suppressed and the diversity gain can be that of $N_A - N_I$ branches. Stated another way, optimum combining with N_A antennas and N_I interferers has similar performance to having $N_A - N_I$ antennas and no interferers.

1.8.2.3 Time diversity (rake receiver) for frequency-selective channels

Diversity techniques are not limited to the use of antennas. For the frequency-selective channel, replicas of the signal arrive at different delay times. Time diversity, or rake reception (Price and Green, 1958), treats the received signal at different time delays as diversity channels. Good diversity performance relies on the powers at each delay tap being similar and the fading at each delay tap being mutually uncorrelated. The concept is illustrated in Figure 1.8.3. The 'rake' refers to the function of gathering replicas of the signal from a tapped delay line, where the tines, or fingers as they are popularly known, represent the tapped delay line. In basic implementations, the tap delay times are fixed and the combiner works on the branch signals according to their instantaneous powers and/or phases.

A more sophisticated combiner may work to alter the delay times between the taps to suit the discrete arrival times of the signals; however, this would normally be implemented with small, fixed tap delays with some of the amplitude weights set to zero. Obviously rake reception is most effective when the delay times are clearly resolvable (cf. discrete echoes such as well defined ghosts on a television), i.e. not too smeared together.

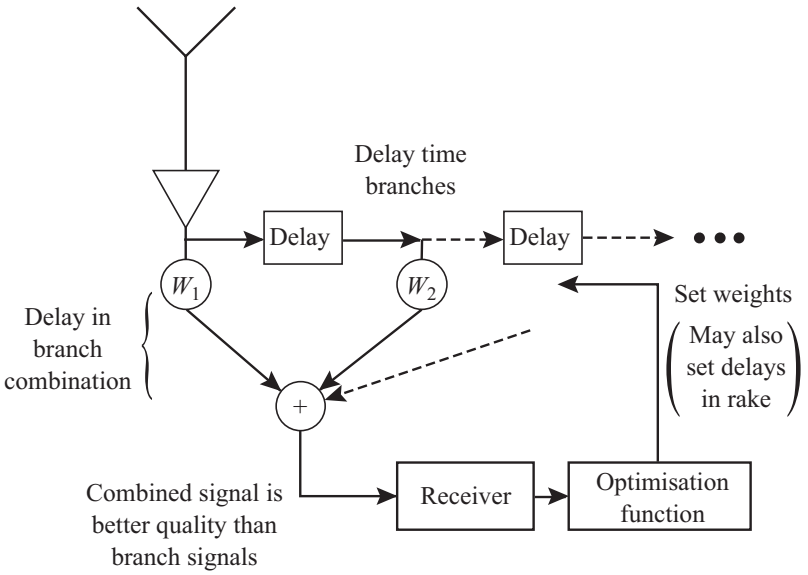


Figure 1.8.3 Time diversity, or rake reception, combines the signals arriving at different times that comprise mutually uncorrelated fading branches

The nature of the propagation environment and the antenna reception affects the distribution of the echoes in the impulse response. The averaged impulse response often displays decreasing power for increasing delay time. This means that the mean powers of the successive delay time branches are different and this decreases the diversity performance relative to a number of branches with equal mean powers. Likewise, if the tap spacing is too small, then the branches become mutually correlated and the diversity performance decreases relative to the same number of branches with no correlation.

The advantage of time diversity is that it does not need multiple antennas to reduce fading by making use of the transmitted energy that is otherwise neglected (the echoes) and which can otherwise create reception problems by causing symbol distortion and intersymbol interference. A disadvantage is that time diversity requires wideband signals in order to allow the time delay time resolution of the echoes in the propagation channel and, in a sense, the system designer is accepting the fading channel and attempting to compensate through the signal processing and coding. This is in contrast with antenna diversity, which, at least in predetection combining, attempts to repair the fading channel before applying other signal processing and coding.

1.8.2.4 Equalisation for ISI in frequency-selective channels

Equalisers have a similar structure to rake reception, i.e. a transversal filter, and the difference between the two techniques is not always well defined. The rake can be viewed as a true diversity technique where mutually uncorrelated branches bearing the same basic information are combined, and a diversity gain (for example, less fading than any one of the branches) is realised as a result. The rake tap spacing is related to the dispersion caused by the propagation environment.

On the other hand, the function of the equaliser is to compensate for the signal distortion introduced by the dispersion of the channel. Typically, the tap spacing is related to the symbol time of the signal being transmitted, rather than to the propagation channel dispersion. The combining algorithm of the equaliser is based on reducing distortion such as amplitude nonlinearity or phase nonlinearity, or intersymbol interference. In this sense, equalisation can be therefore used with any diversity, including with rake reception. A combination of rake reception and equalisation can be implemented with a transversal filter structure and it is in such cases that there is not necessarily a clear difference between the rake and equalisation techniques.

As in rake and antenna diversity where there are several combination algorithms, there are also several equaliser algorithms. Classically, equalisers used a zero-forcing (of the intersymbol interference) algorithm to attempt to remove or compensate for the dispersion within the bandwidth of the channel. In decision feedback equalisers (DFEs), the transversal filter can feature a feedback and feedforward structure, and the algorithm sets the weights to minimise the difference between a known training sequence, which is transmitted regularly, and the corresponding received sequence. In the Viterbi (1967) equaliser, the impulse response of the channel is estimated and, from this, all the possible transmitted symbol sequences are assigned a range of possible sequences which will be detected by the receiver. So for reception of a given sequence

of symbols, the most likely transmitted sequence can be selected from the calculated range of expected output sequences (maximum likelihood sequence estimation). The channel estimation is usually made by measuring the impulse response with a special sequence of symbols whose autocorrelation should have a well-defined peak and low sidelobes. In CDMA systems, the sequence is also the code used to identify uniquely the user, and so low cross-correlation properties between the many sequences also become important. Problems arise when the channel is changing quickly, as in mobile communications, because the estimate of the channel impulse response becomes inaccurate.

Equalisers are used in many systems for combating the spread of the delays imposed by the propagation channel. This means there is a combination of rake action and distortion compensation, and the diversity gain, for a given bit-error rate, is one measure of performance.

1.8.2.5 Linear combination of antenna branches and time-dispersed signals

To combine the antenna branch signals in the frequency-selective channel, frequency-dependent weights are required. This is implemented for digital communications by using an equaliser-type combiner in each branch before combination, as shown in Figure 1.8.4. This is the optimum linear combining structure for frequency-selective

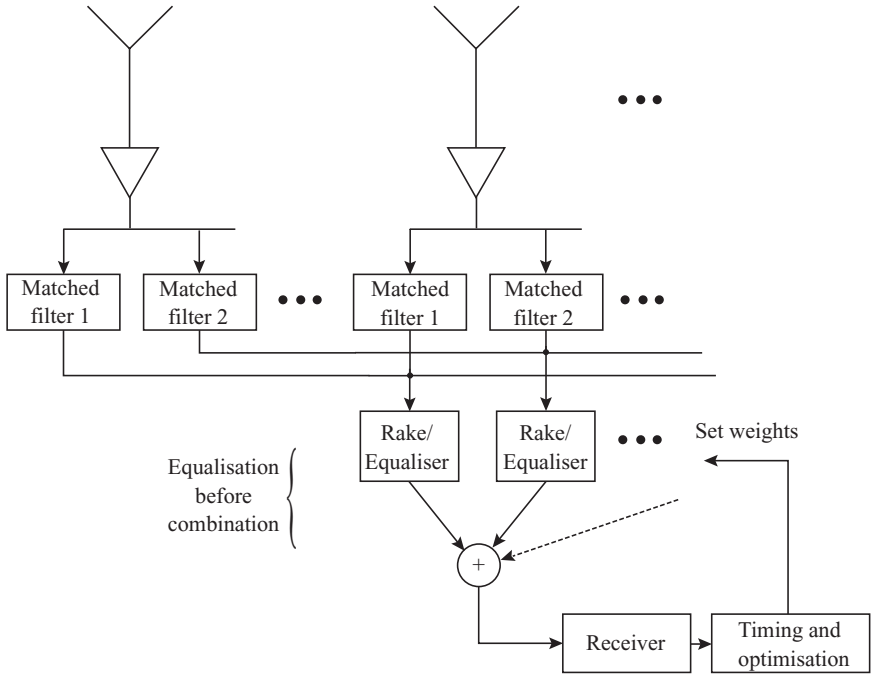


Figure 1.8.4 *A structure for optimum linear signal combination for a digital communications system and a frequency-selective channel. Digital matched filters precede equalisation before the signal is detected. The detection process generates the cost functions for optimisation of timing and sets the equaliser weights*

channels (Clark *et al*., 1994). The performance of the optimum linear combiner is very close to the optimum for the digital channel which requires nonlinear operations such as decision feedback and the Viterbi algorithm. In the linear structure, the antenna branch signals are put through a bank of digital filters that are matched to the alphabet of digital symbols. The outputs of these are added and subjected to equalisation, and these are in turn combined before detection. After detection, sample and symbol timing is optimised, and the equaliser weights are generated.

The concept of degrees of freedom is again useful for getting a feel for the potential performance. For N_A antennas, each with a rake receiver with N_R taps, and with all the antenna branches and rake taps having equal mean power and being mutually uncorrelated, there are $N_A N_R$ DOFs, and these can be divided up between interference suppression and gain improvement. Differences in mean power and finite mutual correlations will degrade the performance. The effective number of DOFs can be likened to an equivalent number of ideal, i.e. equal mean, uncorrelated branches (always less than the number of actual branches in an imperfect implementation). This number gives an indication of the performance of a diversity system, discussed in Chapter 8. As a rule of thumb for the digital channel, M (ideal) branch diversity gives a mean BER improvement over the single branch channel of $BER_M \approx (BER_1)^M$ and similarly for the outage, $P_{OM} \approx (P_{O1})^M$. The inclusion of an equaliser will also improve performance by suppressing ISI. As noted above, the rake and equaliser would probably be implemented in the same way as the transverse filter.

1.8.2.6 Frequency diversity

The frequency-dependent behaviour means that a choice of frequency-spaced channels also offers diversity branches. Frequency diversity has been traditionally used on point-to-point links which have fading. When the link fades, or suffers equivalent quality degradation, the carrier frequency is changed, i.e. the combiner type is switched combining. In mobile systems, *slow frequency hopping* can be used as a form of frequency diversity. The basic condition for channel choice is that the radio frequency spacing between the channels is sufficient to feature mutually uncorrelated fading. This spacing is called the coherence bandwidth. If the frequency spacing is smaller that the coherence bandwidth, then the branches become mutually correlated and the diversity gain reduces. The advantage of using frequency diversity is that there is a diversity gain available without the need for multiple antennas (for antenna diversity) or wideband signals (for time diversity), but the disadvantage with respect to antenna diversity is that, ultimately, extra spectrum is required.

1.8.2.7 Transmit diversity

The above discussion has revolved around diversity as a reception mechanism. However, the basic principle holds for transmission as well. Multiple transmit antennas can transmit to a single receiving antenna with the same diversity gain, i.e. reciprocity holds in this sense, but the combiner (now a power splitter with weights) requires information from the receiving end in order to set the weights. Frequency-diverse branches can be transmitted simultaneously, although this may not be power

or bandwidth efficient; otherwise coordination between transmitter and receiver for the current choice of frequency is required. Frequency-hopping schemes employ a predetermined sequence of frequency choice. Time diversity can be arranged by transmitting a set of replica signals at a set of delay times, which are then combined (e.g. selected from) at the receiver. This technique effectively acts to impose frequency selective fading on the channel before transmission.

1.8.2.8 Transmit and receive diversity

Antenna diversity at both ends of a multipath link transforms the disadvantage of multipath degradation into the advantage of being able to synthesise many (diversity combined) channels which offer independent usage of the same spectral band. The capacity can thus be increased simply by adding more antennas and processing, and not by consuming extra spectrum. Realising this development in a multiple-access system requires an understanding of the propagation mechanisms, antenna principles, the channel behaviour and mathematical models and descriptions, and the signal-processing techniques.

1.8.2.9 Convergence of multiple access, modulation, and multipath mitigation schemes

The functions of modulation, coding, multiple-access technique, and multipath mitigation are all realised through electronic signal processing. Although these functions are quite different, the boundaries between them are sometimes blurred by clever multifunction techniques or else in the implementation. An obvious example is CDMA, which can be viewed as a multiple-access technique, or a coding technique, or as a modulation. Antenna techniques are also being viewed as a multiple-access technique, called space division multiple access (SDMA). In many current systems, a combination of access techniques is used. In GSM (ISM-126), DECT and PHP systems, for example, a combination of both TDMA and FDMA is used. These systems profit from equalisation and antenna diversity. Even DECT systems, which were originally configured not to use diversity or equalisation, can profit from both techniques. Thus diversity and equalisation can be used as *a posteriori* improvements for existing systems, and now that the techniques are becoming more ubiquitous, space-time systems are being deployed that have the techniques as necessary components for system operation.

1.8.2.10 Slow fading and macro diversity

The dynamic range of the received signal is also affected by the *slow fading*, which is also called *long-term fading* or *shadow fading* as noted in Section 1.4. This is superimposed on the fast fading and is caused by a shadowing of the radio signal to the scatterers as the mobile moves behind large obstacles such as hills and buildings, etc. The spatial density of the shadow fading therefore depends on the large-scale nature of the physical environment. The basic short-term multipath mechanism remains unchanged. The dynamic range of the slow fading is much less than that of the fast fading, being confined to about ± 10 dB for most of the time in urban and suburban

environments. The total dynamic range for the fading therefore becomes about 70 dB. The distance-based path loss, as a mobile roams near to, and far from, a base station, adds to this range. For closely spaced diversity antennas, there is little that can be done about the shadow fading because all the antenna elements are affected similarly by the fading. Alternatively stated, the shadow fading is fully correlated for all the receiving antenna elements and there is no diversity gain available. This situation can be improved by *macro diversity* – separating the antennas by large distances. This is not possible at a mobile where the terminal is electrically small, but base stations can be spaced by using physically separated sites (such as different buildings for outdoor, or different walls for indoor, environments) for the antennas. The shadow fading 'seen' at each site is uncorrelated and a diversity gain with respect to both the shadow fading and Rayleigh fading can be achieved by combining the signals from the separated sites.

1.8.2.11 Mean path loss limits diversity action

Diversity action, in its classical form, cannot assist with a path loss associated with decreasing electromagnetic power. The *path loss* is the loss of signal power between matched antennas, i.e. it is the loss of the electromagnetic signal channel of Figure 1.6.1. In free-space conditions, for example, a distance d and wavelength λ, the path gain (reciprocal of path loss) is $(\lambda/4\pi d)^2$. The shadowing and Rayleigh fading both contribute to the path loss. The *mean path loss* normally refers to averaging of the path loss over the Rayleigh fading. An example is shown in Figure 1.8.5 (Butterworth, 1999), which gives an idea of the distribution of mean path loss measured for an indoor environment. The antennas are of the vertical dipole type, so that the patterns are essentially omnidirectional in the plane depicted. The contours cover a dynamic range of some 47 dB and it is clear that there is a tendency for the mean path loss to increase as the distance to the transmitter increases. By superimposing the Rayleigh fading, the dynamic range of path loss over the indoor environment increases by up to $+10$ dB and -40 dB, i.e. to a total of almost 100 dB. The diversity can improve the Rayleigh component, and macro diversity can mitigate the shadow component, but the basic distance-based path loss effectively comes down to the same as the free-space situation: only the antenna gains can improve the situation. This is possible by having enough antenna elements to give sufficient DOFs to deploy for gain increase. As noted above, antenna diversity is the only signal-processing option that can perform this task without using extra spectrum. Using sophisticated signal processing, the antennas can be shared between the transmit and receive ends, and this is covered in Chapter 9.

1.9 Example analysis of simple digital wireless link

The chapter is concluded with an example of a simple wireless digital link between an indoor (say a domestic house) and an outdoor pole-mounted antenna for a cellular-like system. It demonstrates the connection between the antennas, the propagation

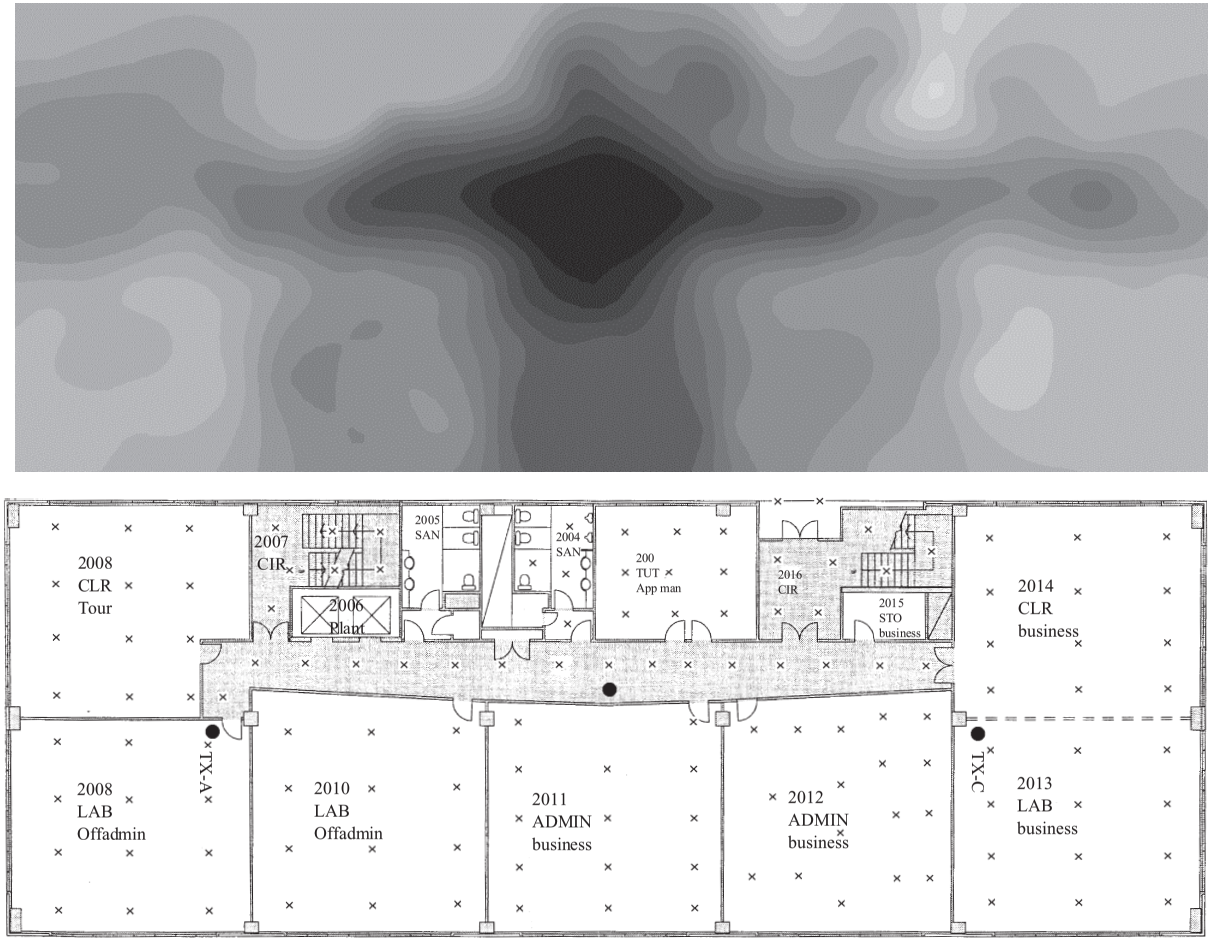


Figure 1.8.5 The mean path loss (averaged over the Rayleigh fading) between omnidirectional antennas in an indoor environment. The dynamic range is about 47 dB (about 2.7 dB per contour) and demonstrates clear distance-based and shadow-based loss with respect to the walls of the building (Butterworth, 1999)

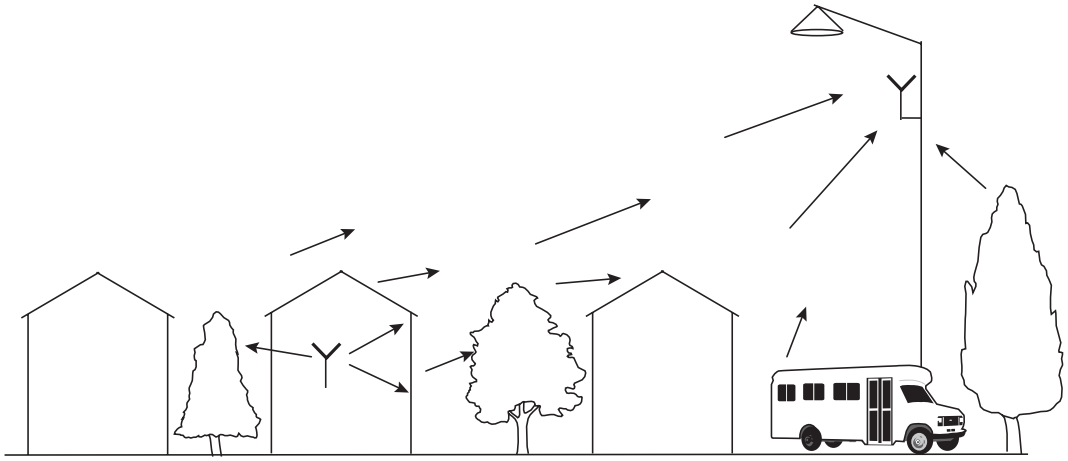


Figure 1.9.1 An indoor/outdoor system has a radio link from an indoor-mounted antenna to an outdoor antenna

and the digital transmission. The issues raised help to illustrate the motivation for the following chapters.

The situation is depicted in Figure 1.9.1. There is a mixed path, comprising multiple wall penetration, scattering, and small free-space distances for the indoor part, and a larger free-space distance for the outdoor part. In practice the outdoor part may comprise mixed paths as well, but this should be avoided if possible in order to keep the path losses to a minimum.

1.9.1 Free space path loss

The free-space situation offers the simplest and lowest power radio link for a given reliability. It is assumed that the antennas are lossless and matched in both impedance and polarisation in the sense that these combined antenna losses are negligible. The basic transmission link equation is discussed in Appendix A, and is simply

$$P_{\mathrm{R}} = \frac{P_{\mathrm{T}} G_{\mathrm{T}} G_{\mathrm{R}}}{L_{\mathrm{P}}}, \tag{1.9.1}$$

where P_{R} is the received power, P_{T} is the transmitted power, G_{T} is the gain of the transmitting antenna in the direction of the receiver and, similarly, G_{R} is the receiver antenna gain, and L_{P} is the path loss. The free-space path gain is just the inverse of the loss, i.e. $G_{\mathrm{P}} = 1/L_{\mathrm{P}}$. In free space, the path loss is

$$L_{\mathrm{P0}} = \left(\frac{4\pi d}{\lambda}\right)^2, \tag{1.9.2}$$

where d is the free-space distance and λ is the carrier wavelength. For convenience, the path loss can be written in logarithmic terms with the frequency in MHz and the distance in km, as

$$L_{\mathrm{P0}}(\mathrm{dB}) = 32.4 + 20\log_{10} f_{\mathrm{MHz}} + 20\log_{10} d_{\mathrm{km}}. \tag{1.9.3}$$

1.9.1.1 Example for basic SNR calculation

As an example, for vertical dipoles with a gain of 2, separated in free space by 10 km, a transmitted power of 4 W at 900 MHz results in a field strength at the receiving dipole of about 15.5 mV/m, a path loss of 111.5 dB, and a received power of

$$\begin{aligned} P_{\mathrm{R}}(\mathrm{dB}) &= 10\log_{10}(4) + 10\log_{10}(2) + 10\log_{10}(2) - 111.5 \\ &= 6 + 3 + 3 - 111.5 \\ &= -99.5\,\mathrm{dBW}. \end{aligned} \tag{1.9.4}$$

The noise at the receiver is taken to be dominated by the noise of the electronic receiving electronics, rather than the antenna noise (discussed in Chapter 7). If the thermal noise level is at $-174\,\mathrm{dBm\,Hz^{-1}} = -204\,\mathrm{dBW\,Hz^{-1}}$ (this is Boltzmann's constant, $k_{\mathrm{B}} = 1.38(10)^{-23}\,\mathrm{J\,K^{-1}}$, times the room temperature in kelvin, 300 K), the bandwidth is say 200 kHz, and the noise figure of the receiver is 10 dB, then the signal-to-noise ratio at the receiver is

$$\begin{aligned} SNR(\mathrm{dB}) &= Signal(\mathrm{dB}) - Noise(\mathrm{dB}) - Noise\ figure(\mathrm{dB}) \\ &= -99.5 - (-203 + 53) - 10 \\ &\approx 40.5\,\mathrm{dB}. \end{aligned} \tag{1.9.5}$$

This large SNR gives a very small BER. Even at a low (for wireless) BER of 10^{-6}, using a simple modulation scheme of incoherent frequency shift keying (FSK) requires only about 14 dB of energy per bit-to-noise ratio. This means that the transmitted power can be reduced to fractions of milliwatts, or, alternatively, for example, the link distance can be extended to hundreds of kilometres (of free space).

1.9.1.2 Modulation and BER

For slightly greater receiver complexity, the modulation can be made more advanced. Coherent FSK allows a decrease of transmitted power of a couple of dB, and higher-order modulations improve things further. For additive white Gaussian noise (thermal noise, for example) and a static, free-space channel, the BER for incoherent FSK is expressed as the probability of a bit error (see, for example, Proakis, 1995)

$$P_{\mathrm{b}} = \frac{1}{2}\exp\left(-\frac{1}{2}\frac{\epsilon_{\mathrm{b}}}{N_0}\right). \tag{1.9.6}$$

For coherent FSK, the situation improves to

$$P_{\mathrm{b}} = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{1}{2}\frac{\epsilon_{\mathrm{b}}}{N_0}}\right), \tag{1.9.7}$$

and stepping up to QPSK, with symbol error probability, $P_{\mathrm{S}} = 2P_{\mathrm{b}}$, and

$$P_{\mathrm{b}} = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\epsilon_{\mathrm{b}}}{N_0}}\right). \tag{1.9.8}$$

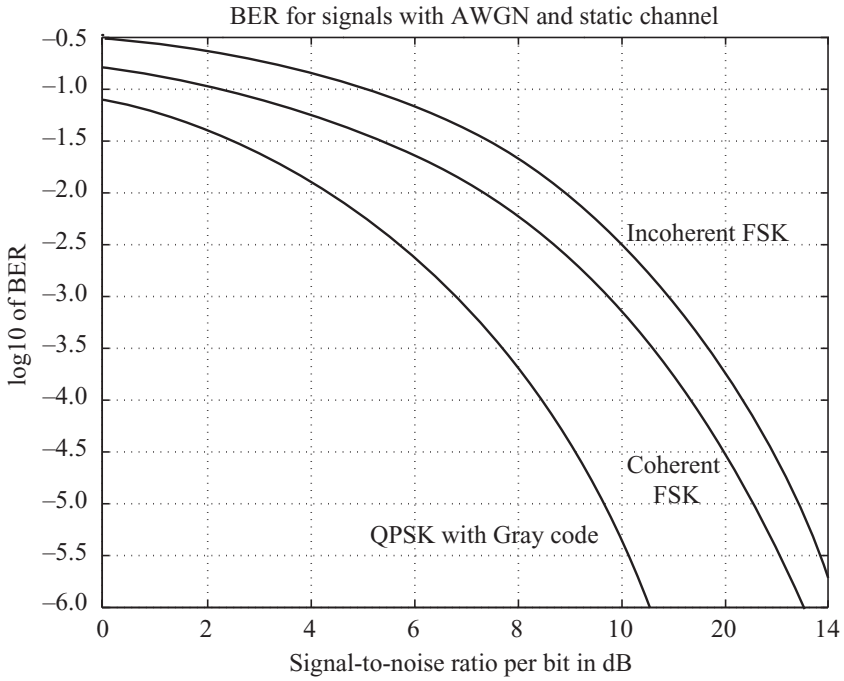


Figure 1.9.2 The BER curves for different modulations show how more-advanced modulations allow a more reliable grade of service for given SNR, or a decreased SNR requirement for a given grade of service

Here, ϵ_b/N_0 is the signal-to-noise ratio per bit and for QPSK, the symbol-to-noise ratio is $2\epsilon_b/N_0$. The curves are given in Figure 1.9.2. The conversion to SNR is $SNR = \epsilon_b R_b/(BN_0)$, where R_b is the bit rate and B is the bandwidth. The decreased SNR requirement for better modulations is clear, with higher-order modulations (not shown) following the same trend.

1.9.1.3 Path loss and BER

The path loss directly affects the SNR, and hence the BER. True free-space conditions are seldom possible in terrestrial communications. Even line-of-sight links often have at least a ground bounce as a path component, as well as the direct path. This can cause a path-loss rate of an inverse-fourth law, as developed in Chapter 3. Many point-to-point systems are well modelled by three paths but the indoor antennas and pole-mounted antennas are probably non-line-of-sight, in which case there are many paths contributing to the link. The many paths cause signal fading and this, plus the extra path loss from wall penetration, etc., contributes to degrade the SNR.

Also, most systems suffer from some sort of interference. The effect of interference is discussed first, followed by the multipath. In a multi-user system, the noise is often dominated by the interference. If the distribution of the interference is Gaussian, then the SNR can be replaced by the SINR in the bit-error ratio expressions.

1.9.1.4 Interference in a multi-user scheme

A multi-user scheme, in which spectrum is shared simultaneously in order to have spectral efficiency, will normally result in the presence of co-channel interference. This has a serious effect on the average BER. Where there is no pressure for spectral efficiency, time slots can be used to avoid co-channel interference. One interferer only is considered (this is a worse situation than several interferers with total power the same as the single interferer), with power S_I

$$P_\mathrm{b}(SNR) = \int P_\mathrm{b}(SNR, S_\mathrm{I})\, p_{S_\mathrm{I}}(S_\mathrm{I})\, \mathrm{d}S_\mathrm{I}. \tag{1.9.9}$$

The interferer is taken to be independent of the wanted signal and to have a uniformly distributed probability density function (pdf) for its envelope. This gives, for QPSK,

$$P_\mathrm{b} = \frac{1}{4\sqrt{\alpha SNR}} \int_{-\sqrt{\alpha SNR}}^{\sqrt{\alpha SNR}} \mathrm{erfc}\left(x + \sqrt{\frac{SNR}{2}}\right) \mathrm{d}x, \tag{1.9.10}$$

where SNR is the symbol signal-to-noise ratio and α is the ratio of interferer power-to-wanted signal power. Figure 1.9.3 gives the graphical form. The degradation cuts in once the interferer comes within 20 dB of the wanted signal and the effect becomes serious below 10 dB. A single interferer at 10 dB below the wanted signal costs about 4 dB of SNR for maintaining the average BER. As the interferer becomes stronger

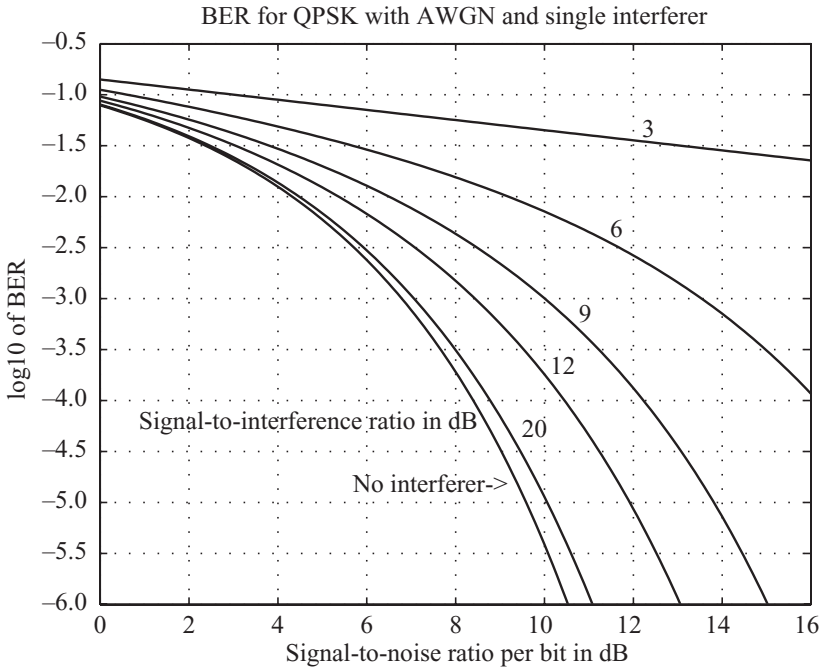


Figure 1.9.3 The effect of an interferer in a free-space link is to reduce the grade of service or alternatively increase the required SNR for a given grade of service

than this, the cost increases quickly. This gives an indication for the configuration of frequency re-use in a cellular-like scheme using line-of-sight links.

1.9.2 Non-line-of-sight

The non-line-of-sight situation creates a larger average path loss and a fading signal. The larger path loss is owing to the extra free-space path loss of the multipath propagation, and the electromagnetic scattering between the free space paths. Some paths involve penetration through walls, etc., and these also suffer excess attenuation relative to the same distance in free space. The SNR becomes smaller, with spatial fading, typically with a Rayleigh distribution imposed on the envelope as discussed earlier. This fading can be time varying, caused by changes in the multiple paths in the link. This can be very slow fading, caused by seasonal vegetation changes; or quicker, caused for example by vehicular or person/animal movement, or wind-driven structure movement; or incremental changes such as a shift of furniture or door, etc., in the premises, or the parking of a vehicle.

The effect of the fading on the average BER is very important. The easiest formulation results from assuming that the fading rate is very slow compared to the signalling rate. Then we can just condition the BER expression with the probability of the SNR,

$$P_{\mathrm{b}}\left(\overline{SNR}\right) = \int P_{\mathrm{b}}(SNR)\, p_{SNR}(SNR, \overline{SNR})\, \mathrm{d}SNR, \tag{1.9.11}$$

where the overbar indicates an average value. For Rayleigh fading, the power is an exponential distribution, i.e. the probability density function (pdf) is

$$p_{SNR}(\gamma) = \frac{1}{\Gamma}\mathrm{e}^{-\gamma/\Gamma} \quad \gamma\ (= SNR) > 0 \tag{1.9.12}$$

and Γ is the mean of the power. A result is depicted in Figure 1.9.4. As a result of the signal fading, the form of the BER curves changes as noted in Section 1.7 and the big returns available from small increases in the SNR are lost. For an average BER of 10^{-3}, the QPSK signal requires a mean SNR of about 17 dB more than that for a non-fading link, and this can be interpreted directly as the required increase in transmitted power. For example, the transmit power would increase from, say, 80 mW to 4 W. As the fading rate increases, the curves will change form again, and level out as shown in Figure 1.7.1. The simple formulation above will not show this. See Korn (1996) for a general approach to the fading channel BER behaviour.

Even if the fading is considered unchanging with time for each link, there is still the same distribution of path losses for the individual terminals. The cumulative probability, or cumulative density function (cdf) can be found by transforming the BER expression and integrating. For example, for coherent FSK, the result is

$$\mathrm{Prob}\,(BER < BER_0) = 2\,(BER_0)^{2/\Gamma_{\mathrm{S}}}\,, \tag{1.9.13}$$

given graphically in Figure 1.9.5. As an example, for a 99 per cent probability (i.e. for 99 per cent of the links) that the BER is less than 10^{-3}, we require a mean value

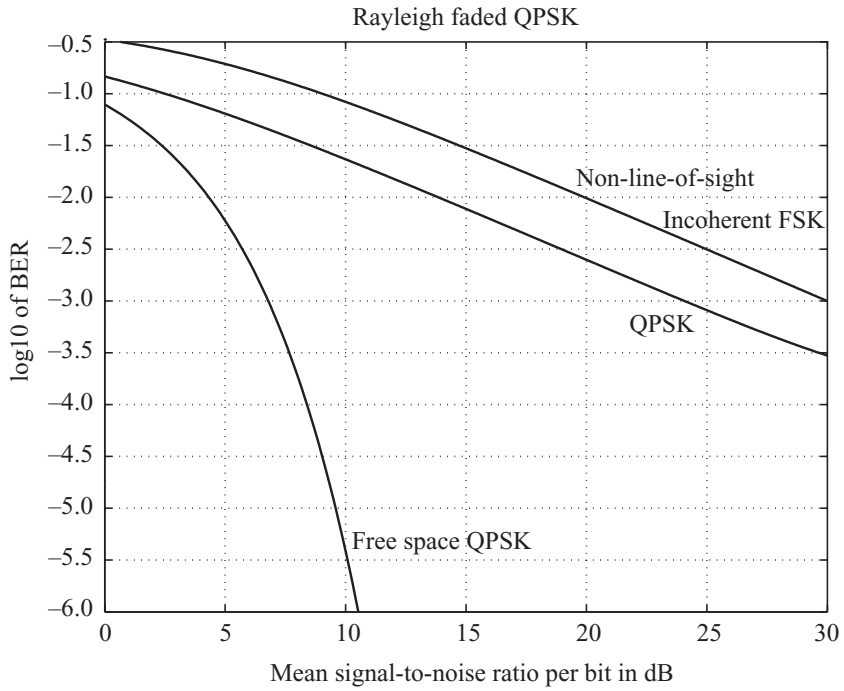


Figure 1.9.4 The effect of the non-line-of-sight is to decrease the grade of service by reducing the SNR for a given transmitted power, and also to reduce the grade of service through the variation of SNR. The probability of the BER is calculated by integrating over the fading density of the SNR

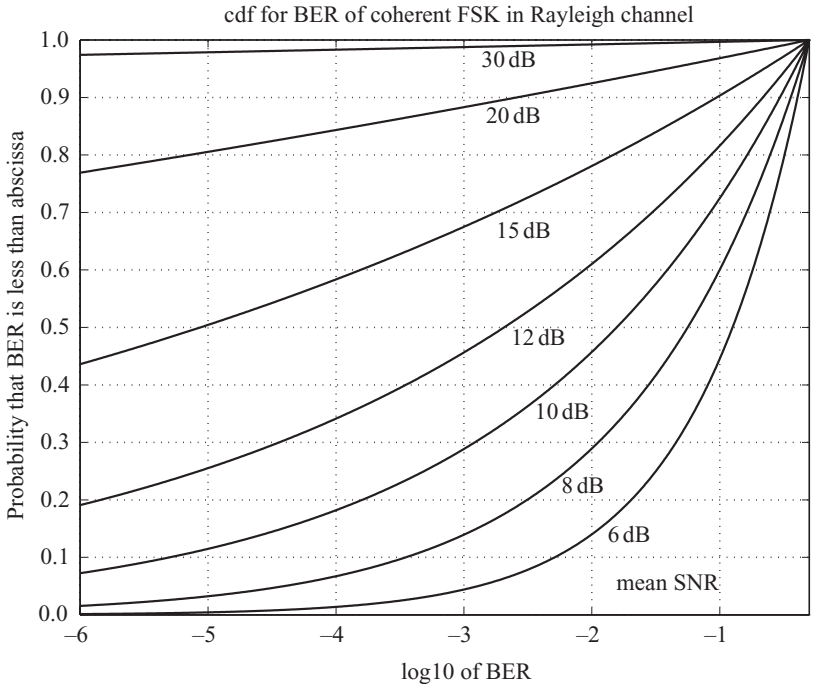


Figure 1.9.5 The cdf for the BER for terminals in Rayleigh channels

for the SNR of 30 dB. This compares with an SNR requirement of about 10 dB for the static (ideal, line-of-sight) channel.

1.9.2.1 Non-line-of-sight and interference

The combined effects of fading and interference, with the above assumptions, here for QPSK, result in the formulation

$$P_{\mathrm{b}} = \frac{1}{4\alpha S} \int_0^{\infty} \int_0^{\infty} \int_{-\sqrt{\gamma_{\mathrm{I}}}}^{\sqrt{\gamma_{\mathrm{I}}}} \operatorname{erfc}\left(x + \sqrt{\frac{\gamma_{\mathrm{S}}}{2}}\right) \mathrm{e}^{-\gamma_{\mathrm{I}}/\alpha S} \, \mathrm{e}^{-\gamma_{\mathrm{S}}/S} \, \mathrm{d}x \, \mathrm{d}\gamma_{\mathrm{I}} \, \mathrm{d}\gamma_{\mathrm{S}}. \tag{1.9.14}$$

No further reduction of the integral seems possible and numerical solution is required. However, the integrand is small over most of the integration region, so the relevant integration region is small, and the integrand is otherwise well-behaved. So for system evaluation, fast and smart numerical integration routines are required. The combined effects are not discussed here, but the trend will be to reduce any grade-of-service measure further by degrading the average BER.

1.9.3 Measurement-based indoor/outdoor path loss

The mixed propagation path for getting from an indoor antenna to an outdoor antenna is dependent on many environmental factors. While it is possible to calculate the path loss through an individual wall or around a well-defined obstacle, and other canonical situations, the mixed-path propagation and absorptive environment is best investigated experimentally. Chapter 4 discusses several models. An example of a measurement-based model uses results from van Loon (1999) and Andersen (1999). Here, measurements of the best-case indoor path loss for five different building types were taken and an empirical formula obtained. The buildings were in Europe, comprising: a simple 1956 medium-quality house; a 1985 high-quality, insulated apartment building; an 1846 high-quality stone-walled house; a low-quality 1971 apartment; and a 1956 farm house built of mixed mortar and wood. Best values (defined here by the 1st percentile) and worst values (99th percentile) were recorded.

An empirical model can be obtained from the measurement data. For a separation distance of d_{m} metres from the transmitter, the best case is given by

$$\mathrm{PL(dB)} = -23 + 90 \log_{10} d_{\mathrm{m}}, \quad d_{\mathrm{m}} > 3\mathrm{m}; \tag{1.9.15}$$

and presented graphically in Figure 1.9.6. The measurement range was confined to 5–10 m, so ranges outside of this (on the figure) are extrapolated. The worst-case path loss can be calculated from the Rayleigh fading to be another 27 dB below this. So the worst-case coverage requires about 27 dB to be added to the path loss.

For other frequencies, the free-space dependency of the path loss still holds. For example, at 866 MHz, the path loss is a factor of about 6 dB greater, or at 2.45 GHz, the path loss is a factor of about 15 dB greater. These factors should be added to the model of Figure 1.9.6 for these frequencies.

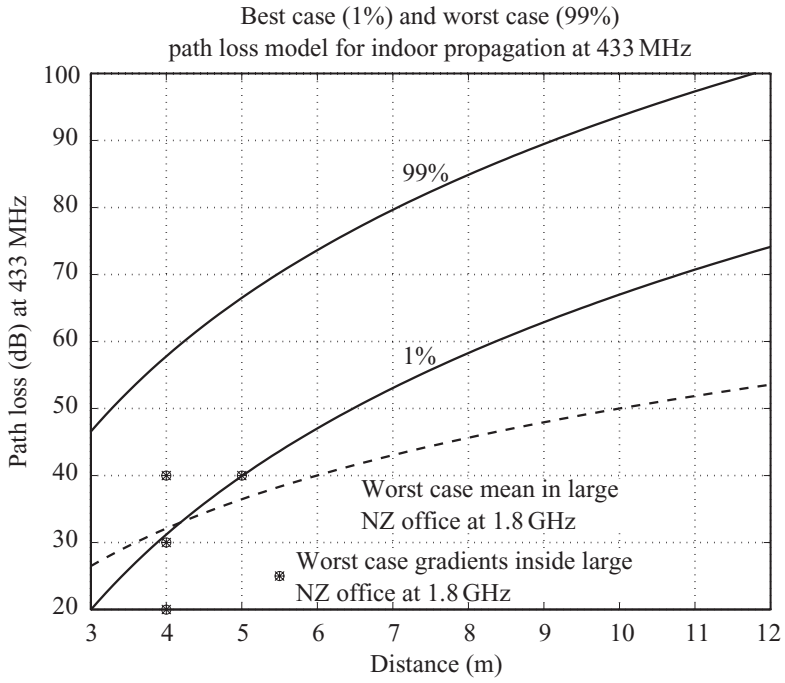


Figure 1.9.6 The measurement-based path loss model for indoor propagation at 433 MHz. The broken curve is a measurement-based model at 1.8 GHz for large offices. The dots are worst case gradients (for example, 40 dB in 5 m) of the path loss for offices measured (Butterworth, 1999) at 1.8 GHz

Another data set on the figure is from measurements at 1.8 GHz in a New Zealand office building (Butterworth, 1999). The offices are large and separated by (open) doors, there are long corridors, and walls of gib-board and wood, and some walls of thick concrete. The worst-case path loss with distance for the median of the signal (the median is often used in measurements because it is quick to calculate an unbiased estimate from samples) is given approximately by

$$\mathrm{PL(dB)} = 5 + 45 \log_{10} d_{\mathrm{m}}, \quad d_{\mathrm{m}} > 3\,\mathrm{m}; \tag{1.9.16}$$

and is shown as the dotted curve. The worst-case gradients in the measurements from New Zealand are given as dots on the plot. These occurred when the direct-path obstacle was a concrete wall or densely stacked steel shelving. These measurements are included to provide the contrast with those at 433 MHz.

1.9.4 Link example

- *The transmitting frequency is 433 MHz. The transmitting antenna is indoors. It has a gain of 1, an efficiency of 50 per cent, has a VSWR of 8, and the matching circuit and balun loss is 3 dB. For linear polarised antennas, we allow a 3 dB polarisation coupling loss*

Transmitting antenna gain: 0 dB
Loss from antenna efficiency: 3 dB
Loss from antenna mismatch: 4 dB
Loss in matching circuit and balun: 3 dB
Loss in polarisation match: 3 dB
Total power loss at transmitter: 13 dB

- *The transmitter is mounted 4 m away from the wall facing the pole-mounted receiver, which is in sight. The distance to the pole from the house is 80 m.*

  Indoor path loss for 99 per cent of cases (Figure 1.9.6): approx. 58 dB
  Free-space path loss to base-station antenna: $32.4 + 52.7 - 21.9 = 63.2$ dB
  Total path loss: approx. 121 dB

- *The base station antenna has a gain of 2, has 1 dB matching loss and is perfectly efficient with no losses in the matching circuit or balun.*

  Receiving antenna gain: 3 dB
  Receiving antenna and total mismatch losses: 1 dB
  Total receiving antennas losses: -2 dB

Total path loss for 99 per cent: approx. $13 + 121 - 2 = 132$ dB

1.9.4.1 Discussion

In the above example, for a transmitted power of 1 W, the received power for 99 per cent of cases is -132 dBW. If the noise level is -141 dBW (from the previous example), and the bit rate is 1 s^{-1} then the energy per bit-to-noise ratio is 9 dB. Even with no interference, and using incoherent FSK, we require about this for a BER of 10^{-3}. So accounting for these degradations, there is a shortfall of signal strength in this example. The biggest problems are in the path losses, which for this example are split approximately evenly between the free-space outdoor path and the worst-case indoor path (i.e. the multipath). The path-loss factors can be attacked directly by antenna techniques, and/or the SNR can be improved by other signal-processing techniques.

Note that by using an externally mounted antenna, the SNR ratio for this example would give plenty of room in the SNR for catering for interference and allowing better coverage. However, there may still be multipath even with outdoor-mounted antennas which must be accounted for.

The following points summarise further design options.

- Using lower carrier frequencies reduces the path loss, but may increase the antenna-loss contributions as the antennas need to become electrically compact.
- Reduction of the bandwidth will reduce the thermal noise contribution, and the SNR will, in turn, increase.
- Using a slower signalling rate, i.e. larger bit durations, makes it possible to increase the energy per bit and this will have a direct impact on the energy-per-bit-to-noise ratio and the BER.

- Higher-level modulations, and also error-correction coding (for example, ARQ or FEC techniques) can also help by imposing a processing gain which can be included in the link budget.
- Anti-multipath techniques (Section 1.7), such as antenna diversity or frequency diversity, can reduce the worst-case path loss by tens of decibels, although they require more sophisticated processing and electronics at each end of the link and more sophisticated antennas.
- Reducing the cell size so that the free-space distance is much smaller will also help by reducing the space loss – this is the beauty of the cellular concept. But this comes with several costs which need to be designed for: it increases the number of pole-mounted antennas and therefore the cost of the infrastructure; it increases the number of frequency re-use channels required; and it increases the interference levels.

1.10 Summary

This introductory chapter gives a coverage of the evolution and basics of mobile and personal communications. Details of the digital communications systems issues will not be pursued in the following chapters, except where necessary for a particular topic. The roles of the propagation and antenna engineering are seen respectively as limiting and mitigating factors in a mobile link. The multipath propagation and the path loss limit communications capacity, and the classical gain and diversity gain from signal processing undertaken by the antennas serve to mitigate the capacity limitations. Some of the subsequent chapters discuss the mechanisms of multipath propagation and the antenna signal-processing action and antenna configurations. It is these processes that govern the performance of mobile links.

1.11 References

ALTGELT, C.W., DUFFY, K.H., and ROWAN, C.W.: 'The evolution of the automobile antenna in the United States and Europe – A historical perspective, Part Two: The last fifty years', *SAE* 880085, February 1988.

ANDERSEN, J. BACH: 'Comments on "Mobile in-home Uhf radio propagation for short-range devices" ', *IEEE Antennas and Propagation Magazine*, October 1999, **41** (5), pp. 122–123.

AULIN, T.: 'Characteristics of a digital mobile radio channel', *IEEE Transactions*, May 1981, **VT-30** (2), pp. 45–53.

BARTON, S.K. and WILKINSON, T.A.: 'HIPERLAN, the high performance radio local area network', COST-231 Document TD(94) 51, 1994.

BUTTERWORTH, K.: 'Performance, planning and deployment of DS-CDMA in-building wireless communications systems', PhD thesis, Auckland University, New Zealand, March 1999.

CCIR {ITU-R}: *CCIR Task Group 8 Document* Working Party 8B, 8/7-E, December 1991.

CCIR {ITU-R}: *CCIR Task Group 8 Document* 8/51, January 1992a.
CCIR {ITU-R}: *CCIR Task Group 8 Document* 8/55E, January 1992b.
CLARK, M.V., GREENSTEIN, L.J., KENNEDY, W.K., and SHAFI, M.: 'Optimum linear diversity receivers for mobile communications', *IEEE Transactions on Vehicular Technology*, February 1994, **VT-43** (1), pp. 47–56.
COE, D.: *Marconi, Pioneer of Radio*, Simon and Schuster, New York, 16th printing 1967.
COST-1999: COST ACTION 231, *Digital Mobile Radio Towards Future Generation Systems*, Directorate-General, Telecommunications, Office for Official Publications of the European Communities, Luxembourg, ISBN 92-828-5416-7, 1999.
COX, D.C.: 'Personal communications – a viewpoint', *IEEE Communications Magazine*, November 1990, pp. 8–92.
DELOGNE, P. and BAAN, W.: 'Spectrum congestion', in Stuchly, M.A. (Ed.) *Modern Radio Science*, pp. 309–327, Oxford University Press, 1999.
FOX, B.: 'Mobile phones fit for the future', *New Scientist*, 8 April, 1992, pp. 23–24.
GAGLIARDI, R.M.: *Introduction to Communications Engineering*, Wiley, 1988.
HEALD, A.B., SCOTT, N.L., and VAUGHAN, R.G.: 'Digital demodulator, implementation parameters evaluated by simulation', *Proceedings of the IEEE Global Telecommunications Conference* (GLOBCOM 98), Sydney, November 1998, **2**, pp. 1303–1308.
HENRY, P.S.: 'Spectrum efficiency of a frequency-hopping DPSK spread-spectrum mobile radio system', *IEEE Transactions*, November 1979, **VT-28**.
ITU (1992): *ITU Press*, ITU/92-2, 3 March 1992.
JAKES, W.C. (Ed.): *Microwave Mobile Communications*, Wiley, 1974.
KAGOSHIMA, K., TAKAHASHI, T., and EBINE, Y.: 'Dual frequency low profile printed dipole yagi-array for a cabin mounted vehicular antenna', *Antennas and Propagation Society Symposium Digest*, 1991, **2**, pp. 962–965.
KORN, I.: 'M-ary CPFSK-DPD with L-diversity maximum ratio combining in Rician fast fading channels', *IEEE Transactions*, November 1996, **VT-45** (4), pp. 613–619.
LEE, W.C.Y.: *Mobile Communications Engineering*, Wiley, 1982.
MACARIO, R.C.V.: *Personal and Mobile Radio Systems*, IEE Telecommunications Series, 25; Peter Peregrinus Ltd, 1991.
MACHINE DESIGN: 'Integral roof antenna improves vehicle design', *Machine Design*, February 1992, p. 102.
MIKI, T. and HATA, M.: 'Performance of 16 kbit/s GMSK transmission with post-detection selection diversity in landmobile radio', *IEEE Transactions*, 1984, **33** (3), pp.128–133.
MUSHIMA, H. and TAGA, T.: 'Mobile antennas and duplexer for 800 MHz band mobile telephone system', *Proceedings of the IEEE APS Symposium*, 1980, pp. 508–511.
NAGY, L.L.: 'A new generation of antennas for automobile use', *SAE* 870092, February 1988.

NYQUIST, H.: 'Certain topics in telegraph transmission speed', *AIEE Transactions*, April 1928, **47**, p. 617.

OSCHNER, H.: 'Future personal telecommunications systems', Chapter 15, in Macario (1991).

PRICE, R. and GREEN, P.E. Jr.: 'A communication technique for multipath channels', *Proceedings of the IRE*, 1958, **46**, pp. 555–570.

PROAKIS, J.G.: *Digital Communications*, Third edition, McGraw-Hill, 1995.

ROWAN, C.W. and ALTGELT, C.A.: 'The evolution of the automobile antenna in the United States and Europe – A historical perspective, Part One: The first forty years', *SAE* 870090, February, 1987.

SAKOH, K., TSYJIMARU, K., KINOSHITA, K., and ADACHI, F.: 'Advanced radio paging service supported by ISDN', *Proceedings of the Nordic Seminar on Digital LandMobile Radiocommunication'*, Espoo, Finland, February 1985, pp. 239–248.

SHANNON, C.E.: 'A mathematical theory of communication', *Bell System Technical Journal*, July and October, 1948, **27**, pp. 379–423 and 623–656.

STEELE, R.: 'Deploying PCNs', *IEEE Communications Magazine*, September 1990.

STEELE, R.: *Mobile Radio Communications*, Pentech, 1992.

VAN LOON, L.J.W.: 'Mobile in-home UHF radio propagation for short-range devices', *IEEE Antennas and Propagation Magazine*, April 1999, **41** (2), pp. 37–40.

VAUGHAN, R.G. and ANDERSEN, J. BACH: 'A multiport patch antenna for mobile communications', *Proceedings of the 14th European Microwave Conference*, September 1984, pp. 607–612.

VITERBI, A.J.: 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE Transactions*, 1967, **IT-13**, pp. 260–269.

WINTERS, J.H., SALZ, J., and GITLIN, R.D.: 'The impact of antennas diversity on the capacity of wireless communications systems', *IEEE Transactions on Communications*, February/March/April 1994, **42** (2–4), pp. 1740–1751.

WU, M.S. and HO, K.: 'Meshed microstrip antennas constructed on a transparent substrate', *IEICE Transactions*, May 1991, **E74** (5).

1.12 Other reading

The following is a sample of relevant books on wireless and mobile communications, although this list is not exhaustive.

BERTONI, H.: *Radio Propagation for Modern Wireless Systems*, Prentice Hall PTR, 2000.

CAVERS, J.K.: *Mobile Channel Characteristics*, Kluwer, 2000.

GIBSON, J.D.: *The Mobile Communications Handbook*, CRC Press, 1996.

HESS, G.: *Land-Mobile Radio System Engineering*, Artech House, 1993.

JANASWAMY, R.: *Radiowave Propagation and Smart Antennas for Wireless Communications*, Kluwer, 2000.

LINNARTZ, J.-P.: *Narrowband Land-Mobile Radio Networks*, Artech House, 1993.

PARSONS, J.D.: *Mobile Communications Systems*, Blackie, London: Halsted Press, New York, 1989.

PARSONS, J.D.: *The Mobile Radio Propagation Channel*, Pentech Press, 1992.

PRASAD, R., MOHR, W., and KONHAUSER, W.: *Third Generation Mobile Communications Systems*, Artech House, 2000.

RAPPAPORT, T.S.: *Wireless Communications Principles and Practice*, IEEE Press, 1996.

SAUNDERS, S.R.: *Antennas and Propagation for Wireless Communications Systems*, Wiley, 1999.

SIWIAK, K.: *Radiowave Propagation and Antennas for Personal Communications*, Artech House, 1995.

STÜBER, G.L.: *Principles of Mobile Communication*, Kluwer, 1996.

TSOULOS, G. (Ed.): *Adaptive Antennas for Wireless Communications*, IEEE Press/Wiley, 2001.

WALKER, J. (Ed.): *Mobile Information Systems*, Artech House, 1990.

Chapter 2

Multipath propagation in mobile communications

When a receiver is moving in a complicated, multipath environment, it is continuously receiving new contributions with changes in delay, amplitude and polarisation. The problem of exactly analysing the multipath problem is a formidable task, but is usually one without interest. What is important is to get a statistical description of the propagation parameters combined with a knowledge of their relevance, and to get an engineering understanding of the variety of multipath phenomena by reducing the problem as much as possible. In this section, the various propagation parameters are introduced for more detailed treatment in later sections.

2.1 Multipath reception and transfer function model

2.1.1 Polarisation of multipath fields

In the general case, the transmitter, which may be either a base or mobile station, emits a field with a well-defined polarisation. After a number of reflections from slanted and rough surfaces, or scatterers, the polarisation of the field is changed in a random way, so the incident field at the receiver, $\boldsymbol{E}_{\mathbf{R}}$, has all three components and is dependent on both frequency, ω, and position, $\boldsymbol{r}$:

$$\boldsymbol{E}_{\mathbf{R}}(\omega, \boldsymbol{r}) = E_x(\omega, r)\hat{x} + E_y(\omega, \boldsymbol{r})\hat{y} + E_z(\omega, \boldsymbol{r})\hat{z}. \tag{2.1.1}$$

A time factor of $\exp(\mathrm{j}\omega t)$ is assumed. Each component, such as E_x, is a complex scalar, and is also a function of frequency and position. In general, at each point in space, the polarisation is elliptical where the plane of the ellipse is skew (see Figure 2.1.1). At the same point in space there will also be three magnetic field components, each with its own plane of polarisation,

$$\boldsymbol{H}_{\mathbf{R}}(\omega, \boldsymbol{r}) = H_x(\omega, \boldsymbol{r})\hat{x} + H_y(\omega, \boldsymbol{r})\hat{y} + H_z(\omega, \boldsymbol{r})\hat{z}. \tag{2.1.2}$$

So there is a total of six complex components in multipath fields.

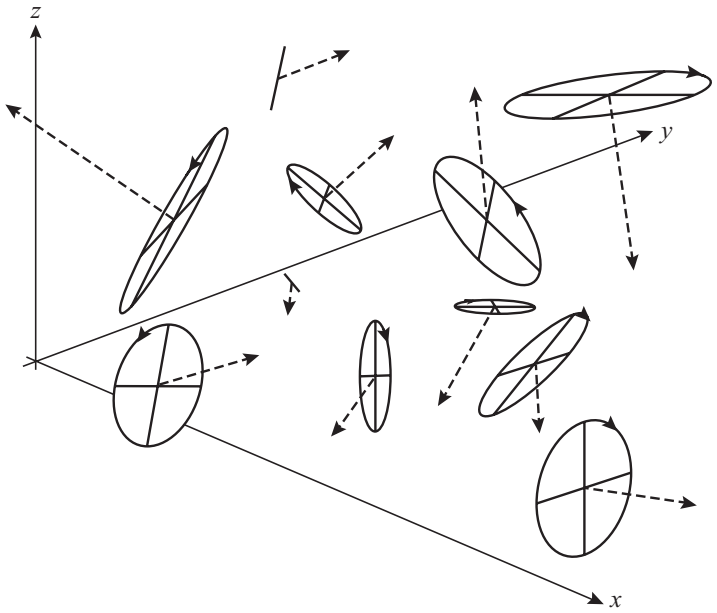


Figure 2.1.1 With multiple random incident waves, the fields at a point in space have elliptical polarisation with the plane of the ellipse having a random direction. To assist visualisation, the principal axes of the ellipses are included, and the sense of the rotation is denoted with right-hand rule arrows with length proportional to the energy density

2.1.2 Fields summed by an antenna

The voltage at the receiving antenna terminal is a function of the spatial fields and the antenna properties. A more complete description would include the magnetic field as well, but this would not add significantly to the understanding of the multipath mechanism. The receiving pattern of the antenna is the complex vector $\boldsymbol{h}(\theta, \phi)$, defined such that the open circuit voltage is given by

$$V_{\text{oc}} = \iint \boldsymbol{E}_{\mathbf{R}}(\theta, \phi) \cdot \boldsymbol{h}(\theta, \phi) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi, \tag{2.1.3}$$

where the integral is over the full angular space. The notation $\boldsymbol{h}$ stems from its units being length and its traditional interpretation of *effective* ***height*** of the receiving antenna. The pattern acts to apply a complex weight to the waves coming from different directions. From the principle of reciprocity, the transmitting pattern has the same form as the receiving pattern, but if $\boldsymbol{h}$ were the transmitting pattern instead of the receiving pattern, it would have to be conjugated in equation (2.1.3). The conjugate comes from a convention to refer the antenna polarisation to the same axis direction as the incoming wave polarisation.

2.1.2.1 Scalar model using discrete effective scatterers

In the multipath case, the total electric field in equation (2.1.3) comprises a summation of contributions from reflectors and scatterers in different directions and with different delays. Often it is advantageous to think of the field contributions in terms of rays, so the situation can be considered as multiple rays with multiple reflections. This description will be developed further in later sections. For the time being, it is sufficient to note that in communication and circuit theory, it is customary to apply a mathematical description in which the underlying physics is simplified and momentarily put aside. In such a context, the complex voltage at the antenna terminal may be viewed as a complex transfer function which has stochastic properties (see Appendix B). This function corresponds to the signal propagation channel in Figure 1.6.1. In the following, units are sometimes attributed to the transfer function quantities to promote insight, with the understanding that proper normalisation is required for formal application as a transfer function.

The antenna pattern can be viewed as a polarisation and spatial filter. Equation (2.1.1) shows explicitly the frequency dependence of the incident fields, and with the frequency dependence of the antenna pattern and matching, etc. (this is understood from equation (2.1.3)) the antenna is also seen to contribute a frequency domain filter to the channel transfer function.

2.1.2.2 Constant point scatterers for localised modelling

Over a region that typically comprises many wavelengths, the signal transfer function for a mobile terminal may be modelled as a summation over many constant, scalar, effective sources, i.e.

$$V_{\mathrm{oc}} = H(\omega, z) = \sum_i a_i \mathrm{e}^{\mathrm{j}\psi_i} \cdot \mathrm{e}^{-\mathrm{j}kr_i}. \tag{2.1.4}$$

The term *effective* is used because the signals from the physical sources (the surfaces of scatterers, etc.) are weighted by the receiving pattern. The expression $a_i \exp(\mathrm{j}\psi_i)$ is the complex amplitude of the ith effective source, and kr_i is the scalar electrical distance (k is the scalar wavenumber, r_i is the physical distance) from the point of observation to the position of the ith scatterer. In equation (2.1.3), the normalisation to keep the transfer function dimensionless is understood.

2.1.3 Moving receiver

The electrical distance to the ith point scatterer, using subscript 'R' here to denote the radio frequency quantities for the time being, can be expressed as

$$k_{\mathrm{R}} r_i(z) = \frac{2\pi}{\lambda_{\mathrm{R}}} r_i(z) = \omega_{\mathrm{R}} \frac{r_i(z)}{c} = \omega_{\mathrm{R}} \tau_i(z), \tag{2.1.5}$$

where c is the velocity of light, τ_i is the time delay from the ith scatterer, ω_{R} is the angular radio frequency and z is the position along the direction of motion. Figure 2.1.2 depicts the situation. The assumptions for the region of validity of this model are that

the magnitude and phase of each scatterer are constant and that there is no distance dependence for the magnitude of the waves from the sources. This means that slowly varying magnitudes are neglected, i.e. only the phase or delay variation with distance is kept, as indicated in equation (2.1.5), in the form of a spatially dependent delay from each scatterer. So only rapid variations of the transfer function with respect to frequency and space are dealt with at the moment, and these stem from the phase mixing of the wave contributions. This phase mixing is the reason for the fast fading of the narrowband signal envelope discussed in Chapter 1.

2.1.3.1 Phase linearisation in localised model

Assuming that the distance to the scatterers is much larger than the distance moved by the mobile receiver, the angle θ between the direction to the scatterer and the direction of movement can be defined as in Figure 2.1.2. Here θ is called the zenith angle. (Note that this definition of θ will not normally coincide with the spherical polar angle θ in an antenna pattern measurement, which is normally referenced to some physical parameter of the antenna itself.)

The electrical distance of equation (2.1.5) can be linearised now and rearranged in the following steps

$$\begin{aligned} k_{\mathrm{R}} r_i &\approx k_{\mathrm{R}} r_{0i} - k_{\mathrm{R}} z \cos(\theta_i) \\ &= \omega_{\mathrm{R}} \tau_i - \frac{\omega_{\mathrm{R}}}{c} \cos(\theta_i) z \\ &= \omega_{\mathrm{R}} \tau_i - u_i z \end{aligned} \tag{2.1.6}$$

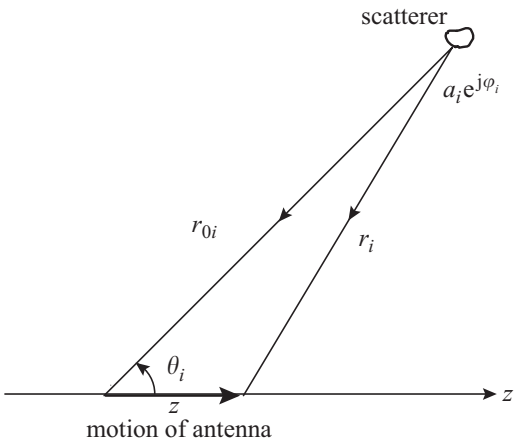


Figure 2.1.2 A single effective scatterer illuminating a receiver. The receiver position is expanded along the direction of motion to give a model for the changing phase. The phase of the signal of the ith effective scatterer is ψ_i and θ_i denotes the angle to the scatterer with respect to the direction of motion of the receiver

in which

$$u_i = k_{\mathrm{R}} \cos\theta_i \tag{2.1.7}$$

is the spatial Doppler shift in rad m^{-1}, so called because of the phase shift $u_i z$ being proportional to distance z. By introducing $z = Vt$, where V is the speed of the mobile in the z-direction, the usual Doppler shift

$$\omega_{\mathrm{D}i} = u_i V = k_{\mathrm{R}} V \cos(\theta_i) \tag{2.1.8}$$

in rad sec^{-1} is obtained. Note that the linearised electrical distance is dependent on only one polar angle, θ, so the model for each scatterer is two-dimensional (2 D).

2.1.4 Baseband equivalent transfer function from discrete scatterers

It is advantageous to distinguish between the actual radio frequency ω_{R} and the baseband frequency ω_{BB}, shifted by the carrier frequency ω_{C}, namely

$$\omega_{\mathrm{R}} = \omega_{\mathrm{BB}} + \omega_{\mathrm{C}} \tag{2.1.9}$$

and similarly for the wavenumber,

$$k_{\mathrm{R}} = k_{\mathrm{BB}} + k_{\mathrm{C}}, \tag{2.1.10}$$

so that the electrical distance to the scatterer, as a function of the receiver position z, may be written from equation (2.1.6), as

$$\begin{aligned} k_{\mathrm{R}} r_i &= \omega_{\mathrm{C}}\tau_i + \omega_{\mathrm{BB}}\tau_i - \frac{\omega_{\mathrm{C}}}{c}\cos(\theta_i)z - \frac{\omega_{\mathrm{BB}}}{c}\cos(\theta_i)z \\ &= \omega_{\mathrm{C}}\tau_i + \omega_{\mathrm{BB}}\tau_i - k_{\mathrm{C}}\cos(\theta_i)z - k_{\mathrm{BB}}\cos(\theta_i)z \\ &= \omega_{\mathrm{C}}\tau_i + \omega_{\mathrm{BB}}\tau_i - u_i z - u_{\mathrm{BB}} z. \end{aligned} \tag{2.1.11}$$

The first term on the right-hand side of the equation is constant for each scatterer and may be included in the phase of the scatterer, ψ_i. The fourth term can be neglected because it is small relative to the third term owing to the small relative bandwidth ($\omega_{\mathrm{BB}}/\omega_{\mathrm{C}}$ usually less than 1 per cent) required of the propagation channel in microwave mobile communications. This leaves the equivalent baseband transfer function of the channel of equation (2.1.4) to be approximately modelled by

$$H(\omega, z) = \sum_i a_i \, \mathrm{e}^{\mathrm{j}\psi_i} \mathrm{e}^{\mathrm{j}(-\omega\tau_i + z u_i)}, \tag{2.1.12}$$

where the subscript 'BB' in ω has been dropped on the right-hand side for convenience. From here on in this section, ω is the baseband angular frequency.

2.1.5 Fourier model using continuous scattering medium

If the scattering medium is considered to comprise or include a continuum, then the summation for all the delays and ray directions should be expressed in integral form, i.e.

$$H(\omega, z) = \frac{1}{2\pi} \int_0^{\infty} \int_{-k_C}^{k_C} a(\tau, u) \mathrm{e}^{-\mathrm{j}\omega\tau} \mathrm{e}^{\mathrm{j}zu} \, \mathrm{d}\tau \, \mathrm{d}u, \tag{2.1.13}$$

where $a(\tau, u)$ is the distribution density of signals (open circuit voltages) received by the antenna from the scattering medium at delay τ and direction u, and ω is the baseband frequency. This distribution is scalar, comprising complex amplitudes, and is referred to as the *effective scatterer distribution*. The subscript 'C' on k_C is a reminder that the wavenumber is for the radio carrier frequency, and does not correspond to the baseband frequency ω. The 2π factor is included in the integral form to formalise the two-dimensional Fourier transform between the transfer function and the effective scatterer distribution. The set of individual scatterer contributions can therefore be related to the effective scatterer distribution by

$$a(\tau, u) = 2\pi \sum_i a_i \mathrm{e}^{\mathrm{j}\psi_i} \delta(u - u_i) \delta(\tau - \tau_i). \tag{2.1.14}$$

The inverse form of equation (2.1.13) is

$$a(\tau, u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(\omega, z) \mathrm{e}^{\mathrm{j}\omega\tau} \mathrm{e}^{-\mathrm{j}zu} \, \mathrm{d}\omega \, \mathrm{d}z. \tag{2.1.15}$$

The importance of the Fourier transform pair is the insight it offers: models of the effective scatterer distribution can be used conveniently to synthesise mobile channel transfer functions, or measured transfer functions can produce an effective scatterer distribution. The relationship of the mathematical model of received signals, $a(\tau, u)$, to the physical scattering environment is discussed further below. The effective scattering distribution and its Fourier relation to the transfer function is depicted in Figure 2.1.3.

The double Fourier transform is usually defined with both exponents in the integrand having the same sign, and $1/(4\pi^2)$ with the integral with negative exponents. Owing to the physical model (namely moving the receiver in the positive z-direction), the double Fourier transform of equations (2.1.11) and (2.1.13) ends up with mixed exponents. Here, the $4\pi^2$ factor is shared in the sense that 2π is attributed to each integral equation. This definition, with a 2π factor associated to each positive exponent and angular frequency integration, ensures conventional electrical engineering definitions for the associated single Fourier transforms. Mathematically, there is no difference between the conventional definition and that used here. However, for computer usage, due care must be taken with the constants when using forward and inverse FFT algorithms for equations (2.1.12) and (2.1.14).

2.1.5.1 Summary

The approximations in developing equations (2.1.12) and (2.1.14) were that the signals have a narrow relative bandwidth, and that the scatterers are fixed and not too

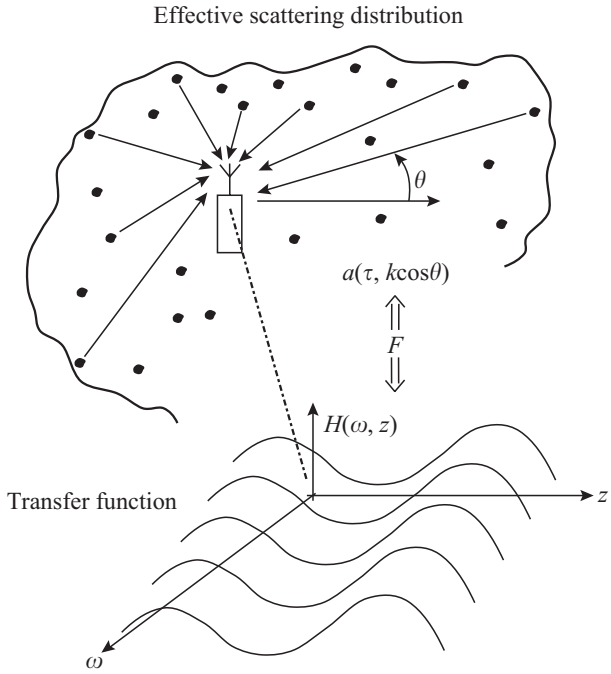


Figure 2.1.3 The effective scattering distribution $a(\tau, u)$ is the complex amplitude density of the multipath illumination of a receiving antenna which includes the effects of the receiving pattern; $u = k_C \cos\theta$ is the polar direction to the distribution and τ is the delay time of the signal. The transfer function, $H(\omega, z)$, and the effective scattering distribution are a Fourier transform pair

close to the receiving position, in which case the linearisation of the phase in equation (2.1.6) breaks down. In this sense, it is the physical geometry (in particular, the linear motion of the receiver, and the distance moved being small relative to the distance to the physical scatterers) rather than the electrical parameters (such as the frequency independence of the scatterer signal and the far-field assumption), that is the limiting aspect of the model.

2.1.6 Resolvability of scatterers

In terms of the effective scattering distribution, $a(\tau, u)$, there are the two variables with which to resolve scatterers – different delays and different angles. Discrete effective scatterers spaced by delay $\Delta\tau$ can be resolved in the time domain by using a sufficiently wide bandwidth ($B > 1/\Delta\tau$) even if they have the same polar direction. The electromagnetic propagation channel and the electromagnetic signal channel (see Figure 1.6.1) can be considered for now as having infinite bandwidth. The restrictions on bandwidth are normally imposed on the signal by the antenna and the radio electronics. The effective scatterers can thus be considered resolvable from the propagation channel viewpoint, but are quantised in the delay variable by the bandwidth restriction of the radio channel and it is here that the resolution is correspondingly restricted. Note also that the physical scatterers feature polarisation, which is not

included in the effective scatterers and so resolvablility of the physical scatterers does not include their polarisation state. This is addressed further below. Analogously to the delay resolution, effective scatterers directionally spaced by angle Δu can be resolved with a sufficient spatial antenna separation ($L > 1/\Delta u$), even if they have the same delay. Super-resolution techniques, which use signal processing to improve on the Fourier resolution, are not considered for now.

In terms of the physical scatterers, these can be resolved only if they are spaced such that they make unique contributions to the effective scattering distribution (the discrete effective scatterer comprises the single physical scatterer only). Several physical scatterers can make up a single (resolvable) effective scatterer. In this case, the physical scatterers are unresolvable. So resolvability of effective scatterers does not necessarily imply resolvability of the physical scatterers.

2.1.7 Time domain representation

The time domain equivalent (Fourier transform) of the discrete scatterers transfer function of equation (2.1.12), is the impulse response

$$h(\tau, z) = \sum_i a_i \, \mathrm{e}^{\mathrm{j}\psi_i} \delta(\tau - \tau_i) \mathrm{e}^{\mathrm{j}zu_i} \tag{2.1.16}$$

in which the discrete effective scatterers are identified directly by their time delay τ_i and two-dimensional polar direction $\theta_i = \cos^{-1}(u_i/k_\mathrm{C})$. A shift in position z of the receiver is seen to impose a new phase at each delay which is proportional to the spatial Doppler shift u_i of the scatterer. The impulse response notation $h(\tau, z)$ should not be confused with the receiving pattern $h(\theta, \phi)$ of equation (2.1.3) or the polar components $h_\theta(\theta, \phi)$ and $h_\phi(\theta, \phi)$ of the pattern (see below). An example of an impulse response is given in Figure 2.1.4.

If each Doppler shift u_i and time delay τ_i in the impulse response of equation (2.1.15) is due to an individual physical scatterer, the impulse response can be regarded as deterministic since it is available from the complex amplitudes, directions and delays of the physical scatterers, the antenna pattern (or alternatively to these, the complex amplitudes of the effective scatterers) and the mobile trajectory. An increment in the receiver position of Δz would result in a movement of the impulse response contribution along the time axis of $\Delta\tau_i = \Delta z_i/c \cos\theta_i$ and a phase rotation in a sense and at a rate found from $\exp(\mathrm{j}\Delta z u_i) = \exp(\mathrm{j}\omega\Delta\tau_i)$. In any band-limited system, i.e. the radio channel, multiple effective scatterers will be lumped together into delay bins, and the resultant components in the impulse response will fade and rotate randomly. In the narrowband channel, all the effective scatterers are lumped and flat fading results.

The time domain equivalent of the continuous form of the transfer function in equation (2.1.12) is

$$h(\tau, z) = \frac{1}{2\pi} \int_{-k_\mathrm{C}}^{k_\mathrm{C}} a(\tau, u) \mathrm{e}^{\mathrm{j}zu} \, \mathrm{d}u, \tag{2.1.17}$$

which is of course the continuous form of equation (2.1.16). Again, the 2π factor is included to formalise the Fourier relationship.

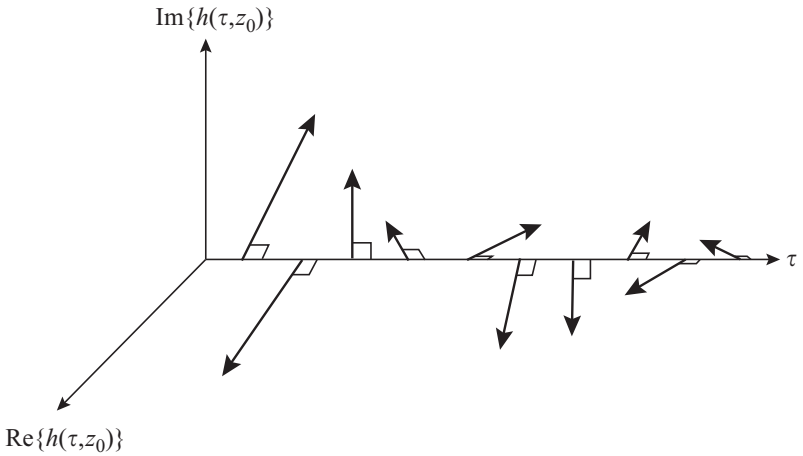


Figure 2.1.4 An impulse response of a multipath propagation channel at a fixed receiver position z_0, with several fixed discrete effective scatterer contributions. As the receiver moves, each component of the impulse response (which corresponds to a discrete effective scatterer) changes phase and delay according to the complex amplitude and position (angle and delay) of the effective scatterer

2.1.8 Polarisation and antenna pattern effect

2.1.8.1 Static receiver with 3D effective scattering distribution

The treatment above was simplified with the scalar effective scattering distribution $a(\tau, u)$ incorporating implicitly the antenna and polarisation effects. In the following, the polarisation is included explicitly in order to emphasise the role of the antenna. With a static receiver, the effective scattering distribution has three-dimensional (3D) directionality in the sense that the distribution a has dependency $a(\tau, k\cos\theta, \phi)$. However, when the moving receiver is reintroduced, the effective scattering distribution returns to having two-dimensional directionality, namely $a(\tau, k\cos\theta)$.

Each physical scatterer may produce cross-polarisation relative to its incident wave. Orthogonal polarisations thus appear, whose ratio, after multiple scattering, can become independent of the transmitted polarisation. Furthermore, each scatterer contribution to the transfer function has a complex weighting for each polarisation from the vector antenna pattern in the direction of the scatterer.

The 3D distribution of the incoming waves, visible from the antenna position, is denoted $\boldsymbol{A}(\tau, \theta, \phi)$. An incident plane wave, arriving at the receiving antenna from the ith physical scatterer, is seen from the receiver position as the contribution

$$\boldsymbol{A}_i(\tau_i, \theta_i, \phi_i) = A_{\theta i}(\tau_i, \theta_i, \phi_i)\hat{\theta} + A_{\phi i}(\tau_i, \theta_i, \phi_i)\hat{\phi}. \tag{2.1.18}$$

The subscripts refer to the polar components in the usual way. The receiving pattern of the antenna, $\boldsymbol{h}(\theta, \phi)$, as in equation (2.1.3), is similarly denoted with its components $h_\theta(\theta, \phi)$ and $h_\phi(\theta, \phi)$. The Fourier transform between the scalar effective scatterer

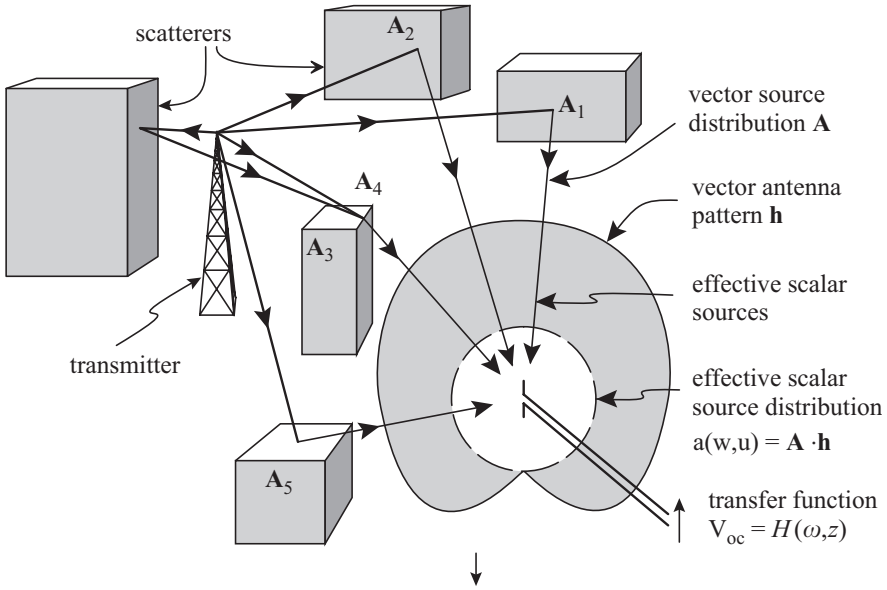


Figure 2.1.5 *The effective scatterer distribution $a(\tau, \theta, \phi)$ in a physical context; θ is the zenith polar angle with respect to the direction of motion and ϕ is the azimuthal angle. The physical scatterers re-radiate vector contributions which have different amplitude, phase, and delay in each polarisation. These waves are seen from the receiver position as the distribution $\boldsymbol{A}_i(\tau_i, \theta_i, \phi_i)$. In this example, the same physical scatterer provides the waves $\boldsymbol{A}_3(\tau_3, \theta_3, \phi_3)$ and $\boldsymbol{A}_4(\tau_4, \theta_3, \phi_3)$, i.e. these two contributions are from the same direction but are otherwise independent. The antenna receiving pattern $\boldsymbol{h}(\theta, \phi)$ weights the vector physical scatterer contributions into scalar contributions making up the three-dimensional effective scatterer distribution $a(\tau, \theta, \phi)$ which then includes the effect of the antenna pattern. The transfer function of the signal propagation channel, $H(\omega, z)$, is equivalent to the normalised open circuit antenna voltage of the antenna*

distribution and the transfer function of equation (2.1.13) becomes

$$H(\omega, z) = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\pi} \int_0^{\infty} \boldsymbol{A}(\tau, \theta, \phi) \cdot \boldsymbol{h}(\theta, \phi) \mathrm{e}^{-\mathrm{j}\omega\tau} \, \mathrm{e}^{\mathrm{j}k_C z \cos\theta} \, \mathrm{d}\tau \, k \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi, \tag{2.1.19}$$

where

$$\begin{aligned} \boldsymbol{A}(\tau, \theta, \phi) \cdot \boldsymbol{h}(\theta, \phi) &= A_\theta(\tau, \theta, \phi) h_\theta(\tau, \theta, \phi) + A_\phi(\tau, \theta, \phi) h_\phi(\theta, \phi) \\ &= a(\tau, \theta, \phi) \end{aligned} \tag{2.1.20}$$

now has the 3D directionality. Figure 2.1.5 depicts the relationship between the waves from the physical scatterers (shown as discrete scatterers) that make up the distribution $\boldsymbol{A}(\tau, \theta, \phi)$, and which, in combination with the antenna pattern, make up the three-dimensional effective scattering distribution $a(\tau, \theta, \phi)$.

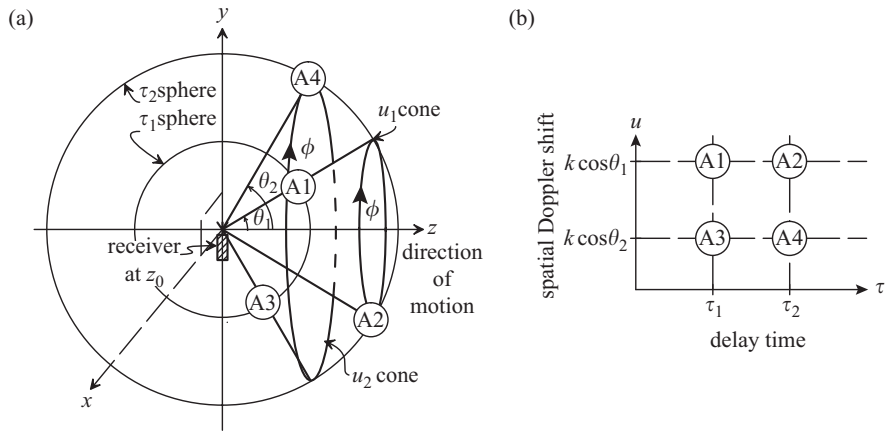


Figure 2.1.6 An example of the effective scatterer distribution $a(\tau, u)$ for a receiver position z_0, depicted in (a) with spherical coordinates where the radius is delay time (i.e. distance to the effective scatterers is proportional to their delay time) and the directions correspond to physical three-dimensional directions to the scatterers, and (b) directly mapped by its variables τ, u. Sections of constant θ form cones of constant Doppler shift. Any physical scatterers on a circular ring of constant τ on the cone (i.e. same delay) – even those in different ϕ directions – cannot be resolved by the receiver

2.1.8.2 Moving receiver with 2D effective scattering distribution

Recalling θ as the angle between the direction of movement and the direction to the scatterer, a constant θ defines a cone about the direction of movement with the vertex of the cone at the receiver as shown in Figure 2.1.6. All the scatterers on this cone will have the same Doppler shift, i.e. the antenna far-field pattern weights the different contributions having the same Doppler shift in the ϕ-plane. It follows that physical scatterers with different ϕ, but same θ and τ, cannot be resolved by a fixed-beam antenna moving in a constant direction. This is because the different physical scatterers having the same delays and same zenith angles with respect to the direction of movement (despite their azimuthal bearings being different) are unresolved, summing to a single effective scatterer. There will be no fading components in the impulse response at these delay times (a discrete effective scatterer is here considered resolved). Conversely, to resolve the physical scatterer contributions at each time delay, there must not be physical scatterers with both the same time delay and same zenith angle with respect to the direction of motion.

As a consequence of the receiver being unable to resolve physical scatterers with the same delay and zenith angle with respect to the direction of motion, the effective scatterer distribution reduces from $a(\tau, \theta, \phi)$ to $a(\tau, u)$, i.e. from three-dimensional to two-dimensional directionality. Specifically, from equation (2.1.19), the ϕ-integration must be incorporated in the $\boldsymbol{A}$ distribution leaving the θ (i.e. the u) integration

$$H(\omega, z) = \frac{1}{2\pi} \int_0^{\infty} \int_{-k_C}^{k_C} \boldsymbol{A}(\tau, u) \cdot \boldsymbol{h}(u) \mathrm{e}^{-\mathrm{j}\omega\tau} \, \mathrm{e}^{\mathrm{j}zu} \, \mathrm{d}u \, \mathrm{d}\tau. \tag{2.1.21}$$

This is the same as the scalar version of equation (2.1.12), since

$$\begin{aligned} \boldsymbol{A}(\tau, u) \cdot \boldsymbol{h}(u) &= A_\theta(\tau, u) h_\theta(u) + A_\phi(\tau, u) h_\phi(u) \\ &= a(\tau, u). \end{aligned} \tag{2.1.22}$$

Multiple physical scatterers can produce signals with the same delay and Doppler frequency. In Figure 2.1.6, such scatterers would occupy a locus of constant distance from the receiver, i.e. a ring centred on the distance axis and in the plane orthogonal to the axis, i.e. all at (u_i, τ_i). Such physical scatterers cannot be resolved by a fixed-beam receiver, and the corresponding single effective scatterer at (u_i, τ_i) comprises a summation of the unresolvable physical scatterers. Because there is a resolvable effective scatterer, the impulse response is again deterministic, but in this case, in terms of knowledge of $a(u, \tau)$, as mentioned above in connection with the impulse response. Note that Figure 2.1.6 gives an interpretation of two pairs of unresolvable physical scatterers. Models for the physical scenario of scatterers around mobile terminals are discussed in a later section.

2.1.8.3 Summary

The vector distribution $\boldsymbol{A}(\tau, \theta, \phi)$ comprises the distribution of carrier waves from the physical scatterers in a scenario. The scalar distribution $a(\tau, \theta, \phi)$ produces the open circuit antenna voltage and is referred to here as the *effective* distribution of scatterers because the component signals include the effect of the vector antenna pattern $\boldsymbol{h}(\theta, \phi)$. When a terminal with a fixed pattern antenna moves linearly within a scenario of fixed physical scatterers, the azimuthal (with respect to direction of motion) dependence of the scatterers drops out, resulting in $\boldsymbol{A}(\tau, u)$ and $a(\tau, u)$. The propagation phenomenon that produces $\boldsymbol{A}$ and the antenna phenomenon that includes $\boldsymbol{h}$ are intimately interwoven to produce the transfer function of the propagation signal channel. Equation (2.1.21) is quite general, and it will be apparent that almost all subjects treated below have their background in this equation.

2.2 Statistical basis of transfer functions: correlation and spectra

At this stage the statistical background for channel description is introduced, since the scattering environment has a stochastic character. For a basic primer of the statistics see Appendix B. The assumption of wide-sense stationarity and uncorrelated scattering is often made in mobile communications, more because it allows elegant Fourier theory for the channel description rather than being justified by exhaustive experimental results. This section develops the Bello relations as applied to the mobile channel.

In Section 2.1, a deterministic relationship was introduced in the form of a two-dimensional Fourier transform between the transfer function $H(\omega, z)$ and the effective scattering distribution $a(\tau, u)$, denoted here by $H(\omega, z) \Leftarrow F \Rightarrow a(\tau, u)$, namely

$$H(\omega, z) = \frac{1}{2\pi} \int_{-k_C}^{k_C} \int_0^\infty a(\tau, u) \mathrm{e}^{\mathrm{j}(-\omega\tau + zu)} \, \mathrm{d}\tau \, \mathrm{d}u. \tag{2.2.1}$$

The variables are recalled here as the delay time, τ, and the spatial Doppler shift, u. The spectral variables are: ω, the baseband angular frequency; and z, the spatial variable (also referred to as position or distance) which, for static scatterers, can be replaced with the time–speed product, $z = Vt$. By transforming just one correlation variable, instead of both at once, two more spectral functions arise. Thus there is a total of four complex functions carrying the same information for characterisation of the channel. These are denoted:

- $a(\tau, u)$ the scattering function in the time delay–Doppler domain (referred to as the effective scattering distribution or effective scatterers);
- $h(\tau, z)$ the impulse response in the delay–space domain (spatial spectrum);
- $A(\omega, u)$ the transfer function in the baseband frequency–Doppler domain (frequency spectrum); and
- $H(\omega, z)$ the transfer function in the baseband frequency–space domain (space–frequency spectrum).

The $A(\omega, u)$ should not be confused with the incident vector wave distribution $\boldsymbol{A}(\tau, \theta, \phi)$. The functions are related by the following single-dimensional Fourier transforms:

$$a(\tau, u) = \frac{1}{2\pi} \int A(\omega, u) \mathrm{e}^{\mathrm{j}\omega\tau} \, \mathrm{d}\omega, \quad A(\omega, u) = \int a(\tau, u) \mathrm{e}^{-\mathrm{j}\omega\tau} \, \mathrm{d}\tau; \tag{2.2.2}$$

$$h(\tau, z) = \frac{1}{2\pi} \int H(\omega, z) \mathrm{e}^{\mathrm{j}\omega\tau} \, \mathrm{d}\omega, \quad H(\omega, z) = \int h(\tau, z) \mathrm{e}^{-\mathrm{j}\omega\tau} \, \mathrm{d}\tau; \tag{2.2.3}$$

$$a(\tau, u) = \int h(\tau, z) \mathrm{e}^{-\mathrm{j}zu} \, \mathrm{d}z, \quad h(\tau, z) = \frac{1}{2\pi} \int a(\tau, u) \mathrm{e}^{\mathrm{j}zu} \, \mathrm{d}u; \tag{2.2.4}$$

$$A(\omega, u) = \int H(\omega, z) \mathrm{e}^{-\mathrm{j}zu} \, \mathrm{d}z, \quad H(\omega, z) = \frac{1}{2\pi} \int A(\omega, u) \mathrm{e}^{\mathrm{j}zu} \, \mathrm{d}u. \tag{2.2.5}$$

2.2.1 Channel correlation functions

The correlation functions of the channel are required when we depart from a deterministic situation. Assuming ergodicity so that the statistics remain second order, the autocorrelation of the effective scatterer distribution with respect to the delay times (but constant direction), is written

$$\begin{aligned} R_{\mathrm{a}}(\tau_1, \tau_2; u) &= \left\langle a(\tau_1, u) \, a^*(\tau_2, u) \right\rangle \\ &= \left\langle \left(\frac{1}{2\pi} \int A(\omega_1, u) \mathrm{e}^{\mathrm{j}\omega_1\tau_1} \, \mathrm{d}\omega_1 \right) \left(\frac{1}{2\pi} \int A(\omega_2, u) \mathrm{e}^{\mathrm{j}\omega_2\tau_2} \, \mathrm{d}\omega_2 \right)^* \right\rangle \\ &= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} S_{\mathrm{A}}(\omega_1, \omega_2; u) \mathrm{e}^{\mathrm{j}(\omega_1\tau_1 - \omega_2\tau_2)} \, \mathrm{d}\omega_1 \, \mathrm{d}\omega_2, \end{aligned} \tag{2.2.6}$$

where

$$S_{\mathrm{A}}(\omega_1, \omega_2; u_1, u_1) = \left\langle A(\omega_1, u_1) \, A^*(\omega_2, u_2) \right\rangle \tag{2.2.7}$$

is recognised as the Fourier transform[1] of the autocorrelation of the transfer function in the frequency–Doppler domain, i.e. the power spectrum, S_{A}. The angle brackets indicate the expectation or averaging (see below). The inverse relation is

$$S_{\mathrm{A}}(\omega_1, \omega_2; u) = \int_0^\infty \int_0^\infty R_{\mathrm{a}}(\tau_1, \tau_2; u) \mathrm{e}^{-\mathrm{j}(\omega_1\tau_1 - \omega_2\tau_2)} \, \mathrm{d}\tau_1 \, \mathrm{d}\tau_2. \tag{2.2.8}$$

The terminology here is for R to represent a correlation function, and S to represent a power spectrum, but either of the functions can be called a spectrum, so the R and S notation can be interchanged where it assists clarity. The subscript denotes the function of the correlation. Below, the symbol P is similarly used for averaged distributions.

2.2.1.1 Wide-sense stationarity and uncorrelated scattering

If the process describing the channels is now assumed to be *wide-sense stationary in the frequency domain*, i.e. the mean and correlation do not depend on the choice of frequency – which seems reasonable in a multipath environment – then the manipulations that show the process is uncorrelated in the delay domain are as follows. Denoting the frequency difference $\Delta\omega = \omega_2 - \omega_1$,

$$R_{\mathrm{a}}(\tau_1, \tau_2; u) = \frac{1}{4\pi^2} \iint S_{\mathrm{A}}(\Delta\omega; u) \mathrm{e}^{-\mathrm{j}\Delta\omega\tau_2} \mathrm{e}^{-\mathrm{j}\omega_1(\tau_2 - \tau_1)} \, \mathrm{d}\omega_1 \, \mathrm{d}\Delta\omega \tag{2.2.9}$$

and applying the delta function, we get

$$\begin{aligned} R_{\mathrm{a}}(\tau_1, \tau_2; u) &= \frac{1}{2\pi} \int S_{\mathrm{A}}(\Delta\omega, u) \mathrm{e}^{-\mathrm{j}\Delta\omega\tau_2} \, \mathrm{d}\Delta\omega \cdot \delta(\tau_2 - \tau_1) \\ &= \frac{1}{2\pi} \int S_{\mathrm{A}}(\Delta\omega, u) \mathrm{e}^{\mathrm{j}\Delta\omega\tau_2} \, \mathrm{d}\Delta\omega \cdot \delta(\tau_2 - \tau_1) \\ &= P(\tau_2, u)\delta(\tau_2 - \tau_1), \end{aligned} \tag{2.2.10}$$

where

$$P(\tau, u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_{\mathrm{A}}(\Delta\omega, u) \mathrm{e}^{\mathrm{j}\Delta\omega\tau} \, \mathrm{d}\Delta\omega \tag{2.2.11}$$

is the power delay profile–spatial Doppler frequency distribution. This uncorrelated nature of the power is referred to as *uncorrelated scattering*. Note from equation (2.2.9) that the WSS process in the frequency domain transforms to a process that is non-stationary in the delay variable (the $P(\tau_2, u)$ term). However, the classical result is the connection between wide-sense stationarity (via the $\Delta\omega$ term) in the baseband frequency domain and uncorrelated scattering in the delay time domain (the $\delta(\Delta\tau)$ term) – one implies the other.

[1] As in Section 2.1, the double Fourier transform is defined here with mixed exponents and the $4\pi^2$ factor is assigned to one of the integrals.

Wide-sense stationarity in the space domain corresponds similarly to uncorrelated scattering in the Doppler domain as follows. Denoting the spatial difference $\Delta z = z_2 - z_1$,

$$\begin{aligned} R_{\mathrm{a}}(\tau; u_1, u_2) &= \iint S_h(\tau; \Delta z) \mathrm{e}^{\mathrm{j}(z_2 u_2 - z_1 u_1)} \, \mathrm{d}z_1 \, \mathrm{d}z_2 \\ &= \iint S_h(\tau; \Delta z) \mathrm{e}^{\mathrm{j}(\Delta z u_2 + z_1 (u_2 - u_1))} \, \mathrm{d}z_1 \, \mathrm{d}\Delta z \\ &= \int S_h(\tau; \Delta z) \mathrm{e}^{\mathrm{j}\Delta z u_2} \, \mathrm{d}\Delta z \, 2\pi\delta(u_2 - u_1) \\ &= P(\tau, u_2) \, 2\pi\delta(u_2 - u_1), \end{aligned} \tag{2.2.12}$$

where

$$\begin{aligned} P(\tau, u) &= \int S_h(\tau, \Delta z) \mathrm{e}^{\mathrm{j}\Delta z u} \, \mathrm{d}\Delta z \\ &= \int S_h(\tau, \Delta z) \mathrm{e}^{-\mathrm{j}\Delta z u} \, \mathrm{d}\Delta z. \end{aligned} \tag{2.2.13}$$

Combining the properties of the wide-sense stationarity conditions into one equation,

$$\begin{aligned} R_{\mathrm{a}}(\tau_1, \tau_2; u_1, u_2) &= \frac{1}{2\pi} \iint S_H(\Delta\omega, \Delta z) \, \mathrm{e}^{\mathrm{j}(-\Delta\omega\tau_2 + \Delta z u_2)} \\ &\qquad \times \mathrm{d}\Delta\omega \, \mathrm{d}\Delta z \, \delta(\tau_2 - \tau_1) \, 2\pi\delta(u_2 - u_1) \\ &= P(\tau_2, u_2)\delta(\tau_2 - \tau_1) \, 2\pi\delta(u_2 - u_1), \end{aligned} \tag{2.2.14}$$

where

$$P(\tau, u) = \frac{1}{2\pi} \iint S_H(\Delta\omega, \Delta z) \mathrm{e}^{-\mathrm{j}(\Delta\omega\tau - \Delta z u)} \, \mathrm{d}\Delta\omega \, d\Delta z, \tag{2.2.15}$$

and the inverse is

$$S_H(\Delta\omega, \Delta z) = \frac{1}{2\pi} \iint P(\tau, u) \mathrm{e}^{\mathrm{j}(\Delta\omega\tau - \Delta z u)} \, \mathrm{d}\tau \, \mathrm{d}u, \tag{2.2.16}$$

or simply $S_H(\Delta\omega, \Delta z) \Leftarrow F \Rightarrow P(\tau, u)$. Thus the above wide-sense stationarity conditions result in the frequency–space correlation function being the Fourier transform of the average power density in time delay and Doppler frequency. The above Fourier relations between the effective source distribution and the channel functions are summarised in Table 2.1 and Figure 2.2.1.

Bello (1963) coined the term WSSUS to describe such a channel. The WSS here refers to the delta function in Doppler frequency (WSS in position, or time for constant velocity of the mobile terminal) and the US refers to the delta function in delay time (sources at different delays are mutually uncorrelated). Bello's work related

Table 2.1 Relationships between the autocorrelation of the effective scattering distribution and the channel spectral functions, under wide-sense stationary conditions; τ is the delay time, ω is the baseband frequency, $u = k_C \cos\theta$ is the spatial Doppler frequency and z is distance

Condition	Autocorrelation of effective source distribution $R_a(\tau_1, \tau_2; u_1, u_2) = E\{a(\tau_1, u_1)a^*(\tau_2, u_2)\}$	Fourier relation between correlation functions of effective source distribution and channel functions
WSS in baseband frequency, ω <=> uncorrelated in delay time, τ	$R_a(\tau_1, \tau_2; u) = P(\tau_2, u)\delta(\tau_2 - \tau_1)$ (US)	$P(\tau, u) \Leftarrow F \Rightarrow S_A(\Delta\omega, u)$
WSS in position, z <=> uncorrelated in Doppler frequency, u	$R_a(\tau; u_1, u_2) = P(\tau, u_2)\delta(u_2 - u_1)$ (WSS)	$P(\tau, u) \Leftarrow F \Rightarrow S_h(\tau, \Delta z)$
WSS in position and baseband frequency <=> uncorrelated in Doppler frequency and delay time	$R_a(\tau_1, \tau_2; u_1, u_2) = P(\tau_2, u_2)\delta(\tau_2 - \tau_1)\, 2\pi\delta(u_2 - u_1)$ (WSSUS)	$P(\tau, u) \Leftarrow F \Rightarrow R_H(\Delta\omega, \Delta z)$
WSS in position and baseband frequency and static receiver (zero position shift)	$2\pi P(\tau) = \int P(\tau, u)\mathrm{d}u$	$P(\tau) \Leftarrow F \Rightarrow R_H(\Delta\omega)$

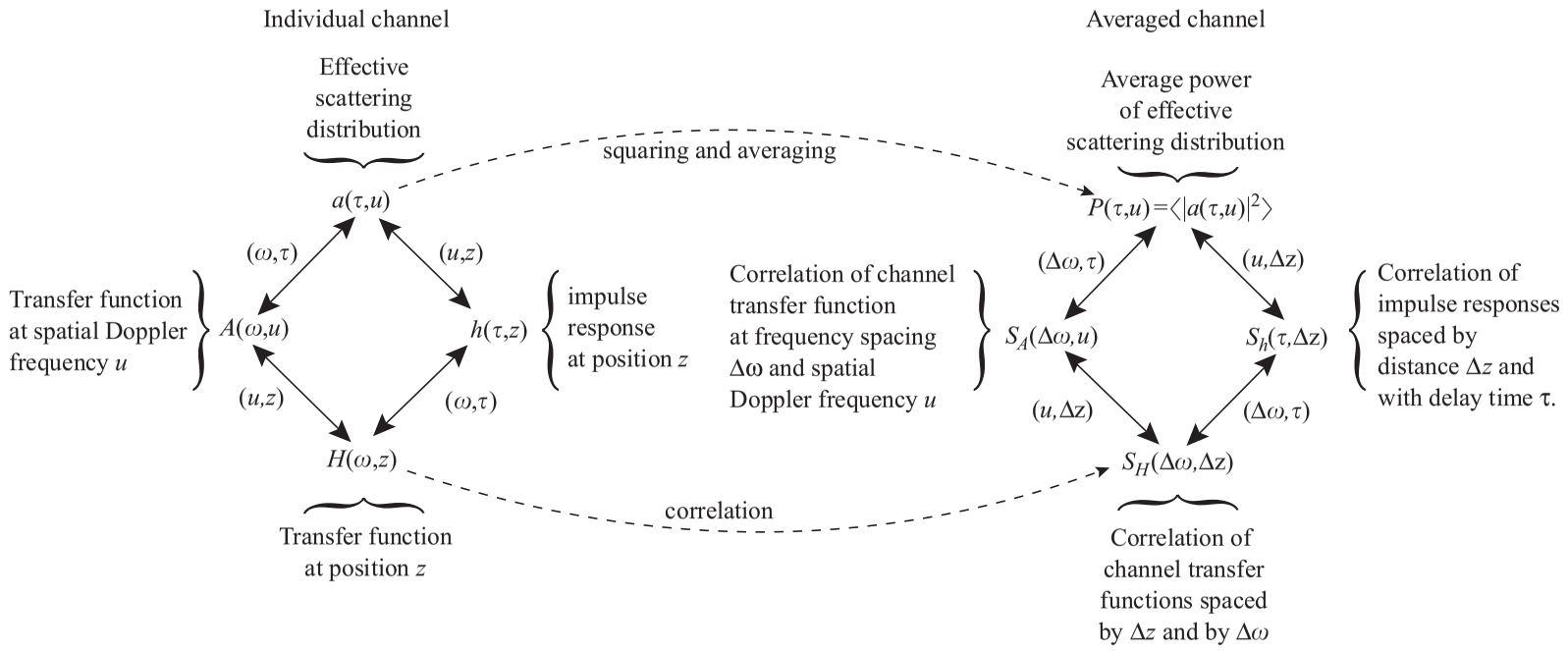


Figure 2.2.1 *Fourier transform relations between the mobile (signal propagation) channel and the effective scatterer distribution, and for their statistical functions under wide sense stationarity in frequency and position. $u = k_C \cos\theta$ is the spatial Doppler frequency with θ the zenith angle to the scatterer with respect to the direction of motion z, k_C is the wavenumber of the radio carrier frequency, ω is the baseband angular frequency and τ is the delay time. The relations are valid only for linear trajectories which are small relative to the distance to the physical scatterers*

to tropospheric channels which contained scintillating scatterers being illuminated by static terminals. For mobile channels, at least one of the terminals is moving and the scatterers are essentially static, and the validity of the WSSUS model should always be questioned for a given situation.

2.2.1.2 Averaging

The assumption of the WSSUS channel allows the convenience of the double Fourier transform relations. The channel can be arranged to be 'sufficiently valid' for gaining useful insight and inferring channel behaviour, by arranging the ensemble averaging. This is often taken as several sampled records over short distances in order to stay within a given physical environment, followed by the averaging. Statistically, the ensemble averaging implies many 'realisations'. We can interpret this as several sampled records which should have different (i.e. independent) spatial paths within the same physical environment, or else as several records in different (independent) physical environments. The two cases are clearly different. One averages within a single environment, the other averages over many different environments. A conventional notation would be $E\{\cdot\}$ for averaging over an ensemble of environments and $\langle\cdot\rangle$ for averaging over data records in a given environment. Here we adhere to the angle bracket form where possible. Strictly speaking, multiple uncorrelated records taken from a common environment do not give truly independent records because the physical scatterers are common to all records. With Gaussian signals, however, uncorrelated implies (statistical) independence.

2.2.1.3 Summary

Equations (2.2.1) and (2.2.15) are key results. They respectively relate, by Fourier transform, the baseband channel transfer function, in multipath conditions, to the effective sources which provide the incident multipath signals. In comparing these key equations there are important differences. In equation (2.2.1), $a(\tau, u)$ is the complex amplitude of a snapshot of effective scatterers; whereas in equation (2.2.15), $P(\tau, u)$ is an averaged power distribution of the effective scatterers. In equation (2.2.1), $H(\omega, z)$ is the transfer function of baseband frequency and position; whereas in equation (2.2.15) $S_H(\Delta\omega, \Delta z)$ is a spectral density and the arguments are shifts in baseband frequency and position. The interrelations are described in more detail below and depicted in Figure 2.2.1.

2.2.2 Power profiles

The total power of the averaged effective scatterers is given by

$$P = \frac{1}{2\pi} \iint P(\tau, u) \mathrm{d}\tau \, \mathrm{d}u. \tag{2.2.17}$$

Similarly, the *average power delay profile* $P(\tau)$ is averaged over all the Doppler frequencies,

$$P(\tau) = \frac{1}{2\pi} \int P(\tau, u) \mathrm{d}u \tag{2.2.18}$$

and the *average Doppler profile* is averaged over all delays

$$P(u) = \int P(\tau, u)\mathrm{d}\tau. \tag{2.2.19}$$

$P(\tau)$ and $P(u)$ are sometimes called the *delay spectrum* and the *Doppler spectrum* respectively.

In measurements of channels, the averaging for these profiles either occurs implicitly or may be arranged in a different way to suit the specific practical situation. The mathematical expressions are necessary as a benchmark for comparing different measurement techniques and results.

2.2.2.1 Fixed position: impulse response correlation

For a static receiver, there is a zero spatial correlation displacement, i.e. $\Delta z = 0$, and from equation (2.2.13), the delay–time correlation function reduces to the average power delay profile

$$\begin{aligned} S_h(\tau;\, \Delta z = 0) &= \frac{1}{2\pi} \int P(\tau, u)\mathrm{d}u \\ &= P(\tau). \end{aligned} \tag{2.2.20}$$

From the correlation definition

$$S_h\left(\tau = \tau_2 - \tau_1,\, \Delta z = 0\right) = \left\langle h\left(\tau_1, z\right) h^*\left(\tau_2, z\right)\right\rangle, \tag{2.2.21}$$

the averaging can be interpreted that the power delay power profile, $P(\tau)$, is the power of the impulse response averaged over space, sometimes denoted explicitly by $P(\tau) = \langle |h(\tau, z)|^2 \rangle_z$. This corresponds directly to mobile channel characterisation by measuring impulse responses at different locations and averaging them in subsequent processing.

2.2.2.2 Fixed position: transfer function correlation

Similarly, for the static receiver, the two-dimensional frequency correlation function reduces:

$$\begin{aligned} S_H\left(\Delta\omega\right) = S_H(\Delta\omega, \Delta z = 0) &= \left\langle H\left(\omega, z\right)\, H^*\left(\omega + \Delta\omega, z + 0\right)\right\rangle \\ &= \frac{1}{2\pi} \iint P(\tau, u)\mathrm{e}^{\mathrm{j}\Delta\omega\tau}\, \mathrm{d}\tau\, \mathrm{d}u \\ &= \int P(\tau)\mathrm{e}^{-\mathrm{j}\Delta\omega\tau}\, \mathrm{d}\tau. \end{aligned} \tag{2.2.22}$$

The inverse is

$$P(\tau) = \frac{1}{2\pi} \int S_H(\Delta\omega)\mathrm{e}^{\mathrm{j}\Delta\omega\tau}\, \mathrm{d}\Delta\omega, \tag{2.2.23}$$

which is recalled as the average power at delay τ, found by integrating over all the spatial Doppler frequencies, u. This is also a familiar-looking and important result: the

frequency correlation function $S_H(\Delta\omega)$ is the Fourier transform of the average power delay profile $P(\tau)$ for the wide-sense stationary channel with uncorrelated scattering. This result is particularly useful for simulations in which power delay profiles are required for given transfer functions with given correlations.

2.2.2.3 Fixed direction

Fixing the spatial Doppler frequency in the effective source distribution, as with an ideal, single-direction antenna, results in

$$S_A(\Delta\omega, u) = \int P(\tau, u)\mathrm{e}^{-\mathrm{j}\Delta\omega\tau}\,\mathrm{d}\tau, \tag{2.2.24}$$

or

$$P(\tau, u) = \frac{1}{2\pi}\int S_\mathrm{A}(\Delta\omega, u)\mathrm{e}^{\mathrm{j}\Delta\omega\tau}\,\mathrm{d}\Delta\omega; \tag{2.2.25}$$

i.e. the correlation of the transfer function in the frequency–Doppler domain is a single Fourier transform of the scattering function. As above, from the correlation definition, the correlation can be written as

$$\begin{aligned} S_\mathrm{A}(\Delta\omega = 0, u) &= \left\langle A(\omega, u)A^*(\omega, u)\right\rangle \\ &= P(u), \end{aligned} \tag{2.2.26}$$

the power Doppler profile averaged over frequency. Figure 2.2.1 summarises the Fourier relations of the correlation functions discussed above.

Many extensions to situations outside of mobile communications may be found in the classic paper by Bello (1963).

2.2.2.4 Spreads

The average power delay profile is the basis for the definition of the *average delay spread*, usually denoted s, which is the standard deviation of the profile,

$$s = \sigma_\tau = \sqrt{\frac{\int \tau^2 P(\tau)\mathrm{d}\tau}{\int P(\tau)\mathrm{d}\tau} - \left(\frac{\int \tau P(\tau)\mathrm{d}\tau}{\int P(\tau)\mathrm{d}\tau}\right)^2}. \tag{2.2.27}$$

Similarly the *average Doppler spread*, often denoted s_u, is defined by the standard deviation of $P(u)$. The spread can be expressed as a local quantity, i.e. a function of another variable, for example,

$$s_u(\tau) = \sigma_u(\tau) = \sqrt{\frac{\int u^2 P(\tau, u)\mathrm{d}u}{\int P(\tau, u)\mathrm{d}u} - \left(\frac{\int u P(\tau, u)\mathrm{d}u}{\int P(\tau, u)\mathrm{d}u}\right)^2}, \tag{2.2.28}$$

which is the Doppler spread at a given delay time.

The spreads are reintroduced in later chapters for further discussion. In this statistical introduction, it is important to note that it is the averaged power profiles, and not individual impulse responses, that are used to produce the spreads. In the terminology used in mobile communications, the spread is often denoted by s. Some canonical profile function examples are given in Table 2.2.

Table 2.2 Simple common examples of the Fourier pairs of power delay profile and frequency correlation function used in simulations; $\tau_0 = \sigma_\tau$ is the standard deviation of the average power delay profile $P(\tau)$, i.e. the delay spread

Power delay profile	$P(\tau)$	$S_H(\Delta\omega)$
One-sided exponential	$(1/\tau_0)\,\mathrm{e}^{-\tau/\tau_0}, \quad \tau \geq 0$	$1/(1 + \mathrm{j}2\pi\Delta\omega\tau_0)$
Gaussian	$(1/\sqrt{2\pi})\,\mathrm{e}^{-(1/2)(\tau/\tau_0)^2}$	$\mathrm{e}^{-2(\pi\Delta\omega\tau_0)^2}$
Double-spike	$\frac{1}{2}(\delta(\tau) + \delta(\tau - 2\tau_0))$	$\frac{1}{2}(1 + \mathrm{e}^{-\mathrm{j}4\pi\Delta\omega\tau_0})$

2.2.3 Averaged transfer function of channel: channel gain

In terms of a given channel transfer function, the total power, or *channel gain*, is given by

$$|H|^2 = \frac{1}{L\omega_\mathrm{B}} \int_L \int_{\omega_\mathrm{B}} |H(\omega, z)|^2 \, \mathrm{d}\omega \, \mathrm{d}z, \tag{2.2.29}$$

where L is an averaging distance and ω_B is an averaging bandwidth. Integrating single variables gives the frequency power transfer function averaged over position,

$$|H(\omega)|^2 = \frac{1}{L} \int_L |H(\omega, z)|^2 \, \mathrm{d}z, \tag{2.2.30}$$

and the position- (i.e. time-) dependent power transfer function averaged over the frequency band,

$$|H(z)|^2 = \frac{1}{\omega_\mathrm{B}} \int_{\omega_\mathrm{B}} |H(\omega, z)|^2 \, \mathrm{d}\omega. \tag{2.2.31}$$

This quantity is often approximated by the RSSI signal in a receiver (cf. Figure 1.4.8).

The gains can be averaged in an ensemble sense as well, giving the *average channel gain*.

2.2.4 Envelope, complex signal, and power signal correlations

A correlation function such as $R_H(\Delta\omega, \Delta z)$ is a complex quantity. In mobile communications the interest is often in the correlations of the envelopes, which are real since the envelopes, derivable from an RSSI signal for example, are real; e.g. for an ergodic signal representing the mobile channel,

$$R_{|H|}(\Delta a) = \mathrm{Lim}_{L\to\infty} \frac{1}{L} \int_0^L |H(z)||H(z + \Delta z)| \, \mathrm{d}z. \tag{2.2.32}$$

This section gives a statistical account of the similarity between envelope correlations and the power signal (envelope squared) and the complex Gaussian transfer function signal correlations.

2.2.4.1 Narrowband Gaussian process and power correlation

For a narrowband process, a complex signal has an envelope, which is meaningful in the sense that it behaves as intuitively expected of an envelope. Wideband signals can be treated the same mathematically, but the envelope is not readily interpreted. Here, the complex Gaussian signal is denoted z, the envelope r;

$$z(t) = x(t) + \mathrm{j}y(t) = r(t)\mathrm{e}^{\mathrm{j}\theta(t)}\,\mathrm{e}^{\mathrm{j}\omega t}; \tag{2.2.33}$$

$$r(t)\mathrm{e}^{\mathrm{j}\theta(t)} = I(t) + \mathrm{j}Q(t); \tag{2.2.34}$$

$$x(t) = I(t)\cos\omega t - \mathrm{j}Q(t)\sin\omega t; \tag{2.2.35}$$

$$y(t) = I(t)\sin\omega t + \mathrm{j}Q(t)\cos\omega t; \tag{2.2.36}$$

where $I(t)$ and $Q(t)$ are the baseband in-phase and quadrature signals. If the process is Gaussian from a 'many-term' summation of scatterer contributions,

$$r(t)\mathrm{e}^{\mathrm{j}\theta(t)} = \sum a_i(t)\mathrm{e}^{\mathrm{j}\theta_i(t)} \tag{2.2.37}$$

i.e.

$$I(t) = \sum a_i \cos\theta_i, \tag{2.2.38}$$

$$Q(t) = \sum a_i \sin\theta_i, \tag{2.2.39}$$

and if I and Q (the time dependence is dropped from now on) are zero mean, equal variance and uncorrelated, then the ergodic averages are

$$\langle I\rangle = \langle Q\rangle = 0, \tag{2.2.40}$$

$$\langle I^2\rangle = \langle Q^2\rangle = \sigma^2, \tag{2.2.41}$$

$$\langle IQ\rangle = 0. \tag{2.2.42}$$

For two variables

$$z_1(t) = x_1(t) + \mathrm{j}y_1(t), \qquad z_2(t) = x_2(t) + \mathrm{j}y_2(t) \tag{2.2.43}$$

the required component correlations and correlation coefficients, ρ, are

$$\langle I_1 I_2\rangle = \langle Q_1 Q_2\rangle = R_{II}(\tau) = \sigma^2\rho_{II}(\tau), \tag{2.2.44}$$

$$\langle I_1 Q_2\rangle = -\langle I_2 Q_1\rangle = R_{IQ}(\tau) = \sigma^2\rho_{IQ}(\tau), \tag{2.2.45}$$

and from the general result for zero mean Gaussians, a, b, c, d:

$$\langle a\,b\,c\,d\rangle = \langle a\,b\rangle\cdot\langle c\,d\rangle + \langle a\,c\rangle\cdot\langle b\,d\rangle + \langle a\,d\rangle\cdot\langle b\,c\rangle, \tag{2.2.46}$$

there results

$$\langle (I_1 I_2)^2 \rangle = \langle I_1^2 \rangle^2 + 2\langle I_1 I_2 \rangle^2. \tag{2.2.47}$$

The correlation function between two complex Gaussian variables is

$$R_z = \langle z_1 z_2^* \rangle = \langle (x_1 + \mathrm{j}y_1)(x_2 - \mathrm{j}y_2) \rangle, \tag{2.2.48}$$

which is readily found as

$$\langle z_1 z_2^* \rangle = 2\langle I_1 I_2 \rangle - \mathrm{j}2\langle I_1 Q_2 \rangle \tag{2.2.49}$$

and note that $\langle z_1 z_2 \rangle = 0$.

The correlation coefficient is

$$\rho_z(\tau) = \frac{\langle z_1 z_2^*(\tau = \tau) \rangle - \langle z_1 \rangle \langle z_2^* \rangle}{\langle z_1 z_2^*(\tau = 0) \rangle - \langle z_1 \rangle \langle z_2^* \rangle} \tag{2.2.50}$$

which reduces to

$$\rho_z = \rho_{II} - \mathrm{j}\rho_{IQ}, \tag{2.2.51}$$

i.e.

$$|\rho_z|^2 = \rho_{II}^2 + \rho_{IQ}^2. \tag{2.2.52}$$

The power correlation is

$$\begin{aligned} R_{r^2} &= \left\langle r_1^2 r_2^2 \right\rangle = \left\langle \left(I_1^2 + Q_1^2 \right) \left(I_2^2 + Q_2^2 \right) \right\rangle \\ &= 4\sigma^4 \left(1 + \rho_{II}^2 + \rho_{IQ}^2 \right), \end{aligned} \tag{2.2.53}$$

and from

$$R_{r^2}(\tau) = \left\langle r_1^2 \right\rangle + \left(R_{r^2}(0) - \left\langle r_1^2 \right\rangle^2 \right) \rho_{r^2}(\tau) \tag{2.2.54}$$

where

$$R_{r^2}(0) = \left\langle r_1^4 \right\rangle = 2\left\langle r_1^2 \right\rangle^2, \tag{2.2.55}$$

we get

$$\rho_{r^2} = \rho_{II}^2 + \rho_{IQ}^2. \tag{2.2.56}$$

Thus the power correlation coefficient is the magnitude squared complex correlation coefficient,

$$\rho_{r^2} = |\rho_z|^2. \tag{2.2.57}$$

For signal simulations, ρ_{II} and ρ_{IQ} can be chosen to provide any required power or envelope correlation between sets of signals: see Chapter 7.

2.2.4.2 Similarity of envelope and power correlations

Pierce and Stein (1960) give the envelope correlation coefficient as a function of the magnitude square of the complex Gaussian correlation coefficient (same as the power signal correlation coefficient). From their tabular data the approximation has a maximum error of 0.026 in the correlation coefficient, and a maximum relative error of less than 10 per cent. The envelope correlation coefficient can therefore be well approximated,

$$\rho_r \approx |\rho_z|^2 . \tag{2.2.58}$$

In fact the exact relation is given by Middleton (1960) as the correlation function of the envelopes,

$$R_r = \frac{\pi}{2} b_0 \, {}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; 1; \rho^2\right) \tag{2.2.59}$$

where b_0 is a moment (namely the variance of each of the Gaussian components, σ^2) and ${}_2F_1$ is the Gaussian hypergeometric function. This relation allows the correlation

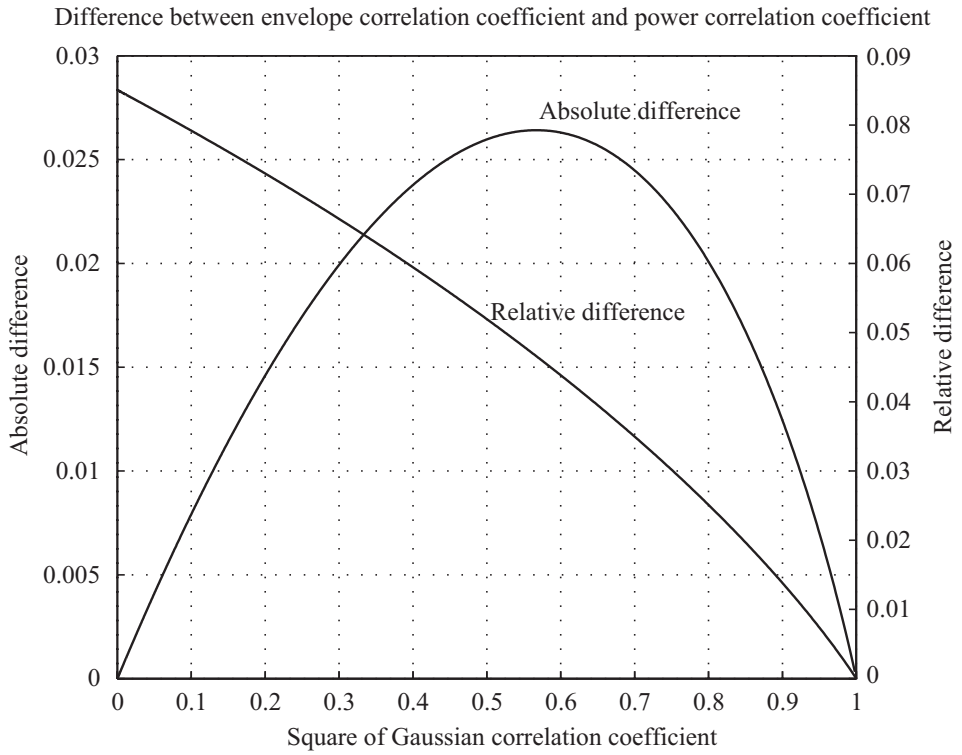


Figure 2.2.2 The absolute error, $\left(\rho_{r_1^2 r_2^2} - \rho_{r_1 r_2}\right)$, *and the relative error,* $\left(\rho_{r_1^2 r_2^2} - \rho_{r_1 r_2}\right) / \rho_{r_1^2 r_2^2}$, *between the power correlation coefficient* $\left(\rho_{r_1^2 r_2^2} = |\rho_{z_1 z_2}|^2\right)$ *and the envelope correlation coefficient*

coefficient to be written as (see also Chapter 7)

$$\rho_r = \frac{2E\left(\rho^2\right) - (1-\rho^2)K(\rho^2) - \pi/2}{2 - \pi/2}. \tag{2.2.60}$$

Here $K(m)$ and $E(m)$ are the complete elliptical functions (Abramowitz and Stegun, 1970) of the first and second kind respectively, and m is the square of the modulus parameter. The similarity of the envelope squared correlation coefficient and the envelope correlation coefficient, here denoted for cross-correlation coefficients,

$$|\rho_{z_1 z_2}|^2 = \rho_{r_1^2 r_2^2} \approx \rho_{r_1 r_2}, \tag{2.2.61}$$

is depicted in Figure 2.2.2, where it is clear that the approximation is worst for decreasing absolute values of the correlation coefficient. In summary, the accuracy of the approximation in between the limiting correlation coefficient of the envelope and power signals, is less than about 10 per cent maximum relative error.

2.2.4.3 Estimated correlation coefficient

For finite interval (sample record size) correlation coefficient function estimates, a typical estimation formula of normalised zero mean complex Gaussians, w_1 and w_2, is

$$\hat{\rho}_{w_1 w_2}(\chi) = \frac{1}{N-1} \sum_{n=1}^{N-\chi} w_1(n) w_2^*(n+\chi), \tag{2.2.62}$$

where χ is a positive correlation displacement and the measurement interval comprises N independent, or uncorrelated (since the signals are Gaussian) samples. The estimates of the two equated correlation coefficients, ρ_{r^2} and $|\rho_z|^2$, are different for a finite sample size, but they converge for increasing sample size, or measurement interval. The estimate of $|\rho_z|^2$ tends to be more accurate than that of ρ_{r^2}, but the calculation requires complex data of course.

2.3 References

ABRAMOWITZ, M. and STEGUN, I.A.: *Handbook of mathematical functions*, Dover Publications, 1970.

BELLO, P.A.: 'Characterisation of randomly time-variant linear channels', *IEEE Transactions*, December 1963, **CS-11**, pp. 360–393.

MIDDLETON, D.: *An Introduction to Statistical Communication Theory*, McGraw-Hill, 1960.

PIERCE, J.R. and STEIN, S.: 'Multiple diversity with non-independent fading', *Proceedings of the IRE*, January, 1960, **48**, pp. 89–104.

Chapter 3

Basic multipath mechanisms

3.1 Reflection from a smooth, planar surface

3.1.1 Basic configuration

The simplest case of multipath propagation comprises one reflected path in addition to a direct path. This section discusses the fields resulting from a reflection from a smooth planar surface, which has direct relevance to propagation along streets and across rural areas, and provides the basic reflection mechanism used in later sections. The geometry is given in Figure 3.1.1 where the transmitter and receiver are at heights h_T and h_R respectively, and are separated by a horizontal distance d. The quantity

$$\Delta = R_i - R_d \tag{3.1.1}$$

is the path difference between the indirect path length R_i and the direct path length R_d, and E_0 is the field at the receiver position resulting from the direct free-space path.

The reflection coefficients are given by the Fresnel coefficients R_v and R_h in which the subscripts 'v' and 'h' denote vertical and horizontal polarisation respectively.

For smooth surface reflection, the surface where the reflection occurs is assumed to be electrically large, i.e. covering several Fresnel zones (see Appendix D). Although the configuration is expressed as a classical outdoor scenario, the results can also apply to indoor surfaces such as floors, walls and ceilings. The vertical polarisation case is also referred to as parallel polarisation, because the plane containing the electric field is parallel to the plane of incidence. Likewise, the horizontal polarisation case is also called perpendicularly polarised. The use of 'vertical' presents an unfortunate clash of traditional engineering terminologies. In propagation terms, the vertical component is the z-directed component of the wave. In traditional antenna terminology, the vertical component refers to the parallel polarisation, which is denoted the θ-component, referred to standard spherical polar coordinates at the antenna. For the horizontal surface (in the x–y-plane) of Figure 3.1.1, the antenna's *vertically polarised* component is in the vertical direction (z-coordinate) only when the grazing angle is zero. In fact, at normal incidence ($\psi = 0$) the 'vertical polarisation' is geographically horizontal!

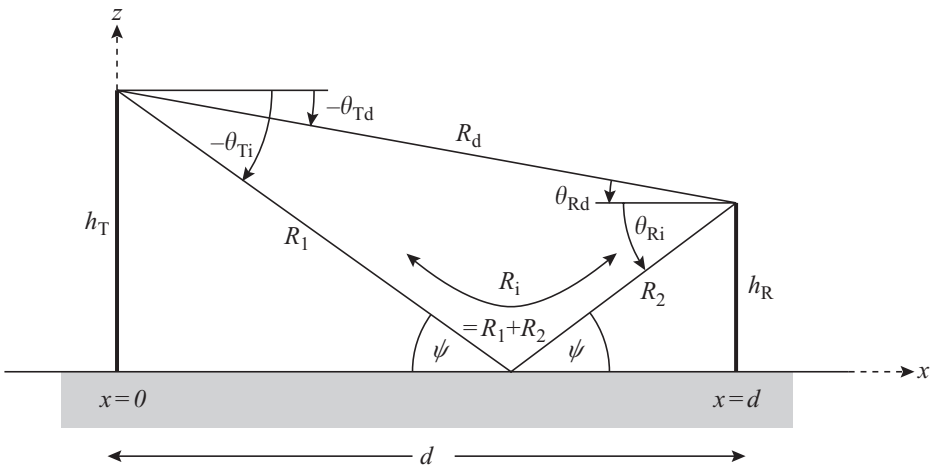


Figure 3.1.1 *A single reflection of a smooth surface and a direct signal path comprise the simplest case of multipath propagation: ψ is the grazing angle, and grazing incidence is when ψ is small; Δ is the path length difference between the direct path R_d and the indirect reflected path, R_i. The planar surface is in the x–y plane, and the z-coordinate is directed upwards. 'Vertical' polarisation refers to the z-component of the electric field in traditional propagation terminology, but means 'parallel polarisation' in traditional antenna terminology*

Clarification of 'vertical' is made where necessary in the following sections. For the Fresnel coefficients, 'vertical' is always the θ-direction, or parallel polarisation.

The medium below the planar surface is considered as homogeneous dielectric with a relative dielectric constant

$$\epsilon_r = \epsilon_r' - j\epsilon_r'' \tag{3.1.2}$$

so that ϵ_r' is the relative permittivity and

$$\epsilon_r'' = \frac{\sigma}{\omega\epsilon_0} \tag{3.1.3}$$

is the imaginary, or lossy part, with σ the conductivity and ω the angular radio frequency. The permittivity is then denoted

$$\epsilon = \epsilon_r\epsilon_0 = \epsilon' - j\epsilon''. \tag{3.1.4}$$

The ratio of the dielectric part to the conductive part,

$$\delta = \frac{\epsilon''}{\epsilon'} \tag{3.1.5}$$

is the loss factor, or the loss tangent if its tangent is taken.

3.1.2 Sum of vertical and horizontal fields at the receiver

The exact expression for the field over a lossy ground from a vertical (z-directed) dipole is quite complicated. However, for reasonably large values of the relative permittivity ϵ_r, which includes most surfaces of interest (building materials and dry ground at microwave frequencies), the expressions can be simplified. In this case, the following expressions for the vertical (z-directed) and horizontal fields from transmitting dipoles are accurate for the fields at the receiver position (Sommerfeld, 1909; Norton, 1937; Jordan and Balmain, 1968):

$$E_z = E_\mathrm{v} \approx E_0 \left(\cos^2 \psi \left(\mathrm{e}^{\mathrm{j}k_0 \Delta} + R_\mathrm{v} \right) + (1 - R_\mathrm{v})\, F(p_\mathrm{v}) \right), \tag{3.1.6}$$

$$E_y = E_\mathrm{h} \approx E_0 \left(\left(\mathrm{e}^{\mathrm{j}k_0 \Delta} + R_\mathrm{h} \right) + (1 - R_\mathrm{h})\, F(p_\mathrm{h}) \right). \tag{3.1.7}$$

The first term in each equation is the summation of the direct and reflected paths, depicted in Figure 3.1.1. E_0 is the free-space direct path contribution to the fields at the receiver,

$$E_0 = \frac{\mathrm{e}^{-\mathrm{j}kR_\mathrm{d}}}{R_\mathrm{d}} \tag{3.1.8}$$

and R_v is the reflection coefficient at the surface. The $\cos^2 \psi$ factor arises partly from the $\cos \psi$ far-field pattern of the vertical transmitting dipole. For the horizontal case, the factor is unity, which is related to the omnidirectional pattern (in the plane of propagation) of the horizontal transmitting dipole. For other directional antennas, the equations would change. The second term in the equations is a surface wave contribution, sometimes called the Norton surface wave. The space wave contributions are discussed further below, following a brief review of the surface wave.

3.1.3 Norton surface wave

The Norton surface wave transfers energy between the transmitter and the receiver by travelling only along the dielectric interface. It attenuates more quickly than a free-space wave. Here, p_v and p_h are called the numerical distance, a complex quantity in general, which, for grazing incidence, is approximated as

$$p_\mathrm{v} \approx -\mathrm{j}\frac{k_0 d}{2\epsilon_\mathrm{r}} \tag{3.1.9}$$

for the vertical polarisation case. Thus, p_v is proportional to the electrical distance along the surface between the antennas and inversely proportional to the permittivity of the medium. $F(p)$ is Sommerfeld's attenuation function

$$F(p_\mathrm{v}) = 1 - \mathrm{j}\sqrt{\pi p_\mathrm{v}}\mathrm{e}^{-p_\mathrm{v}}\mathrm{erfc}\left(\mathrm{j}\sqrt{p_\mathrm{v}}\right) \tag{3.1.10}$$

in which

$$\mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty \mathrm{e}^{-t^2}\, \mathrm{d}t \tag{3.1.11}$$

is the complementary error function.

At microwave frequencies, p_v is numerically very large for typical physical distances in mobile communications, $F(p_v)$ is small, and the surface wave field becomes negligible. However, when both antennas are on the ground, then $\Delta = 0$ and $R_v = -1$, so the two space wave contributions of the first term cancel leaving only the surface wave for energy transfer. As soon as the transmitting dipole or the receiving position (or both) become raised above the ground, the space waves do not cancel and they rapidly dominate the surface wave in defining the behaviour of the total field at the receiver.

A similar argument may be used for horizontal polarisation from a horizontal dipole, which produces a weaker surface wave than that of the vertical polarisation case. The formulation for horizontal polarisation uses a different expression for the numerical distance, p_h, which gives larger values for a given physical distance, d, than for the vertical polarisation case. Also, the $\cos^2 \psi$ term becomes unity as mentioned above.

In summary, the surface wave is the only means of transmission when both receiver and transmitter are on the surface. For frequencies lower than those typically used in mobile communications, and with a conducting dielectric earth, the surface wave can be important. This is evident in some radio broadcasting, for example. For microwave frequencies, the surface wave propagation can be neglected.

3.1.4 Sum of direct and reflected space waves at grazing incidence

The space waves arrive at the receiver from different directions in general. The combination of the waves occurs in the receiving antenna, giving the open circuit voltage as

$$V_{oc} = F_d \frac{e^{-jkR_d}}{R_d} + F_i \frac{e^{-jkR_i}}{R_i} \tag{3.1.12a}$$

$$\approx E_0 \left(F_d + F_i e^{-jk\Delta} \right), \tag{3.1.12b}$$

where F_d and F_i are the coupling terms for the direct and indirect waves. These terms are the product of the directional complex amplitude gains of the antennas. The patterns in the propagation plane are denoted as in Section 2.1, here for the transmitting antenna,

$$\boldsymbol{h}_T(\theta) = h_{T\theta}(\theta)\hat{\theta} + h_{T\phi}(\theta)\hat{\phi} \tag{3.1.13}$$

with θ corresponding to the vertical (parallel) polarisation and ϕ corresponding to the horizontal polarisation. The 'T' and 'R' subscripts again denote transmit and receive antennas. Then

$$F_d = h_{T\theta}(-\theta_{Td}) h_{R\theta}(\theta_{Rd}) + h_{T\phi}(-\theta_{Td}) h_{R\phi}(\theta_{Rd}) \tag{3.1.14}$$

for the direct wave; and including the Fresnel reflection coefficients (discussed below) for the indirect wave,

$$F_i = h_{T\theta}(-\theta_{Ti}) h_{R\theta}(\theta_{Ri}) R_v + h_{T\phi}(-\theta_{Ti}) h_{R\phi}(\theta_{Ri}) R_h. \tag{3.1.15}$$

The reflection coefficients also have an angular dependence, which is discussed in the following section. If the reference direction ($\theta = 0$) for each antenna is horizontal in the direction of the other antenna, then

$$\theta_{\mathrm{Rd}} = -\theta_{\mathrm{Td}} = \tan^{-1}\left(\frac{h_{\mathrm{T}} - h_{\mathrm{R}}}{d}\right) = \cos^{-1}\left(\frac{d}{R_{\mathrm{d}}}\right) \tag{3.1.16}$$

$$\theta_{\mathrm{Ri}} = -\theta_{\mathrm{Ti}} = \psi = \tan^{-1}\left(\frac{h_{\mathrm{T}} + h_{\mathrm{R}}}{d}\right) = \cos^{-1}\left(\frac{d}{R_{\mathrm{i}}}\right). \tag{3.1.17}$$

The formulation is revisited below in describing the *image model*. The implicit approximation that $R_{\mathrm{d}} \approx R_{\mathrm{i}}$ only counts for the inverse distance amplitude reduction, which is a slowly varying term. The difference between the distances, Δ, is preserved in the phase that dominates the behaviour of the received signal.

Although an antenna is required to combine wave signals, fields can be summed if they originate from the same direction. For very small grazing angles, or *grazing incidence*, the direct and indirect waves can be approximated as coming from the same direction. Combining the vertical (z-direction) and horizontal polarisation subscripts for brevity, the total field at the receiving antenna, expressed as a summation, is then

$$E_{\mathrm{v,h}} = \frac{\mathrm{e}^{-\mathrm{j}kR_{\mathrm{d}}}}{R_{\mathrm{d}}} + R_{\mathrm{v,h}}\frac{\mathrm{e}^{-\mathrm{j}kR_{\mathrm{i}}}}{R_{\mathrm{i}}} \tag{3.1.18a}$$

$$\approx E_0\left(1 + R_{\mathrm{v,h}}\mathrm{e}^{-\mathrm{j}k\Delta}\right). \tag{3.1.18b}$$

This is the same as equation (3.12a,b) but with no antenna dependencies.

3.1.5 Fresnel coefficients

The reflection coefficients are given by

$$R_{\mathrm{h}} = \frac{\sin\psi - \sqrt{\epsilon_{\mathrm{r}} - \cos^2\psi}}{\sin\psi + \sqrt{\epsilon_{\mathrm{r}} - \cos^2\psi}} \tag{3.1.19}$$

for the horizontal polarisation and

$$R_{\mathrm{v}} = \frac{\epsilon_{\mathrm{r}}\sin\psi - \sqrt{\epsilon_{\mathrm{r}} - \cos^2\psi}}{\epsilon_{\mathrm{r}}\sin\psi + \sqrt{\epsilon_{\mathrm{r}} - \cos^2\psi}} \tag{3.1.20}$$

for the vertical (parallel) polarisation. It has been assumed that the medium is non-magnetic. In Section 3.2, the complex permittivity of the medium will be modified to help the Fresnel equations cater for rough, instead of smooth, surfaces.

For the vertical polarisation, the Brewster angle phenomenon influences the propagation behaviour. The angle corresponds to when $|R_{\mathrm{v}}|$ is a minimum, i.e. when the grazing angle is

$$\psi_{\mathrm{B}} = \Re\left\{\tan^{-1}\left(\frac{1}{\sqrt{\epsilon_{\mathrm{r}}}}\right)\right\}. \tag{3.1.21}$$

When the permittivity is purely real, then $R_v = 0$ at the Brewster angle and taking the real part is unnecessary. This case is also called the polarising angle from the fact that only horizontally polarised power is reflected at ψ_B, independent of the incident polarisation. The Brewster grazing angle is always less than 45°.

The phase of the vertical reflection coefficient near the Brewster angle changes quickly. At grazing incidence, the phase is 180°, and at normal incidence, the phase is 0°. The phase at the Brewster angle (assuming finite magnitude of the reflection coefficient) is always 90°. If the receiver has a signal contribution from a Brewster angle reflection, then as the receiver moves, that contribution will change more quickly than usual; namely, the fading and the spatial Doppler frequency will be higher than normally expected from simple geometric considerations.

As the conductivity of the medium increases, the minimum reflection coefficient at the Brewster angle increases and the Brewster grazing angle itself decreases. When the medium is perfectly conducting ($\sigma \to \infty$, $\varepsilon_r \to -j\infty$) the Brewster grazing angle is effectively zero (or strictly speaking, does not exist), and the vertical reflection coefficient is always +1, and −1 for horizontal polarisation. Also, media with higher relative permittivities have Brewster angles at smaller grazing angles. This behaviour is shown in Figure 3.1.2 where the Brewster angle is plotted against the relative permittivity of the medium with the loss factor $\delta = \varepsilon''/\varepsilon'$ as a parameter.

In contrast to the vertical polarisation case, the horizontal polarisation has no Brewster angle phenomena with the phase virtually constant, being 180° for grazing incidence and close to 180° for all other grazing angles.

For both hands of circular polarisation, the co-polar and cross-polar reflection coefficients respectively are

$$R_{co} = \tfrac{1}{2}(R_v + R_h), \quad R_{cx} = \tfrac{1}{2}(R_v - R_h). \tag{3.1.22}$$

The magnitude of the four reflection coefficients are plotted in Figure 3.1.3 for a medium of 2.5–j0.25, corresponding to some light plastering concretes at microwave frequencies.

The plot shows the strikingly different behaviour for the different polarisations. The horizontal coefficient drops off with increasing grazing angle. At normal incidence, about 5 per cent (from 0.23^2) of the power is reflected, and the remainder travels into the dielectric medium. For the vertical case, the Brewster angle is obvious, at about 32° for this example. This is where there is a rapid phase change with changing reflection angle. For example, at $\psi = 31.5°$ the phase is 9°, and at $\psi = 33°$, the phase is 170°.

3.1.6 Example of Brewster angle effect with a moving receiver

The following example may help crystallise some of the concepts discussed. For a microcellular base station with $h_T = 5$ m and a personal mobile approaching with $h_R = 1.5$ m, the distance $d = (h_T + h_R)/\tan\psi$ changes from 10.6 m to 9.9 m, i.e. an additional spatial Doppler shift of $(170 - 9)^*(\pi/180)/0.7 = 4$ rad m^{-1} occurs over

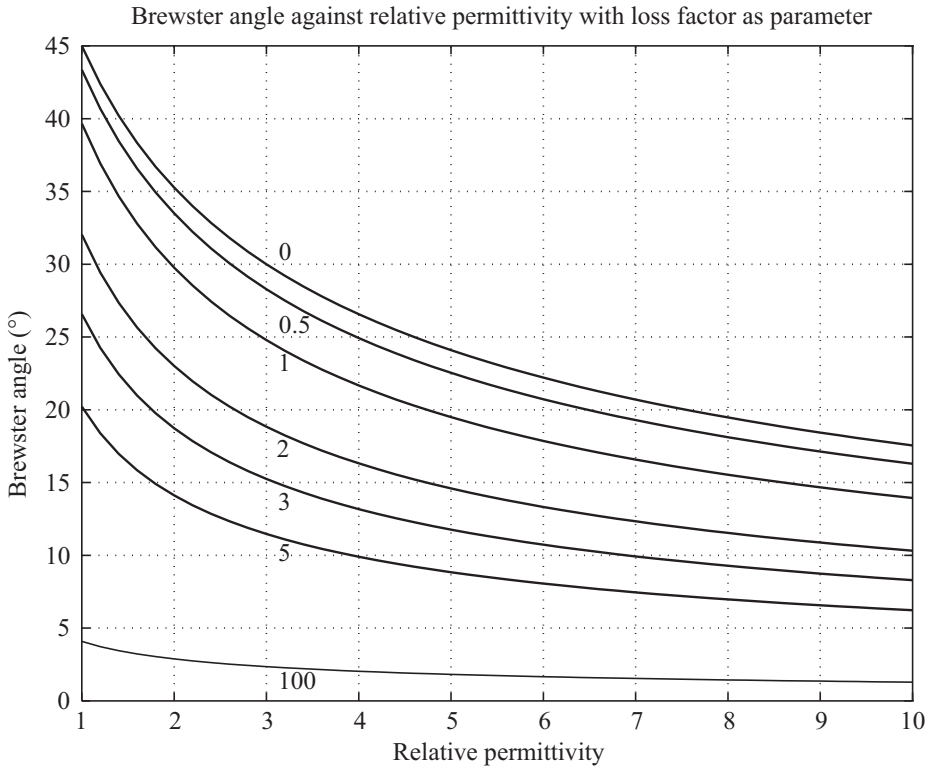


Figure 3.1.2 The Brewster grazing angle as a function of the relative permittivity and with the loss factor (ratio of the imaginary part to the real part of the permittivity) as a parameter. An increase in the magnitude of the complex permittivity reduces the Brewster grazing angle. Most surfaces encountered in mobile communications have loss factors much smaller than unity

this distance. The example is shown in Figure 3.1.4, in which the change in the instantaneous Doppler profile[1] is given as the mobile passes through the Brewster angle.

3.1.7 Phase of the reflection at normal incidence

Returning to the Fresnel equations; as the grazing angle approaches normal incidence, the linear polarisations become geometrically alike and the coefficients converge in amplitude. However, their phases are opposing. This is a consequence of the directional definitions for the two polarisations, which are shown in Figure 3.1.5. For the vertical (parallel) polarisation, it is seen that, at normal incidence, the incident and reflected waves have defining vectors in opposite directions. A mathematical phase inversion during reflection is therefore interpreted as resulting in the reflected and incident waves having the same phase. Similarly, a mathematical zero phase change during reflection means that the incident and reflected waves have opposite phase. So although the horizontal and vertical polarisations have identical geometry

[1] The Doppler profile is normally a statistical quantity but here, in a deterministic situation, it is treated as an instantaneous, deterministic function.

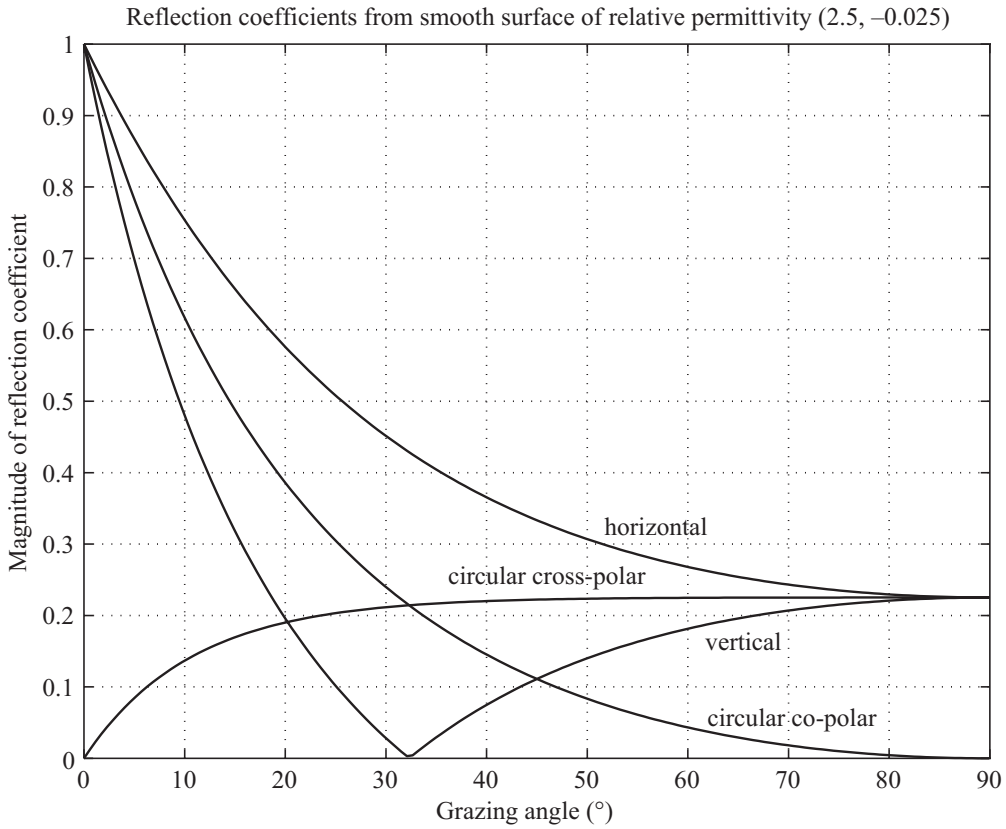


Figure 3.1.3 *The magnitude of reflection coefficients for a smooth-surfaced medium, similar to light plastering concrete at microwave frequencies. At grazing incidence, virtually all the power is reflected, whereas much of the power is transmitted into the medium at other angles. The Brewster angle is at about* 32°. *The phase of the vertical reflection is essentially* 0° *for* $\psi > 32°$ *and* 180° *for* $\psi < 32°$

at normal incidence, the reflection coefficients have opposite signs, resulting from the directional definitions for the fields.

3.1.8 Circular polarisation

The results for circular polarisation are also of interest. In Figure 3.1.3, the co-polar coefficient drops from virtually total reflection to zero reflection with increasing grazing angle. Some of the reflected power appears as cross-polarised, which grows with increasing grazing angle, from pure cross-polarised reflection at normal incidence to pure co-polarised reflection at grazing incidence. This is because the vertical component undergoes no phase change for reflection at normal incidence, whereas the horizontal component is inverted, as discussed above. Likewise, at grazing incidence, both linear components are inverted so no cross-polarisation appears. At the Brewster angle, the co- and cross-polar reflection coefficients are virtually the same, since the reflected polarisation is essentially horizontal. Also, over a highly conductive

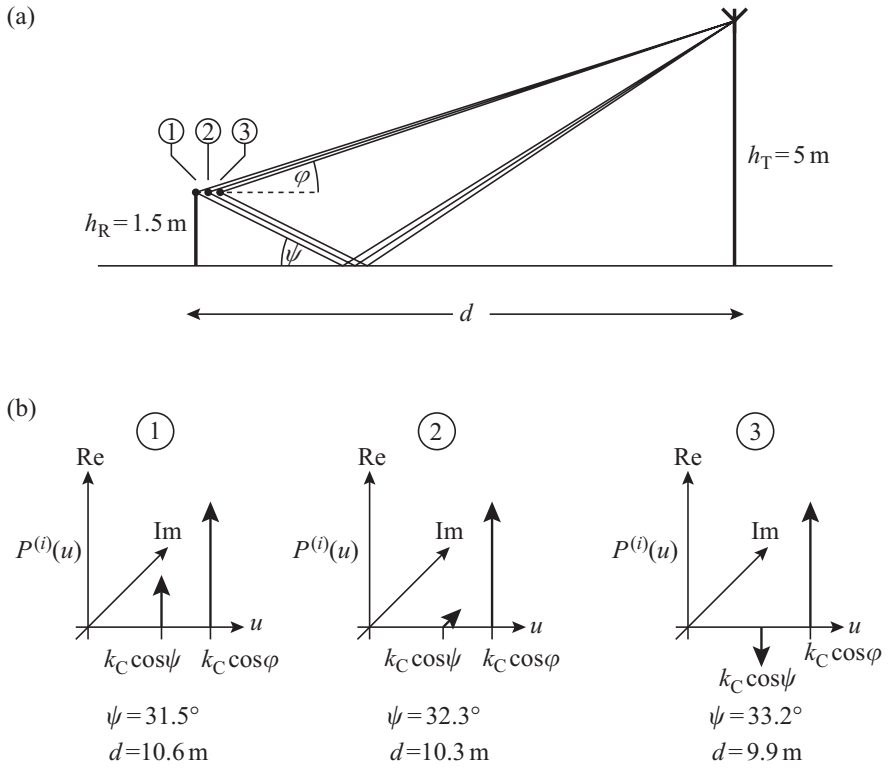


Figure 3.1.4 *An example where the reflection passes through the Brewster angle. The configuration is like an urban microcell with a smooth concrete ground, an elevated base station and a handheld mobile. For a moving receiver, the instantaneous Doppler profiles at points 1, 2, and 3 show the changing contribution of the reflection, and the essentially static contribution from the direct path*

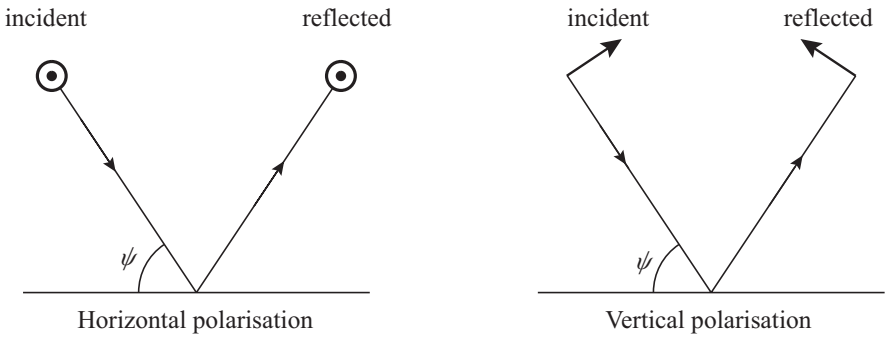


Figure 3.1.5 *The defining directions for the reflection of horizontally and vertically polarised waves. At normal incidence, the horizontally polarised incident and reflected waves are defined in the same direction, whereas for vertically polarised waves, the incident and reflected waves are defined as having opposite directions*

surface, only the vertical polarisation can exist adjacent to the ground, and the circular polarisation cannot be preserved here.

The phase of the circular co-polar is essentially constant at 180°, whereas the cross-polar phase is similarly 0°. This is because the reflected horizontal component is always larger than the vertical, and so the phase of the circular polarisations follows the phase of the horizontal component.

3.1.9 Image model

For image models, the surface and the dielectric medium is removed and an extra source is added to cater for the received energy from the reflected path. As indicated in Figure 3.1.6, the source is the image in the surface plane of the transmitter, with its power and phase modified by the reflection coefficient. It follows that the pattern of the source $\boldsymbol{h}_{\mathbf{Ti}}(\theta)$ is also mirrored for the image, again modified by the directionality (i.e. grazing angle dependence) of the reflection coefficient, as expressed in equations (3.1.19) and (3.1.20). The coordinate system for the formulation uses horizontal as the reference direction, corresponding to $\theta = 0$.

Denoting the reflection coefficient components using the vector

$$\boldsymbol{R} = \hat{\theta} R_{\mathrm{v}} + \hat{\phi} R_{\mathrm{h}}, \tag{3.1.23}$$

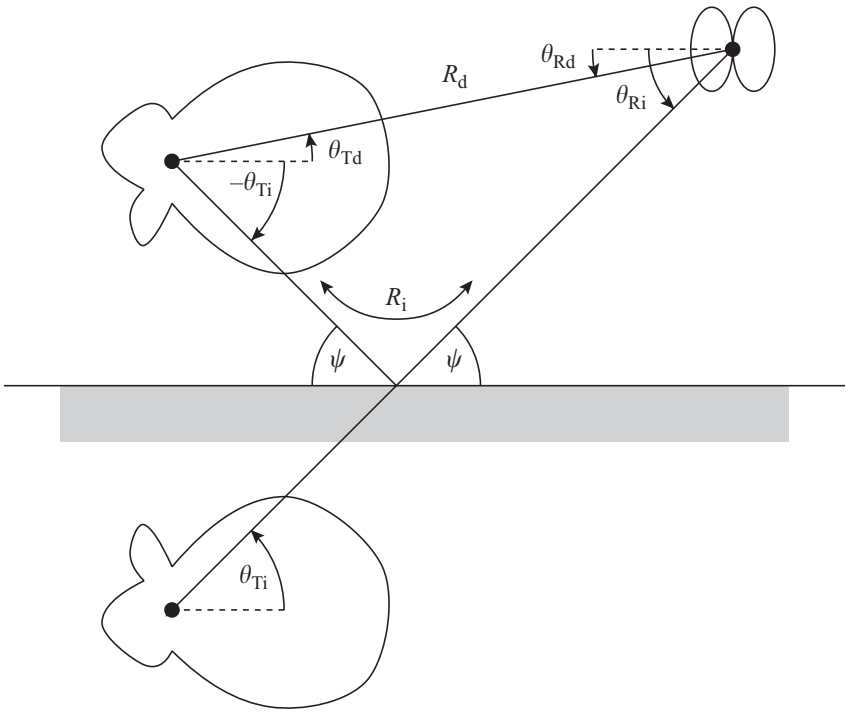


Figure 3.1.6 The image model is a free-space (without reflections) configuration which, as far as the receiving end of the link is concerned, is equivalent to the real-world configuration with reflections. The model uses an image of the transmitter with mirrored directionality, modified by the reflection coefficient, to reproduce the wave from a reflection

the directional pattern of the image is

$$\boldsymbol{h}_{\mathbf{Ti}}(\theta) = \hat{\theta} R_{\mathrm{v}}(\psi) h_{\mathrm{T}\theta}(\theta_{\mathrm{Ti}} = \psi) + \hat{\phi} R_{\mathrm{h}}(\psi) h_{\mathrm{T}\phi}(\theta_{\mathrm{Ti}} = \psi), \tag{3.1.24}$$

where $\boldsymbol{h}_{\mathrm{T}}(\theta)$ is the vector pattern of the transmitter in the propagation plane and $\theta = \psi$ is the pattern angle from the horizontal direction as indicated in Figure 3.1.6. The signal gathered by the receiving antenna is proportional to (cf. equation 3.1.12)

$$V_{\mathrm{oc}} = \boldsymbol{h}_{\mathbf{T}} \cdot \boldsymbol{h}_{\mathbf{R}}^{*} \frac{\mathrm{e}^{-\mathrm{j}kR_{\mathrm{d}}}}{R_{\mathrm{d}}} + \boldsymbol{h}_{\mathbf{Ti}} \cdot \boldsymbol{h}_{\mathbf{R}}^{*} \frac{\mathrm{e}^{-\mathrm{j}kR_{\mathrm{i}}}}{R_{\mathrm{i}}}. \tag{3.1.25}$$

The image model is particularly useful for analysing some aspects of indoor communications scenarios where there are multiple planar surfaces with multiple reflections.

3.1.10 Grazing incidence

The grazing incidence is when ψ is small. Exactly how small is not just a question of geometry, it also depends on the polarisation. At grazing incidence we have $|R_{\mathrm{h}}| \approx |R_{\mathrm{v}}| \approx 1$. A small ψ leads to the approximation

$$\sin\psi \approx \tan\psi = \frac{h_{\mathrm{T}} + h_{\mathrm{R}}}{d}, \tag{3.1.26}$$

and it is also clear from the Fresnel equations that both R_{v} and R_{h} approach -1. Here, the basic difference between the reflection coefficients may be seen from their expansion around -1, which gives

$$R_{\mathrm{v}} \approx -1 + \frac{2\,(h_{\mathrm{T}} + h_{\mathrm{R}})}{d} \frac{\epsilon}{\sqrt{\epsilon - 1}}, \tag{3.1.27}$$

$$R_{\mathrm{h}} \approx -1 + \frac{2\,(h_{\mathrm{T}} + h_{\mathrm{R}})}{d} \frac{1}{\sqrt{\epsilon - 1}}. \tag{3.1.28}$$

So for grazing incidence, R_{h} is closer to -1 than R_{v} is. This is evident in Figure 3.1.3 where $|R_{\mathrm{h}}| > |R_{\mathrm{v}}|$ during their approach to $\psi = 0$.

3.1.11 Propagation effects at grazing incidence

To get some appreciation of the propagation effects of the two rays, the approximations $R_{\mathrm{v,h}} = -1$ are used in equation (3.1.18b), to offer the simple form for the summed fields

$$|E_{\mathrm{v,h}}|^2 \approx |E_0|^2 4 \sin^2\left(\frac{k\Delta}{2}\right). \tag{3.1.29}$$

When the path difference Δ varies, the direct and the reflected fields interfere to give a total field comprising fades alternating with maxima. In practice, the fades are deeper for horizontal polarisation than for vertical polarisation, since R_{h} is closer to -1 than R_{v} is.

The small-ψ assumption also corresponds to $h_1, h_2 << d$, and so the electrical path difference may be approximated by

$$k\Delta \approx \frac{4\pi}{\lambda}\frac{h_\mathrm{T}h_\mathrm{R}}{d} \tag{3.1.30}$$

and the sum of the fields can now be written as an explicit function of the antenna heights and separation distance, and frequency:

$$|E_{\mathrm{v,h}}|^2 \approx |E_0|^2 4\sin^2\left(\frac{2\pi}{\lambda}\frac{h_\mathrm{T}h_\mathrm{R}}{d}\right). \tag{3.1.31}$$

3.1.12 Distance dependence at grazing incidence

Equation (3.1.31) shows that for large values of separation distance d, the fields decay asymptotically. For smaller values of d, the fields fade and peak. These regions of different behaviour meet at a breakpoint, denoted $d = d_\mathrm{b}$. The breakpoint is at the separation distance of the last field maximum (for increasing d), and occurs when

$$\frac{k\Delta}{2} = \frac{\pi}{2}, \tag{3.1.32}$$

i.e.

$$d_\mathrm{b} = \frac{4h_\mathrm{T}h_\mathrm{R}}{\lambda}. \tag{3.1.33}$$

Note that the breakpoint is identical to the border of the first Fresnel zone (Appendix D), agreeing with the notion that when an obstacle (in this case the ground) is inside the first Fresnel zone, major modifications occur to the fields at the receiver.

For $d < d_\mathrm{b}$, i.e. before the breakpoint, the fades occur along the distance coordinate, at the decreasing distances

$$d^{(\min)} = \frac{2h_\mathrm{T}h_\mathrm{R}}{\lambda} \times \left(1, \frac{1}{2}, \frac{1}{3}, \ldots\right), \tag{3.1.34}$$

which it is recalled is valid as long as there is grazing incidence. This direction along the distance coordinate is also called, self-evidently, the radial direction.

The fading and peaking of the fields along the distance coordinate can be averaged in equation (3.1.31) to give

$$\langle |E_{\mathrm{v,h}}|^2 \rangle = |E_0|^2\, 2 = \frac{2}{d^2}, \tag{3.1.35}$$

since

$$\left\langle \sin^2\left(\frac{k\Delta}{2}\right)\right\rangle = \frac{1}{2}, \tag{3.1.36}$$

so the averaged power decays as d^{-2}, independent of the antenna heights.

For $d > d_b$, i.e. after the breakpoint, equations (3.1.31) and (3.1.35) give

$$|E|^2 = \frac{1}{d^2}\left(\frac{4\pi h_T h_R}{\lambda d}\right)^2, \tag{3.1.37}$$

i.e. the power of the field decays as d^{-4}. This additional attenuation is not a result of absorption in the medium, it is a cancellation of the two space-wave contributions occurring for smooth surfaces at grazing incidence. Surface roughness and other larger objects in this first Fresnel zone will spoil the cancellation. In this region, the antennas both offer a 'height gain' factor of 6 dB per octave: a doubling of the height of either antenna increases the channel gain, or decreases the path loss (see below) by 6 dB.

Equation (3.1.37) gives the impression that the field is zero (or the path loss is infinite) when one antenna height is zero, but this is only true in the limit when the reflection coefficients are -1. A more accurate expression for the total field, taking one more term in the expansions of equations (3.1.27) and (3.1.28), and assuming that $\epsilon' \gg \epsilon''$, yields

$$\left|\frac{E_h}{E_0}\right|^2 = \frac{1}{d^2}\left[\frac{4}{\epsilon'_r}(h_T + h_R)^2 + (2kh_T h_R)^2\right] \tag{3.1.38}$$

for horizontal polarisation, and

$$\left|\frac{E_v}{E_0}\right|^2 = \frac{1}{d^2}\left[4\epsilon'_r(h_T + h_R)^2 + (2kh_T h_R)^2\right] \tag{3.1.39}$$

for vertical polarisation. The inverse fourth law is still present because of the inverse-square dependence of the free space field, E_0. Equations (3.1.38) and (3.1.39) show that the vertically polarised wave is slightly larger than the horizontally polarised wave, independent of any absorption in the scattering surface. However, the second term, which is common to both equations, often dominates the behaviour and there may be no noticeable asymptotic difference between the polarisations.

3.1.12.1 Receiver height dependence at grazing incidence

It is also of interest to study the variation of the summed fields in the vertical direction. This is in a plane that is at right angles to the distance coordinate. The field minima are spaced by

$$\Delta^{(\min)} = \frac{\lambda d}{h_T}, \quad n = 1, 2, \ldots. \tag{3.1.40}$$

3.1.13 Frequency dependence at grazing incidence

The frequency dependence is best studied by introducing the free-space path loss, L_0 (see Appendix A), which is defined using the ratio of received power P_R to transmitted power P_T,

$$\frac{P_R}{P_T} = \frac{G_T G_R}{L_0}, \tag{3.1.41}$$

and is given by

$$L_0 = \left(\frac{4\pi d}{\lambda}\right)^2, \tag{3.1.42}$$

where G_T and G_R are the antenna gains. For the asymptotic decay region $d > d_b$, the additional decay factor due to the reflection (equation (3.1.37)), is combined with the space loss to give an overall path loss L as

$$\begin{aligned} L &= \left(\frac{4\pi d}{\lambda}\right)^2 \left(\frac{\lambda d}{4\pi h_T h_R}\right)^2 \\ &= \frac{d^4}{(h_T h_R)^2}, \end{aligned} \tag{3.1.43}$$

which depends only on the geometry, and is independent of frequency! However, the breakpoint distance, d_b, from equation (3.1.33), increases linearly with frequency.

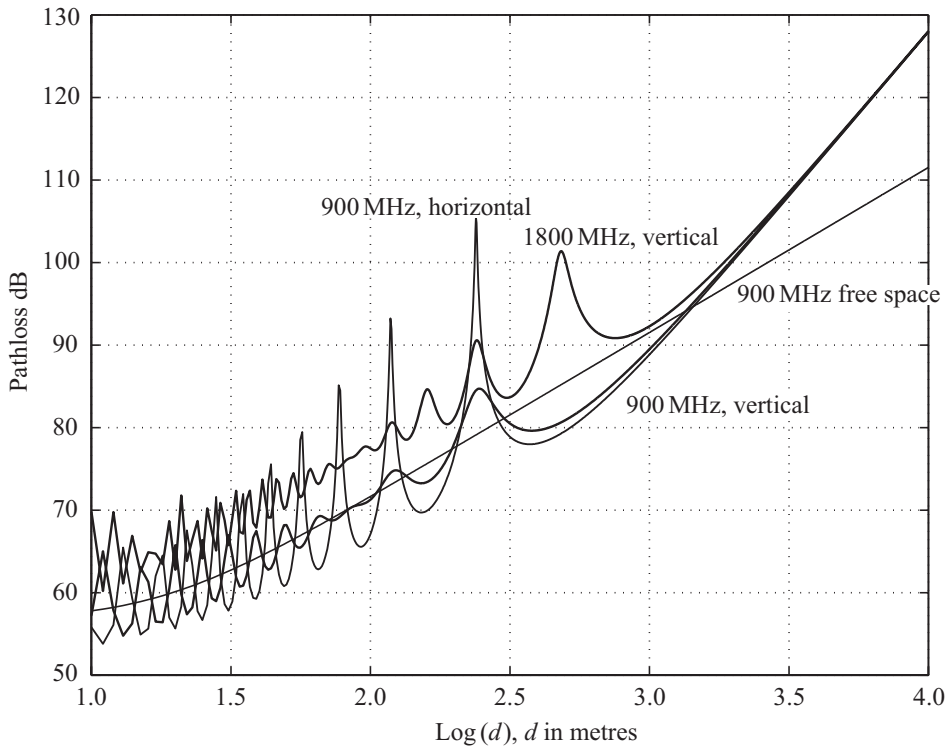


Figure 3.1.7 *Path loss vs. log distance (d in metres), $\epsilon = 15, \sigma = 0.005$ S m^{-1}, $h_T = 20$ m, $h_R = 2$ m. Note the distinction between the fading and the decaying region, the Brewster angle occurring at* $\log d \approx 1.9$, *and the independence of frequency and polarisation in the decay region*

Table 3.1 Combinations of mixed polarisations for path loss over a smooth surface. Four reflection coefficients with corresponding path loss expressions cover all the combinations

Transmitted polarisation	Receiving polarisation	Reflection coefficient	(Total received signal)/E_0
RHCP	RHCP	$R_{\mathrm{co}} = (R_{\mathrm{v}} + R_{\mathrm{v}})/2$	$1 + R_{\mathrm{co}}\mathrm{e}^{-\mathrm{j}kD}$
$(\hat{\theta} - \mathrm{j}\hat{\phi})/\sqrt{2}$	LHCP	$R_{\mathrm{cx}} = (R_{\mathrm{v}} - R_{\mathrm{v}})/2$	$R_{\mathrm{cx}}\mathrm{e}^{-\mathrm{j}kD}$
	Slant (+45°)	$R_{\mathrm{lc}-} = (R_{\mathrm{v}} - \mathrm{j}R_{\mathrm{v}})/2$	$(1 + \mathrm{j})/2 + R_{\mathrm{lc}+}\mathrm{e}^{-\mathrm{j}kD}$
	Slant (−45°)	$R_{\mathrm{lc}+} = (R_{\mathrm{v}} + \mathrm{j}R_{\mathrm{v}})/2$	$(1 - \mathrm{j})/2 + R_{\mathrm{lc}-}\mathrm{e}^{-\mathrm{j}kD}$
LHCP	RHCP	R_{cx}	$R_{\mathrm{cx}}\mathrm{e}^{-\mathrm{j}kD}$
$(\hat{\theta} + \mathrm{j}\hat{\phi})/\sqrt{2}$	LHCP	R_{co}	$1 + R_{\mathrm{co}}\mathrm{e}^{-\mathrm{j}kD}$
	Slant (+45°)	$R_{\mathrm{lc}-}$	$(1 - \mathrm{j})/2 + R_{\mathrm{lc}-}\mathrm{e}^{-\mathrm{j}kD}$
	Slant (−45°)	$R_{\mathrm{lc}+}$	$(1 + \mathrm{j})/2 + R_{\mathrm{lc}+}\mathrm{e}^{-\mathrm{j}kD}$
Slant (+45°)	RHCP	$R_{\mathrm{lc}+}$	$(1 + \mathrm{j})/2 + R_{\mathrm{lc}+}\mathrm{e}^{-\mathrm{j}kD}$
$(\hat{\theta} + \hat{\phi})/\sqrt{2}$	LHCP	$R_{\mathrm{lc}-}$	$(1 - \mathrm{j})/2 + R_{\mathrm{lc}-}\mathrm{e}^{-\mathrm{j}kD}$
	Slant (+45°)	R_{co}	$1 + R_{\mathrm{co}}\mathrm{e}^{-\mathrm{j}kD}$
	Slant (−45°)	R_{cx}	$R_{\mathrm{cx}}\mathrm{e}^{-\mathrm{j}kD}$
Slant (−45°)	RHCP	$R_{\mathrm{lc}-}$	$(1 - \mathrm{j})/2 + R_{\mathrm{lc}-}\mathrm{e}^{-\mathrm{j}kD}$
$(\hat{\theta} - \hat{\phi})/\sqrt{2}$	LHCP	$R_{\mathrm{lc}+}$	$(1 + \mathrm{j})/2 + R_{\mathrm{lc}+}\mathrm{e}^{-\mathrm{j}kD}$
	Slant (+45°)	R_{cx}	$R_{\mathrm{cx}}\mathrm{e}^{-\mathrm{j}kD}$
	Slant (−45°)	R_{co}	$1 + R_{\mathrm{co}}\mathrm{e}^{-\mathrm{j}kD}$

3.1.14 Example of distance dependence: outdoor cell

To get an appreciation of typical values of d_{b}, a transmitting height of 20 m and a receiving height of 2 m produce the break point at 480 m at 900 MHz, and at 960 m at 1800 MHz. Figure 3.1.7 shows the path loss for a typical situation with the breakpoint clearly showing on the log distance scale. It is clear that the path loss after the break-point is almost independent of frequency and polarisation. Also the Brewster angle is seen to occur at a distance of about 80 m ($\log(d) = 1.9$), where the interference between the direct wave and reflected wave is least, i.e. where there is least fading of the vertical polarisation. At distances greater than when the Brewster angle occurs, the horizontal and vertical reflection coefficients have similar phases, and so the fading of the summed field is coincident for both linear polarisations. For distances closer than when the Brewster angle occurs, the phase of the reflection coefficients for the polarisations is opposite, and their fading is correspondingly 'out of phase'.

The circular polarisation behaviour is also of interest, as are the fields from slant polarisation. A sloping dipole at 45° gives an equal mix of in-phase vertical and horizontal polarisation. Table 3.1 gives a summary of the reflection coefficients and fields

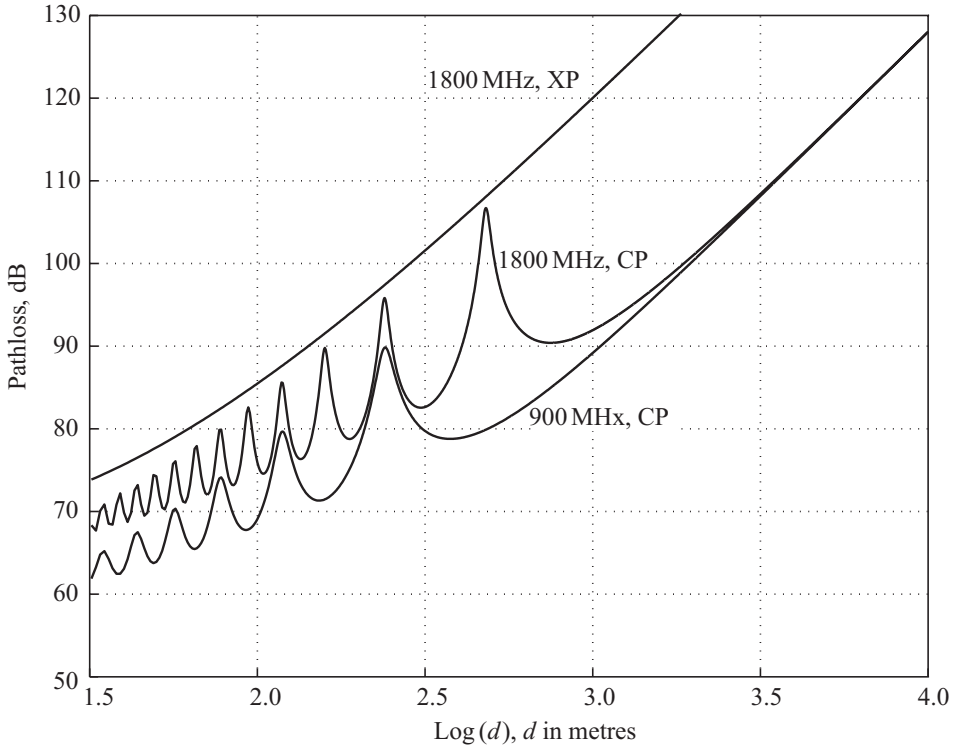


Figure 3.1.8 Path loss for circular polarisation. The cross-polarised wave is generated only from the reflection, so there is no fading

at the receiver for combinations of slant and circular polarisations. No such combination offers an overall advantage compared to, say, vertical polarisation. Figure 3.1.8 gives the path loss for circular polarisation. Circular polarisation does not seem to have any advantages over linear polarisation. The cross-polar (XP) experiences no fading since only the reflected wave is detected.

3.1.15 Fade lengths with distance

The 3 dB length of the fades may be found from combining equations (3.1.31) and (3.1.33), i.e.

$$|E_{\mathrm{v,h}}|^2 \approx |E_0|^2 4 \sin^2\left(\frac{\pi}{2}\frac{d_\mathrm{b}}{d}\right). \tag{3.1.44}$$

The last maximum occurs for $d = d_\mathrm{b}$; the penultimate minimum for $d = d_\mathrm{b}/2$; and the last fade before the break point (i.e. the first full minimum of equation (3.1.44)) has a 3 dB length of $d_\mathrm{b}/4$, i.e. some 480 m in the 900 MHz example above. The reason for this extremely long fade is the almost parallel direction of the incident and reflected waves, i.e. the vertical Doppler spread is very small. Similarly, the path lengths have only a small difference leading to small delay spread. In the example above, the path length difference is $\Delta = (2h_\mathrm{T}h_\mathrm{R})/d \approx 15$ cm, giving a time-of-arrival difference of 0.5 ns. If this fade were to be countered with some form of equalisation in the

receiver, the signal bandwidth should exceed 2 GHz in order to be able to resolve the two components. The carrier would need to be at a much higher frequency than in the above examples to allow such a bandwidth. For the 900 MHz carrier example, the two paths provide effective scatterers that are unresolvable in time.

3.2 Propagation along a rough surface

In this section, reflection from a rough surface is treated by modification to the planar, smooth-surface formulation of the previous section. Here, surface roughness is considered to mean small-scale, random fluctuations in the local height. This is in contrast to large, deterministic variations in the height, as in a hilly terrain which is treated in Section 3.3. The treatment draws heavily on established texts; in particular, Beckmann and Spizzichino (1963), Bass and Fuks (1979), DeSanto and Brown (1986), Ogilvy (1991), and Blaunstein (2000).

In the image model of Section 3.1, the scattering from the smooth surface was expressed by a simple Fresnel reflection coefficient. When the surface is rough, it is better to interpret the reflection as a continuous process with scattering from the many small parts of the surface. In general, this is an exceedingly complicated problem involving shadowing and multiple scattering. However, some approximate solutions will be sufficient for showing the main effect of surface roughness for mobile communications situations. Kirchhoff's solution is presented first, since it highlights the problem through its limitations.

3.2.1 Kirchhoff's solution for scattering from a rough surface

Consider the situation in Figure 3.2.1 where the surface is random in nature with an average height of $z = 0$. The Kirchhoff solution applies when there is no shadowing, no multiple scattering, and when the surface is locally smooth, i.e. no sharp edges and corners. In the following it is understood, unless otherwise stated, that the value of the grazing angle is tempered by the nature of the surface roughness, in order to avoid shadowing. Under the above conditions, the total field on the surface is well described by the incident field and the local reflected field. For a complete discussion the reader is referred to Beckmann and Spizzichino (1963); here, the emphasis is on the physical interpretation.

The Kirchhoff approximation starts with the exact formulation of the Helmholtz scattering formula. This expresses the scattered field above a surface as an integral of known and unknown quantities over the surface S. The approximation comes in through the way we approximate the unknown quantity, the surface field E_S. The scalar version for the scattered field is

$$E_{\text{scat}} = \int_S \left(E_S \frac{\partial \Psi}{\partial n} - \Psi \frac{\partial E_S}{\partial n} \right) \mathrm{d}S, \tag{3.2.1}$$

where $\Psi = \mathrm{e}^{-\mathrm{j}kR_2}/(4\pi R_2)$ is the free-space Green's function, R_2 being the distance from the scattering point to the receiving antenna, and the integration is over the whole surface.

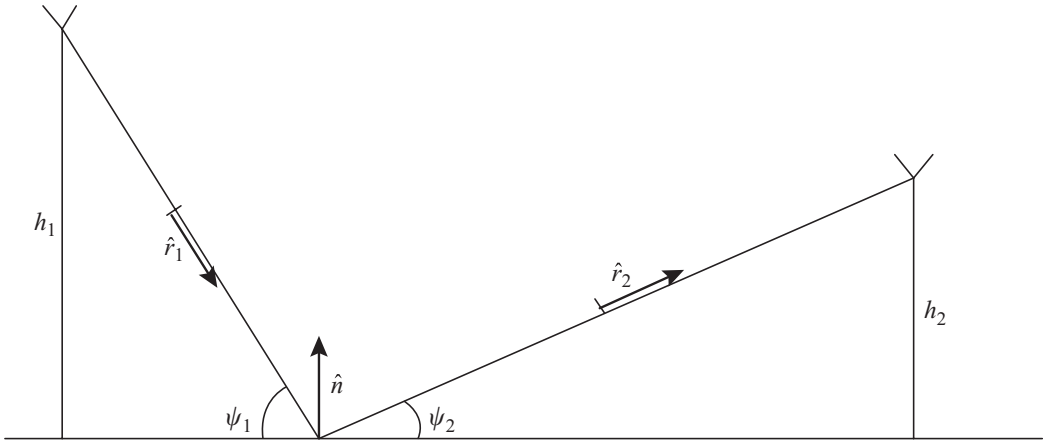


Figure 3.2.1 Geometric relations for an incident wave along $\hat{r}_1$ with associated transmitter height h_1 and grazing angle ψ_1, with a reflection along $\hat{r}_2$ defined by ψ_2 for a two-dimensional surface, propagating to a receiving height h_2, and where $\hat{n}$ is the normal to the local surface. R_1, R_2 are the general distances along $\hat{r}_1$ and $\hat{r}_2$, and R_{10}, R_{20} are the values of R_1, R_2 in the plane of incidence, $y = 0$

In accordance with the Kirchhoff approximation, it is now assumed that there is a smooth and locally plane surface, and that the total field on the surface only depends on the local properties, i.e. the local reflection coefficients. It is clear that we are then neglecting multiple scattering, and that we need to 'see' the whole surface from the antenna, i.e. no shadowing. Thus

$$E_{\mathrm{S}} = (1 + R_{\mathrm{v,h}})E_{\mathrm{inc}} = (1 + R_{\mathrm{v,h}})\frac{\mathrm{e}^{-\mathrm{j}kR_1}}{R_1} \tag{3.2.2}$$

where 'v, h' relates to the relevant polarisation, vertical or horizontal, including the local material parameters and local angle of incidence. The derivative with respect to the local normal is found to be

$$\frac{\partial E_{\mathrm{S}}}{\partial n} = -\mathrm{j}k(1 - R_{\mathrm{v,h}})\hat{r}_1 \cdot \hat{n}E_{\mathrm{inc}}. \tag{3.2.3}$$

It is interesting that the two terms in the Helmholtz integral depend so differently on the reflection coefficient and on the angle of incidence ($\hat{r}_1 \cdot \hat{n}$). For vertical polarisation and perfect conductivity $R_{\mathrm{v}} = +1$, and the second term of equation (3.2.1) disappears. For small angles of incidence, $R_{\mathrm{v,h}}$ is close to -1 for most natural media as discussed earlier. This approximation will be used in the following in which case the first term in equation (3.2.1) disappears, resulting in

$$E_{\mathrm{scat}} = \int_S \left(-\Psi\frac{\partial E_{\mathrm{S}}}{\partial n}\right)\mathrm{d}S = \frac{\mathrm{j}k}{2\pi}\int_S (\hat{r}_1 \cdot \hat{n})\frac{\mathrm{e}^{-\mathrm{j}k(R_1+R_2)}}{R_1R_2}\,\mathrm{d}S. \tag{3.2.4}$$

If it is assumed that the surface does not vary in the y-direction (a two-dimensional rough surface) the integral reduces further. The transverse integration (i.e. in the y-direction) can then be done analytically, using

$$\begin{aligned} R_1 &= R_{10}\left(1+\frac{y^2}{R_{10}^2}\right)^{1/2} \simeq R_{10}+\frac{y^2}{2R_{10}} \\ R_2 &= R_{20}\left(1+\frac{y^2}{R_{20}^2}\right)^{1/2} \simeq R_{20}+\frac{y^2}{2R_{20}}, \end{aligned} \tag{3.2.5}$$

where R_{10} and R_{20} are the values of R_1 and R_2 in the $y = 0$ plane, the plane of incidence. In an asymptotic evaluation, k large, the phase terms will dominate over the amplitude terms, so the following line integral results:

$$E = \frac{\mathrm{e}^{\mathrm{j}\pi/4}}{\sqrt{\lambda}}\int_L (\hat{r}_1\cdot\hat{n})\frac{\mathrm{e}^{-\mathrm{j}k(R_1+R_2)}}{\sqrt{R_{10}R_{20}(R_{10}+R_{20})}}\,\mathrm{d}l. \tag{3.2.6}$$

The dot product is the projection of the incident wave direction onto the normal of the local surface. For small slopes of the surface, this may be approximated as

$$\begin{aligned} \hat{r}_1\cdot\hat{n} &= -\sin\psi_1 \simeq \frac{z-h_1}{x}-z' \\ &= -\frac{h_1}{x}+\left(\frac{z}{x}-z'\right), \end{aligned} \tag{3.2.7}$$

where $z' = \mathrm{d}z/\mathrm{d}x$. In the second line of equation (3.2.7), smooth (i.e. flat) surface and rough surface parts have been separated. The smooth surface can be interpreted as the mean of the rough surface.

In a shadow region, the grazing angle becomes negative (Figure 3.2.2) and the expression (3.2.7) changes sign, indicating shadowing. For the low grazing angles characterising propagation along the surface, this sign change will occur often, even for moderate heights and slopes. Kirchhoff theory cannot be expected to give good results in these cases where shadowing occurs, because the assumptions on the surface field are not fulfilled.

3.2.1.1 Phase-only effects

Emphasis is on the phase terms in this section, since the Kirchhoff solution fails anyway in situations with strong amplitude effects like shadowing. The basic idea is that the incident-plus-locally-reflected field determines the distribution of effective sources over the surface, i.e. a physical optics approximation. Integration over these surface fields then gives the scattered field. Neglecting constants and slowly varying factors, the incident (scalar) field at the position (x,z) is just the phase term

$$E_{\mathrm{inc}} \approx \mathrm{e}^{-\mathrm{j}kR_1}, \tag{3.2.8}$$



Figure 3.2.2 The local surface normal on a rough surface. Note that the grazing angle becomes negative in a shadow region, where the Kirchhoff theory is not valid

where

$$R_1^2 = (h_1 - z(x))^2 + x^2. \tag{3.2.9}$$

Introducing an average fixed distance x_0 and expanding around it by introducing x' gives the equations

$$x' = x - x_0$$
$$R_{01}^2 = (h_1 - z(x_0))^2 + x_0^2.$$

If the height fluctuations are small relative to R_{01} , then

$$\begin{aligned} R_1 &\approx R_{01} + \frac{x' x_0}{R_{01}} - \frac{h_1 z}{R_{01}} \\ &= R_{01} + \cos\psi_1\, x' - \sin\psi_1\, z(x'), \end{aligned} \tag{3.2.10}$$

where ψ_1 is the average angle from the surface scattering section to the antenna. Similarly,

$$R_2 \approx R_{02} - \cos\psi_2\, x' - \sin\psi_2\, z(x'). \tag{3.2.11}$$

The essence of the Kirchhoff integral (see equation (3.2.6)) is

$$E_{\text{scat}} = \int_L \mathrm{e}^{-\mathrm{j}k(R_1+R_2)}\, \mathrm{d}x \tag{3.2.12}$$

where now

$$R_1 + R_2 \approx R_{01} + R_{02} + x' (\cos\psi_1 - \cos\psi_2) - z(x') (\sin\psi_1 + \sin\psi_2)\,. \tag{3.2.13}$$

It is readily seen that for the smooth surface ($z = 0$), the phase of the integrand is constant (stationary) for $\psi_1 = \psi_2$. This is the point of specular reflection, and valid approximately for a region around the point corresponding to the first Fresnel zone on the surface. Thus the integral will have a large contribution from this region, corresponding to a large specular reflection. A stationary-phase integration will lead to the smooth-surface reflected field of Section 3.1.

3.2.1.2 Numerical illustrations

An example of the integration result is illustrated in Figure 3.2.3(a), which shows the real part of equation (3.2.12) for $x = 500\,\text{m}$ as a function of the upper limit of the integral. The point of stationary phase is at $x = 417\,\text{m}$. Defining the limits of the first Fresnel zone where the additional path length is $\lambda/4$ gives $x_\text{F} = 227\,\text{m}$ and 484 m. Outside the first Fresnel zone, the scattered fields from the successive Fresnel zones essentially cancel as the successive contributions are in alternating phase. This manifests itself as the oscillatory character of the Kirchhoff integrand.

3.2.1.3 Effect of surface roughness on the specularly reflected field

Introduction of the random surface will have two effects on the Kirchhoff integral. The coherent phase addition of all the contributions from inside the first Fresnel zone will be partially destroyed by the random surface, and similarly, the cancellation of scattered field contributions from successive Fresnel zones will no longer take place to the same degree (Figure 3.2.3(b)). It is now shown how the Fresnel concept loses some of its significance depending on the size and nature of the surface roughness.

3.2.1.4 Specular scattering with small roughness factor

Assuming a small roughness factor so that the expansion of equation (3.2.13) is valid, and considering only the specular direction, $\psi_1 = \psi_2$, the scattered scalar field becomes

$$\begin{aligned} E_\text{scat} &= \mathrm{e}^{-\mathrm{j}k(R_{01}+R_{02})} \int_L \mathrm{e}^{-\mathrm{j}2k\sin\psi\ z(x)}\,\mathrm{d}x \\ &= \mathrm{e}^{-\mathrm{j}kR_\text{r}} \int_L \mathrm{e}^{-\mathrm{j}k_z z(x)}\mathrm{d}x. \end{aligned} \tag{3.2.14}$$

Here, $R_\text{r} = R_{01} + R_{02}$ is the distance to the image point (since we have chosen $\psi_1 = \psi_2$), and $k_z = 2\,k\sin\psi$ is twice the vertical component of the wave vector. The exponential of the integrand is expanded as

$$\mathrm{e}^{-\mathrm{j}k_z z} \approx 1 - \mathrm{j}k_z z - \tfrac{1}{2}k_z^2\,z^2, \tag{3.2.15}$$

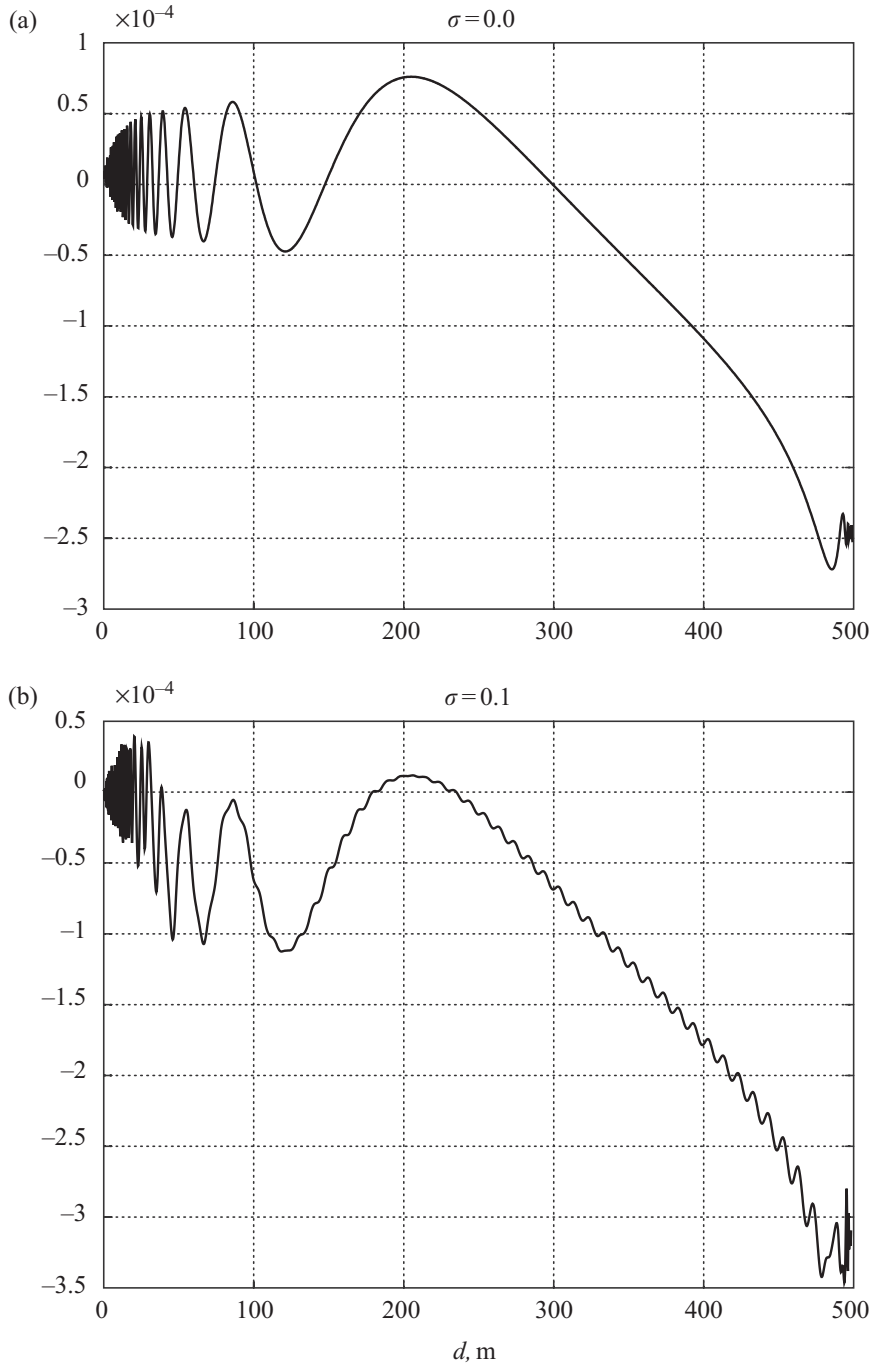


Figure 3.2.3 Illustration of the stationary-phase principle, with the integral in equation (3.2.12) as a function of its upper limit, in (a) a smooth surface where the integrand oscillates outside the first Fresnel zone and does not contribute to the integral, (b) a periodic two-dimensional sinusoidal surface with amplitude $\sigma = 0.1$ m and period 10 m, where the phase cancellations are partly destroyed

allowing a statistical approach for expressing the mean scattered scalar field:

$$\begin{aligned}\langle E_{\text{scat}}\rangle &\approx \int_L \left(1 - \mathrm{j}k_z\langle z\rangle - \frac{1}{2}k_z^2\langle z^2\rangle \cdots\right)\mathrm{d}x \\ &\approx \int_L \left(1 - \frac{1}{2}k_z^2\sigma^2\right)\mathrm{d}x \\ &\approx \left(1 - \frac{1}{2}k_z^2\sigma^2\right)E_{\text{smooth}};\end{aligned} \tag{3.2.16}$$

where σ is the standard deviation of the random surface and the mean surface height is zero. E_{smooth} represents the field reflected from a smooth surface and is included explicitly to allow comparison with the rough surface case. So from equation (3.2.16), the first-order effect of the small surface roughness is a reduction in amplitude and no change in phase in the average scattered field.

3.2.1.5 Specular scattering from a surface with statistically described height fluctuations

By assuming that a probability density function $p(z)$ can describe the height fluctuation in the surface roughness, the above result can be generalised. This paves the way for incoherent scattering.

Treating z as a random variable, the surface position dependence is removed from the solution of equation (3.2.14), which becomes

$$\langle E_{\text{scat}}\rangle = E_{\text{smooth}}\langle \mathrm{e}^{-\mathrm{j}k_z z}\rangle. \tag{3.2.17}$$

The average scattered field is thus the same as in the smooth surface case, but multiplied by a complex term which is the mean of the fluctuating phasor in the specular direction. This term is determined by the characteristic function $\tilde{p}(\kappa)$ of the surface height distribution, namely

$$\tilde{p}(\kappa) = \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{j}\kappa z}p(z)\,\mathrm{d}z. \tag{3.2.18}$$

The transform variable κ has dimension of inverse length so it is natural to call it a wavenumber. The expression $\tilde{p}(\kappa)$ is also seen to be the spectrum of the surface. The mean of the fluctuating phasor is then simply

$$\langle \mathrm{e}^{-\mathrm{j}k_z z}\rangle = \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{j}k_z z}p(z)\,\mathrm{d}z = \tilde{p}(k_z). \tag{3.2.19}$$

3.2.1.6 Gaussian distributed surface height

It is common to take the zero-mean Gaussian function as an example,

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma}e^{-z^2/(2\sigma^2)}. \tag{3.2.20}$$

For the zero-mean Gaussian pdf, the characteristic function is also Gaussian,

$$\tilde{p}(\kappa) = \mathrm{e}^{-\kappa^2\sigma^2/2}. \tag{3.2.21}$$

Thus for a Gaussian height distribution of the surface, equation (3.2.17) becomes the well-known formula for the scalar reflection coefficient of a rough surface,

$$\begin{aligned} R_{\text{rough}} &= R_{\text{smooth}}\, \mathrm{e}^{-\frac{(k_z\sigma)^2}{2}} \\ &= R_{\text{smooth}}\, \mathrm{e}^{-2(k_0\sigma\sin\psi)^2}. \end{aligned} \tag{3.2.22}$$

Note that this is in agreement with equation (3.2.16) in the case of small roughness. Equation (3.2.16) required no assumptions about the surface height following a specific probability distribution, it just required that the height fluctuations be small. The term

$$k_z\sigma = 2k_0\sigma\sin\psi \tag{3.2.23}$$

is called the Rayleigh roughness factor. It represents the electrical path difference between specular rays reflected from horizontal surface heights differing by one standard deviation of the height distribution.

The formula (3.2.22) is often used in engineering practice, but it should be noted that its validity is limited to the effect on the specular scalar field only and assumes: that the surface height distribution is Gaussian; that there is no shadowing; and that the correlation length of the surface is large (see below).

3.2.1.7 Numerical examples of specular scattering from a rough surface

Table 3.2 gives some indicative results for the effect on the specular scattering relative to that from a smooth surface.

The most significant trend is the enormous variation between the normal incidence and grazing incidence cases. Care must be taken in interpreting the results, or rather the validity of the theory with respect to the geometry of the cases. Specifically, recall that when shadowing occurs, the theory breaks down, so the slopes of the surface relative to the incident wave must now be considered. This is addressed in the following sections where the perturbation theory is applied. As long as there is no shadowing, the above results are valid, independent of the slopes of the surface.

To complete this section results from another simple distribution are presented to get a feel for the effect on the scattering. Table 3.3 shows analogous results for a uniform distribution, where the surface heights vary between $-h/2$ and $h/2$. The standard deviation is $\sigma = h/\sqrt{12} = 0.289\,\text{h}$.

As was shown earlier the reduction of the specular reflection due to roughness is independent of the actual distribution for small values of the Rayleigh parameter. This follows from the tables which have a common row with $\sigma = 2.89\,\text{cm}$ $(10/\sqrt{12})$; only for 90°, normal incidence, is there a small difference.

Table 3.2 Change (in dB) of the reflection coefficient for a Gaussian distributed surface for different grazing angles and roughness at $f = 900$ MHz. The table values are valid as long as the correlation length is large (see text)

σ	$\psi = 90°$	45°	10°	1°
100 cm	−6172	−3086	−186	−1.88
50 cm	−1544	−772	−46	−0.48
10 cm	−62	−31	−2	−0.02
2.89 cm	−5.1	−2.6	−0.16	−0.02
1 cm	−0.62	−0.31	−0.02	∼0

Table 3.3 Change (in dB) of the reflection coefficient for a uniformly distributed surface for different grazing angles and roughness, where $f = 900$ MHz

	$\psi = 90°$	45°	10°	1°
$h = 100$ cm $\sigma = 29$ cm	−328	−25.7	−27.9	−0.16
$h = 50$ cm $\sigma = 14.5$ cm	−328	−25.1	−4.3	−0.04
$h = 10$ cm $\sigma = 2.89$ cm	−5.9	−2.7	−0.16	∼0
$h = 1$ cm $\sigma = 2.89$ mm	−0.5	−0.03	∼0	∼0

3.2.1.8 Limitations of the Kirchhoff model

To summarise, the reflected specular wave is reduced *on average* when the reflection takes place over a rough surface. The average is to be interpreted as an ensemble average over many surfaces obeying the same statistics, or a time average for a changing surface such as water. The formulation may also be used in directions other than the specular direction, in which case the term *coherent* field is used. The reason is that the phase is preserved, and averaging over many realisations will give a non-zero answer. The random phases that we have not considered give rise to zero mean fields, which have a certain intensity, and this is called the *diffuse* or *incoherent* field. The

situation is analogous to the coherent signal and noise in signal theory, zero-mean Gaussian noise, but a noise power exists; see Figure 3.2.4(a) and (b).

It is clear from equation (3.2.22) that for values of the Rayleigh roughness factor larger than about one, the specular reflections disappear completely. In this case, all

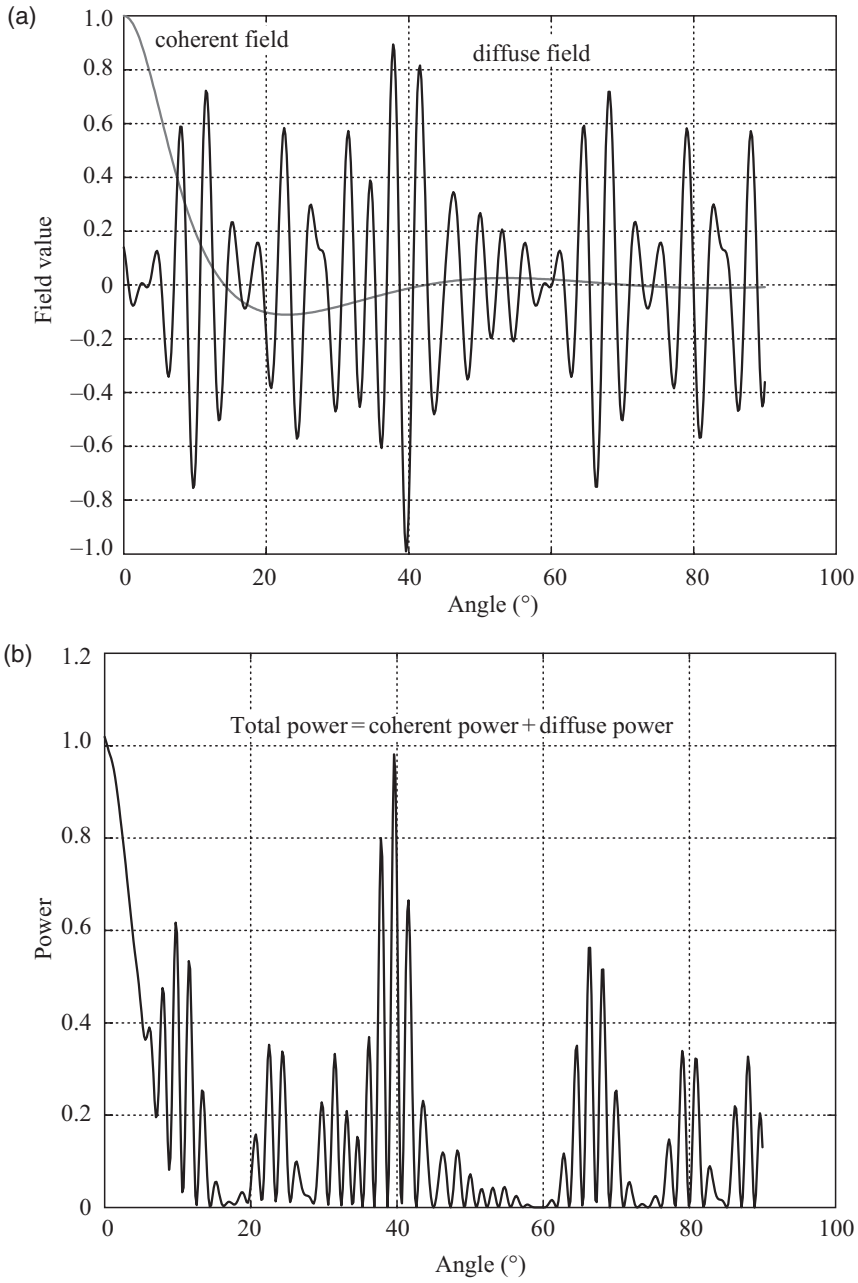


Figure 3.2.4 Illustrating the difference between the coherent field and the diffuse field. The diffuse field has zero mean (a) when averaged over many realisations of surfaces. Nevertheless, for any one realisation the diffuse field may add considerably to, or even dominate, the total scattered power (b)

the scattered power is in the diffuse field, which is addressed in Section 3.2.3. The main problem with the Kirchhoff model is that it goes wrong when shadowing is involved, and this is often the case for the low angles involved in mobile communications. For the low angles we therefore treat another theory, the perturbation theory, which does include shadowing, although only for small roughness.

Horizontal scaling of the surface roughness is absent in the above treatment of the coherent field. The reason follows from equation (3.2.13), where it is shown that for the specular direction ($\psi_1 = \psi_2$) the x-dependence disappears. However, when the average power is involved away from the specular direction, the horizontal dependence matters, as will be clear in the following.

3.2.2 Perturbation theory for the field scattered from a rough dielectric medium

3.2.2.1 Statistics of horizontal properties of a rough surface

To progress from the Kirchhoff theory, more information is required on the statistics of the surface – primarily its horizontal correlation properties. A horizontal correlation function for the surface height fluctuation is defined (DeSanto and Brown, 1986) as

$$\sigma^2 W(\boldsymbol{x}) = \langle z(\boldsymbol{x}_0) z(\boldsymbol{x} + \boldsymbol{x}_0) \rangle \tag{3.2.24}$$

with its spectrum

$$S(\boldsymbol{k}) = \sigma^2 \iint W(\boldsymbol{x}) \mathrm{e}^{-\mathrm{j}\boldsymbol{k}\cdot\boldsymbol{x}} \mathrm{d}\boldsymbol{x}, \tag{3.2.25}$$

where $\boldsymbol{x}$ is a transverse vector (x,y) and $\boldsymbol{k}$ is the transform variable (k_x, k_y). These functions gauge how rapidly the height fluctuations vary along the surface. In analogy with the time–frequency relationships (see Appendix B), S is the spatial power spectral density of the surface height fluctuations.

Assuming that the surface roughness does not have any preferred directions, W will depend on distance, not on angle. The correlation function is then written $W(\rho)$ and is characterised by a correlation length L, where the correlation coefficient function has decayed to some value, say 0.5. The results for an isotropic surface are stated below without derivation, after Bass and Fuks (1979).

3.2.2.2 Effective surface parameters for coherent field

The effect of a small degree of surface roughness on the average specular field can be couched as an *effective reflection coefficient*, or an *equivalent surface impedance* (or admittance). This effective surface impedance may have a real part corresponding to a loss of power from the specular field to the diffuse field, and an imaginary part corresponding to a reactive, stored energy near the surface. The results are complicated and only the asymptotic cases are presented, corresponding to small and large correlation lengths, and non-grazing and grazing incidence.

3.2.2.3 Horizontal polarisation: asymptotic cases, isotropic surface

For horizontal polarisation the equivalent relative surface impedance is

$$\frac{1}{\sqrt{\epsilon_{\text{eff}}}} = \frac{1}{\sqrt{\epsilon_{\text{r}}}} + \text{j}\frac{k\sigma^2}{2L}\int_0^\infty \frac{\text{d}x}{x}\frac{\text{d}\hat{W}(x)}{\text{d}x}, \quad kL \ll 1; \tag{3.2.26a}$$

$$= \frac{1}{\sqrt{\epsilon_{\text{r}}}} + (k\sigma)^2 \sin\psi, \quad kL \gg 1, \ \ \psi \gg \frac{1}{\sqrt{kL}}; \tag{3.2.26b}$$

$$= \frac{1}{\sqrt{\epsilon_{\text{r}}}} - (k\sigma)^2 \text{e}^{-\text{j}\pi/4}\frac{\sqrt{2}}{\sqrt{\pi kL}}\int_0^\infty \frac{\text{d}x}{\sqrt{x}}\frac{d\hat{W}(x)}{\text{d}x}, \quad \sqrt{kL} \gg 1, \ \ \psi \ll \frac{1}{\sqrt{kL}}; \tag{3.2.26c}$$

where $\hat{W}(x)$ is the correlation function with normalised argument, $\hat{W}(\rho/L)$. Case (a) is for short surface correlation lengths – many surface height deviations per wavelength, (b) is for long correlation lengths relative to wavelength and non-grazing wave incidence (the incident angle may be small, but not too small, depending on the surface correlation length, in order to avoid shadowing), and (c) is for long correlation lengths and grazing incidence.

A practical example may illustrate the different situations. Consider the side of a brick-walled building with balconies, doors and windows. The wall itself would probably have a correlation length of a few centimetres ($kL \sim 0.2$ at 900 MHz), so version (a) should be used independent of angle. Note that in this region Kirchhoff failed, since ϵ_{eff} depends on L. Because the integral is negative ($W(x)$ has negative curvature near the origin) the effective permittivity increases in magnitude. For the larger structures, windows and doors, the correlation length would probably be of the order a metre, leading to $kL \sim 20$ at 900 MHz. Translating this into an angle we find that version (b) should be used for angles much larger than 10°, and (c) for angles much less than that. In both (b) and (c) the effective ϵ is reduced.

It is worthwhile comparing the result with that of the previous phase-only Kirchhoff solution for the case of small surface roughness (equation (3.2.16)), which can simplify as

$$\begin{aligned} R_{\text{rough}} &= R_h \text{e}^{-2(k\sigma\psi)^2} \\ R_{\text{rough}} &\approx -\left(1 - \frac{2\psi}{\sqrt{\epsilon}}\right)\left(1 - 2(k\sigma)^2\psi^2\right) \\ &\approx -\left(1 - 2\psi\left(\frac{1}{\sqrt{\epsilon}} + (k\sigma)^2\psi\right)\right) \\ &= -\left(1 - \frac{2\psi}{\sqrt{\epsilon_{\text{eff}}}}\right). \end{aligned} \tag{3.2.27}$$

This agrees exactly with equation (3.2.26b). So if the conditions of (3.2.26b) are not fulfilled, the Kirchhoff theory and the simple use of the Rayleigh condition fail.

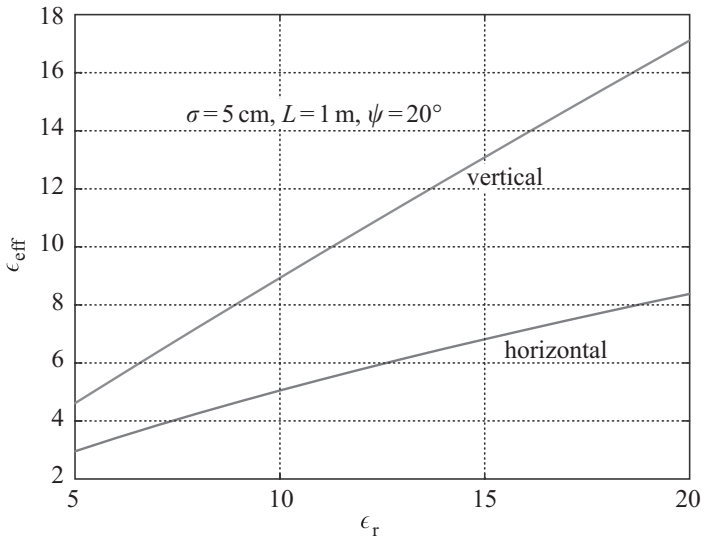


Figure 3.2.5 Effective permittivity for rough surface for a fixed angle of 20° at 900 MHz, $\sigma = 5$ cm, $L = 1$ m for vertical (upper curve) and horizontal (lower curve)

This was expected from the assumptions made for the Kirchhoff solution, but equations (3.2.26) quantify the Kirchhoff limitations. To reiterate, the Kirchhoff solution for the coherent case is only valid for large correlation lengths and not too small grazing angles, the actual value depending on the correlation length.

It is apparent from equation (3.2.26c) that the effective permittivity of the rough surface for horizontal polarisation is reduced. Figure 3.2.5 shows some examples for $f = 900$ MHz and $\psi = 20°$; for lower angles the equation is invalid for the chosen correlation length.

Returning to equations (3.2.26), the change in surface impedance for horizontal polarisation is strictly reactive for short correlation lengths and all incident angles, resistive for large correlation lengths and large angles, and a mixture of both for near grazing incidence. An important point is that the Rayleigh roughness factor has restricted validity.

3.2.2.4 Vertical polarisation: asymptotic cases, isotropic surface

For vertical polarisation, the relative surface impedance is

$$\begin{aligned}
\frac{1}{\sqrt{\epsilon_{\text{eff}}}} &= \frac{1}{\sqrt{\epsilon_{\text{r}}}} + \text{j}\frac{k\sigma^2 \sin(2\psi)}{2L}\int_0^\infty \frac{\text{d}x}{x}\frac{\text{d}\hat{W}(x)}{\text{d}x}, \quad kL \ll 1; \\
&= \frac{1}{\sqrt{\epsilon_{\text{r}}}} + (k\sigma)^2 (\sin\psi)^3, \quad kL \gg 1, \ \ \psi \gg \frac{1}{\sqrt{kL}}; \\
&= \frac{1}{\sqrt{\epsilon_{\text{r}}}} - \frac{\sigma^2}{2L^2}\frac{\sqrt{kL}}{\sqrt{2\pi}}\text{e}^{\text{j}\pi/4}\int_0^\infty \frac{\text{d}x}{x^{3/2}}\frac{\text{d}\hat{W}(x)}{\text{d}x}, \quad kL \gg 1, \ \ \psi \ll \frac{1}{\sqrt{kL}};
\end{aligned} \tag{3.2.28}$$

where the three cases are similar to those for the horizontal polarisation. The same general behaviour occurs as for horizontal polarisation, except that now the Rayleigh roughness factor is never valid (cf. condition (3.2.26b) for the horizontal polarisation). The influence of the roughness is much less than for the horizontal polarisation, and the frequency dependence is different.

To summarise, for grazing incidence the specular reflection coefficient from a rough surface tends to -1 in all cases, but the way it approaches this value depends on the polarisation, the surface correlation length and the degree of surface roughness.

3.2.3 Effect of surface roughness on diffuse field

The surface roughness also influences the contributions from outside the first Fresnel zone and gives rise to a randomly scattered contribution that is called the diffuse field. This is also referred to as clutter, especially in the radar discipline. Recall that for a smooth surface, the contributions from outside the first Fresnel zone essentially cancel owing to the phase inversions of the contributions from successive Fresnel zones. When the surface becomes rough, this cancellation decreases and there will be a more random addition of fields.

Within the bounds of the perturbation theory of the previous section, the intensity of the diffuse field may be evaluated. The average value over many realisations of the diffuse field is zero; however, the average power is non-zero. Quoting from Ogilvy (1991) the elemental power $\mathrm{d}J$ of the diffuse field u scattered from an elemental surface area $\mathrm{d}A$ is

$$\mathrm{d}J = \langle u \times u^* \rangle = \frac{k^4}{(2\pi)^2 R_1^2 R_2^2} \Phi(\psi_1, \psi_2, \psi_3) S(\boldsymbol{k}_2 - \boldsymbol{k}_1)\, \mathrm{d}A, \tag{3.2.29}$$

corresponding to an incident power of $1/R_1^2$. Figure 3.2.6 depicts the situation.

The incident energy propagating with wave vector $\boldsymbol{k_1}$ and decaying as $1/R_1^2$, is scattered from an area $\mathrm{d}A$ characterised by a spectral density S (equation 3.2.25) and roughness variance σ^2, scatters and propagates in the scattering direction $\boldsymbol{k_2}$, decaying as $1/R_2^2$. The scalar wavenumber is $k = |\boldsymbol{k_1}| = |\boldsymbol{k_2}|$. The Φ-function is to cater for effects of the boundary conditions on the surface and is discussed below.

Inclusion of the spectral density may be understood if we ask which wavenumber a scatterer should possess in order to interact in the best possible way to send a wave back to a receiver. The answer must be that the phases add up in that particular direction, a coherent addition of the fields, i.e.

$$\begin{aligned} &\mathrm{e}^{-\mathrm{j}\boldsymbol{k_{1h}}\cdot r}\mathrm{e}^{-\mathrm{j}\boldsymbol{k_{sc}}\cdot r} = \mathrm{e}^{-\mathrm{j}\boldsymbol{k_{2h}}\cdot r} \\ &\boldsymbol{k_{sc}} = \boldsymbol{k_{2h}} - \boldsymbol{k_{1h}}. \end{aligned} \tag{3.2.30}$$

If this condition is satisfied there is a type of spatial resonance, reflections from different parts of the surface adding in phase. This was a deterministic explanation; in the random case we are treating, the return depends on the spectral density of the surface at that wavenumber. A reflector with a normal in the direction of $\boldsymbol{k_{sc}}$ will

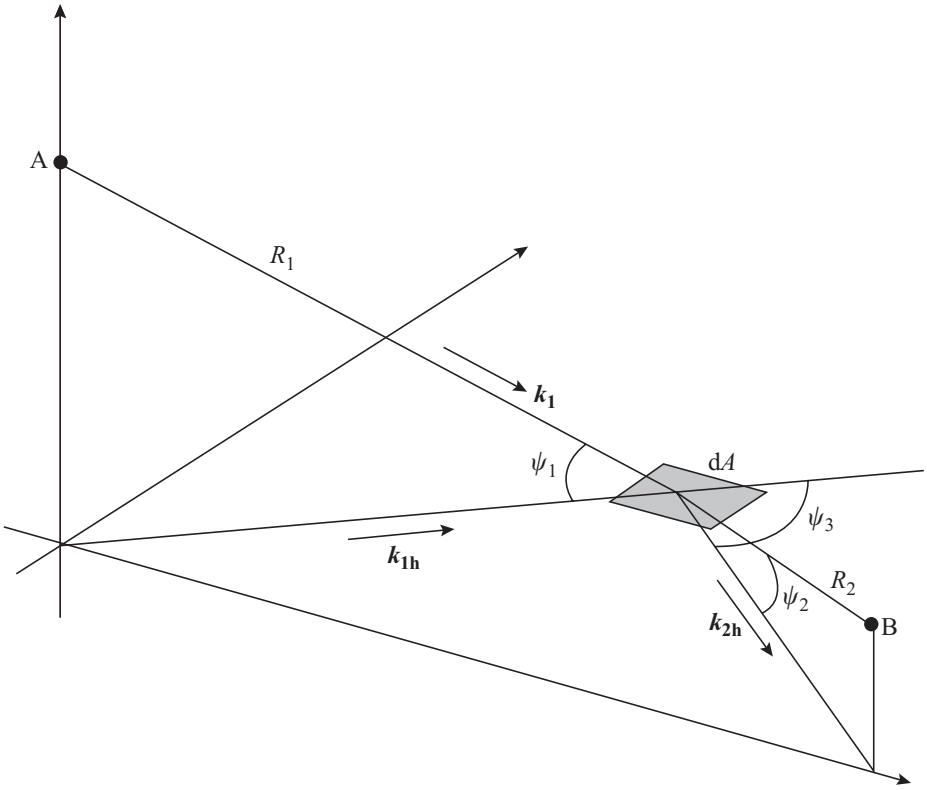


Figure 3.2.6 Relations for incoherent scattering from an elemental area of rough surface. The incident wave with wavevector $\boldsymbol{k}_1$ *from point A, scatters from area* dA *and propagates along* $\boldsymbol{k}_2$ *to point B. The horizontal components* $\boldsymbol{k}_{1\mathrm{h}}$, $\boldsymbol{k}_{2\mathrm{h}}$ *appear in the scattering equation*

also reflect in that direction, satisfying the condition of equal incidence and reflection angles (Figure 3.2.7). Taking as an example a Gaussian correlation function and corresponding spectrum,

$$\begin{aligned} \sigma^2 W(\rho) &= \frac{1}{L\sqrt{\pi}} \mathrm{e}^{-(\rho/L)^2} \\ S(k_{\mathrm{sc}}) &= \sigma^2 L\sqrt{\pi} \mathrm{e}^{-(k_{\mathbf{sc}}L)^2/4} \end{aligned} \tag{3.2.31}$$

it is clear that the largest contributions from a uniform surface are obtained for small values of $\boldsymbol{k}_{\mathbf{sc}}$.

In order to get an overview of the effect of the variation of $\boldsymbol{k}_{\mathbf{sc}}$ we assume that the grazing angles are so small that the wavevector equals its horizontal projection, and $\boldsymbol{k}_{\mathbf{sc}}$ is then simply found as the free-space wavenumber $\boldsymbol{k}_0$ times 2 cosine half the angle γ between the 'outgoing' $\boldsymbol{k}_{\mathbf{1}}$ and the incoming $\boldsymbol{k}_{\mathbf{2}}$ (Figure 3.2.7).

$$k_{sc} = 2k_0 \cos(\gamma/2) \tag{3.2.32}$$

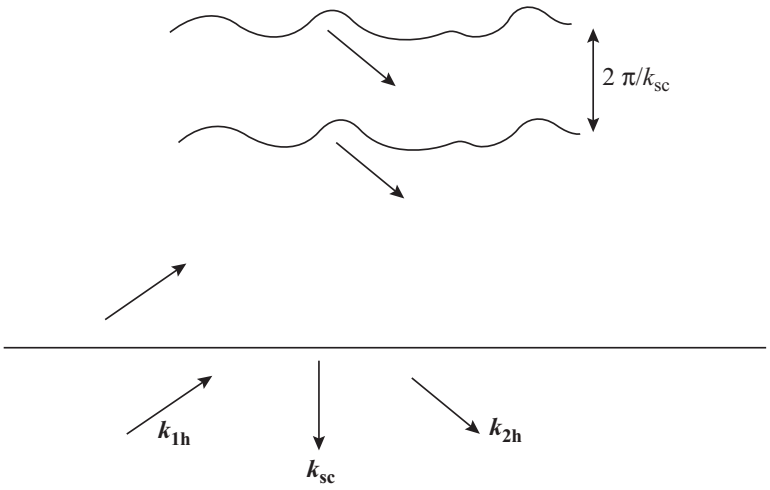


Figure 3.2.7 Scattering wave numbers

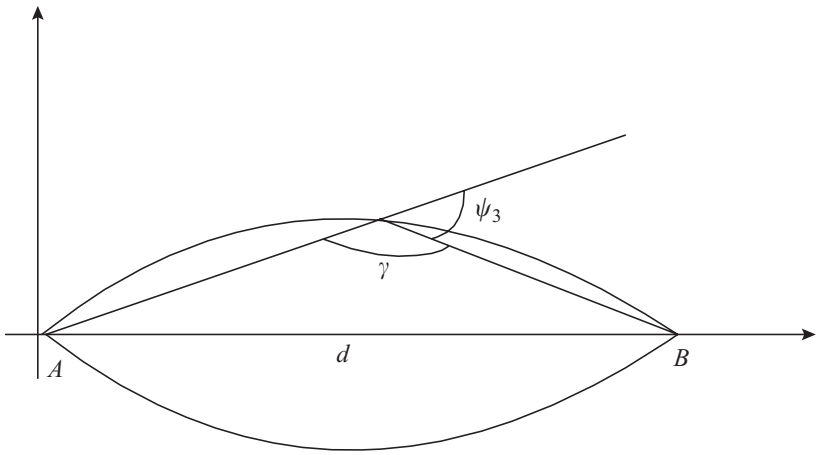


Figure 3.2.8 Curves of constant spectral contribution on the scattering surface

Curves of constant $\boldsymbol{k_{sc}}$ are thus circles (Figure 3.2.8) going through receiver and transmitter, parametric curves from which the baseline is seen under the angle γ.

Note that k_{sc} equals zero on the line between transmitter and receiver, and that it approaches $2k_0$ at distances large compared with d. At these large distances there is a backscattering situation while close to the baseline there is forward scatter.

Since the forward scattering is dominant, it deserves a closer look on the base line, and a more correct version without assuming small grazing angles is

$$k_{sc}/k_0 = \cos\psi_1 - \cos\psi_2 \tag{3.2.33}$$

with the conclusion that the scattering cross section (another word for $J/\mathrm{d}A$) is largest at the point of specular reflection, where $\psi_1 = \psi_2$. In fact if L is very large compared with the wavelength this is the only point contributing. It is interesting to compare with the coherent case. In the coherent case the main contributions to the specular

component come from an area around the point of reflection, a rather narrow zone (Fresnel zone) defined by phase differences. In the incoherent case the main contributions also come from a rather narrow zone around the point of reflection, defined by the spectrum of the surface. The distance factors of course also have a major influence ($1/R^4$), so the contributions near the lower antenna will dominate. Later this relationship will be used to study path loss over rough terrain based on this single scattering theory. In many situations in practice we will also have multiple scattering, which is a major complication in the theories and not accounted for here.

3.2.3.1 Polarisation aspects of rough surface scattering

Electromagnetic scattering is a vector process, so the polarisation of the field will in general be changed on reflection from a surface. A simple case is the reflection of an arbitrary polarised wave from a smooth surface with given material parameters. The wave may be decomposed into horizontal and vertical components relative to the surface, and each component is reflected independently with different reflection coefficients. Put together again for the composite field, the polarisation has changed. Similarly, it can be stated that if the fields are purely horizontal or vertical (the wavevectors are in the plane of the surface or in a plane normal to it) then the surface will not depolarise the field.

For a rough surface the situation is quite complicated. The results for a general dielectric surface (Valenzuela, 1967) will be quoted, where the Φ-factors are given by

$$\Phi = 4 \sin^2 \psi_1 \sin^2 \psi_2 Q_{xx}^2 \tag{3.2.34}$$

$$Q_{\mathrm{hh}} = \frac{(\epsilon_{\mathrm{r}} - 1) \cos \psi_3}{(\sin \psi_1 + \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_1})(\sin \psi_2 + \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_2})} \tag{3.2.35}$$

$$Q_{\mathrm{hv}} = \frac{(\epsilon_{\mathrm{r}} - 1) \sin \psi_3 \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_2}}{(\sin \psi_1 + \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_1})(\epsilon_{\mathrm{r}} \sin \psi_2 + \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_2})} \tag{3.2.36}$$

$$Q_{\mathrm{vh}} = \frac{(\epsilon_{\mathrm{r}} - 1) \sin \psi_3 \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_1}}{(\epsilon_{\mathrm{r}} \sin \psi_1 + \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_1})(\sin \psi_2 + \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_2})} \tag{3.2.37}$$

$$Q_{\mathrm{vv}} = \frac{(\epsilon_{\mathrm{r}} - 1)(\epsilon_{\mathrm{r}} \cos \psi_1 \cos \psi_2 - \cos \psi_3 \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_1} \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_2})}{(\epsilon_{\mathrm{r}} \sin \psi_1 + \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_1})(\epsilon_{\mathrm{r}} \sin \psi_2 + \sqrt{\epsilon_{\mathrm{r}} - \cos^2 \psi_2})}. \tag{3.2.38}$$

It should be noted that the depolarisation is zero in the plane of incidence ($\psi_3 = 0$). For ϵ tending to infinity the expressions reduce to those above for perfect conductivity. Examples are given in Figure 3.2.9. In part (a) the forward scattering ($\psi_3 = 0$) is shown for horizontal and vertical polarisation as a function of ψ_1 ($= \psi_2$) for $\epsilon_r = 5$. The power goes to zero rapidly with ψ_1 with the horizontal polarisation dominating. Part (b) shows the off-axis scattering for $\epsilon_r = 5$ and $\psi_1 = \psi_2 = 10°$. The vertical

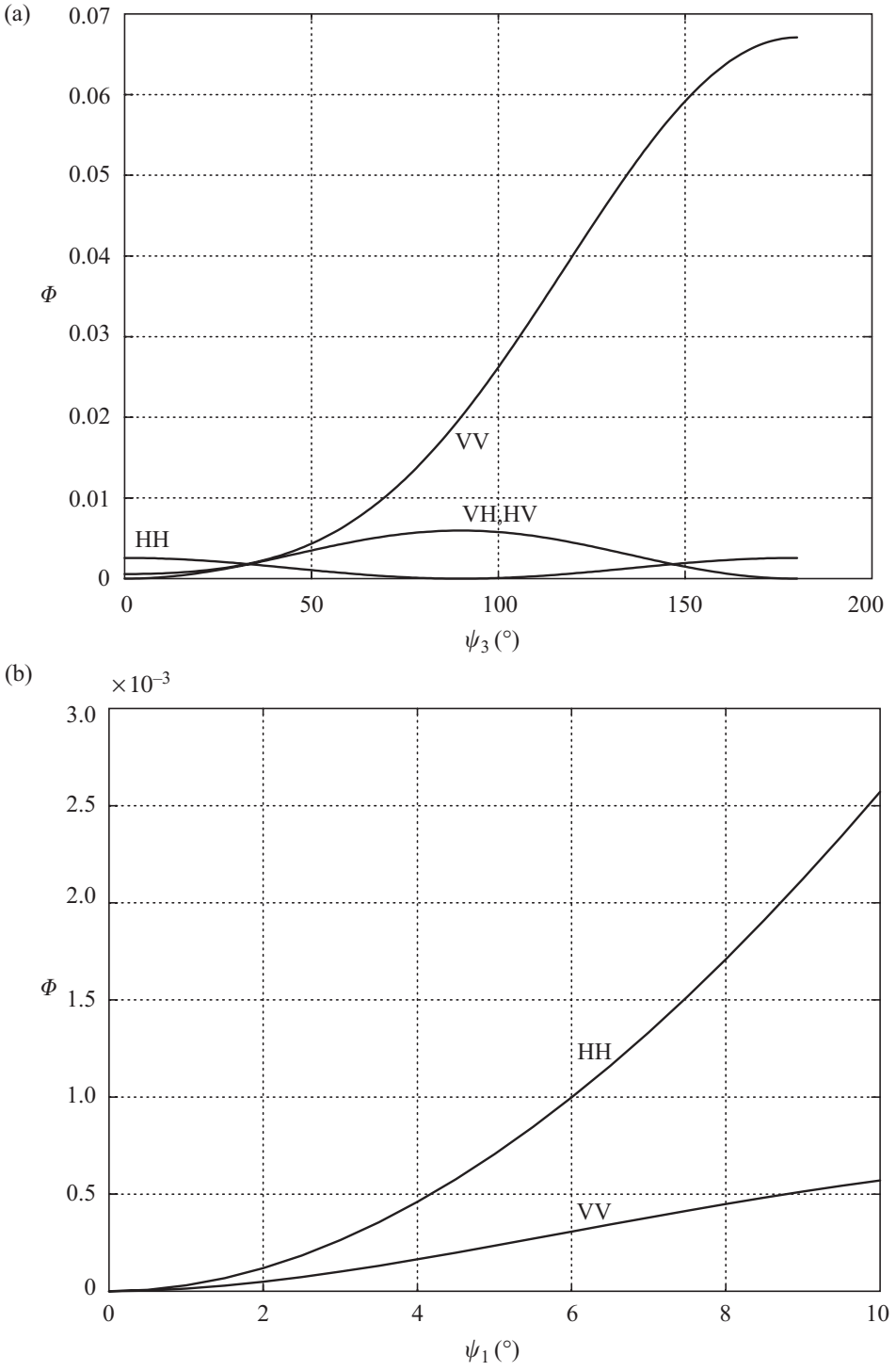


Figure 3.2.9 *(a) Φ values for forward scattering , $\epsilon_r = 5$. (b) Φ values for off-axis scattering, $\epsilon_r = 5$, $\psi_1 = \psi_2 = 10°$*

co-polarisation is dominating for large values of ψ_3 with a maximum value in the back-scattering direction. It should be recalled that it is not the total power, the spectral function S should also be considered, and this will have a damping effect in the off-axis scattering.

3.3 Diffraction

After reflection and scattering, diffraction is an important mechanism, and various diffraction theories are often used to explain propagation phenomena in mobile environments, or used in computer modelling. The treatment will be limited to rather simple situations; those that are important for mobile communications. These are propagation in hilly terrain, microcellular propagation in urban environments, and indoor situations. The diffraction is here concerned with abrupt changes such as a sharp edge or a corner. Since the concern is in the microwave region where the sizes of obstacles tend to be many wavelengths, it is natural to concentrate on high-frequency approximations such as geometrical optics (GO), physical optics (PO), geometrical theory of diffraction (GTD), and uniform theory of diffraction (UTD), which are all explained in the following. The basic problem in diffraction is best explained by a simple example, an absorbing screen illuminated by an antenna (Figure 3.3.1).

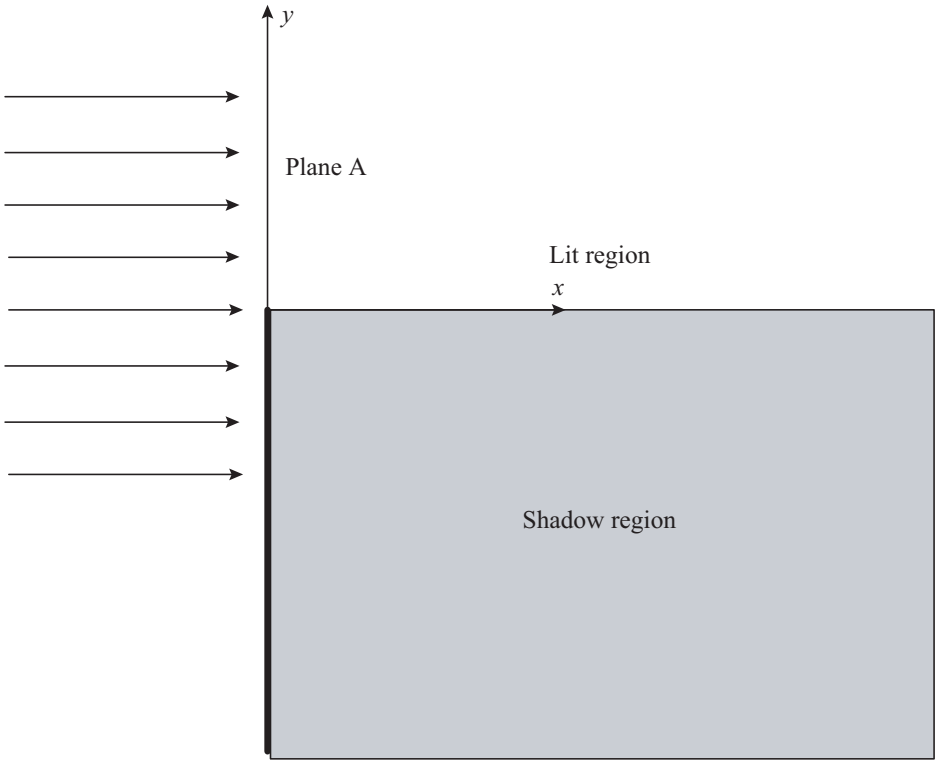


Figure 3.3.1 Diffraction of a plane electromagnetic wave by an absorbing screen. Space is divided into a lit region and a shadow region. Diffraction gives the continuous transition of the fields between the two regions

Huygens' principle gives a good explanation as to why the fields behind the screen are changed. A Kirchhoff integration over the plane A ($y > 0$) would give the field equal to the incident field if the screen was not there. The absorber removes (or sets to zero) part of the integration, and the remaining integration gives the field to the right of the screen, including the shadow region. Over plane A a PO approximation is applied, i.e. locally the field is equal to the incident field. The exact, total field above the screen depends on the presence of the screen; however, as seen below, the approximation is quite good. For an incident plane wave, the result is

$$E_{\text{total}} = \mathrm{e}^{-\mathrm{j}kx}\left(\frac{1}{2} - \frac{\mathrm{e}^{\mathrm{j}\pi/4}}{\sqrt{2}}\,\text{Fres}\left(-\frac{2y}{\sqrt{\lambda x}}\right)\right) \quad x > 0 \tag{3.3.1}$$

for an incident field of $E_{\text{inc}} = \mathrm{e}^{-\mathrm{j}kx}$. Fres($z$) is the Fresnel integral given by

$$\text{Fres}(z) = C(z) - \mathrm{j}\,S(z) = \int_0^z \mathrm{e}^{-\mathrm{j}(\pi/2)u^2}\,\mathrm{d}u \tag{3.3.2}$$

with the following values worth noting

$$\begin{aligned} \text{Fres}(\mathrm{y} = 0) &= 0, \quad \Rightarrow \quad E_{\text{total}} = \frac{E_{\text{inc}}}{2} \\ \text{Fres}(\mathrm{y} \to \infty) &= -\frac{\mathrm{e}^{-\mathrm{j}\pi/4}}{\sqrt{2}}, \quad \Rightarrow E_{\text{total}} = E_{\text{inc}} \\ \text{Fres}(\mathrm{y} \to -\infty) &= \frac{\mathrm{e}^{-\mathrm{j}\pi/4}}{\sqrt{2}}, \quad \Rightarrow E_{\text{total}} = 0, \end{aligned} \tag{3.3.3}$$

so at the shadow boundary the total field strength is exactly one half of the incident field strength (which is expected since we only integrate over half of the plane). In the optical limit with zero wavelength, the incident field is recovered in the lit region and total darkness occurs in the shadow zone. The amplitude of the total field is illustrated in Figure 3.3.2, indicating interference in the lit region.

In order to understand the phenomena more clearly, the diffracted field is *defined* as the difference between the total field and the geometrical optics (GO) field,

$$E_{\text{total}} = E_{\text{GO}} + E_{\text{d}}. \tag{3.3.4}$$

The GO field is discontinuous across the shadow boundary, since it is the high-frequency asymptotic limit. The diffracted field thus corrects the GO field for the finite frequency, and since the total field is continuous, the diffracted field must be discontinuous as well. The real and imaginary parts of normalised E_{d} are shown in Figure 3.3.3 clearly indicating the jump of unity in the real part.

It is also apparent that the diffracted field behaves as a wave-like field further away, when the dimensionless argument $2y/\sqrt{(\lambda x)}$ is numerically larger than one. This is the basis for the geometrical theory of diffraction (GTD), which describes in a rigorous way the diffraction rays emanating from edges and corners, but which is unable to

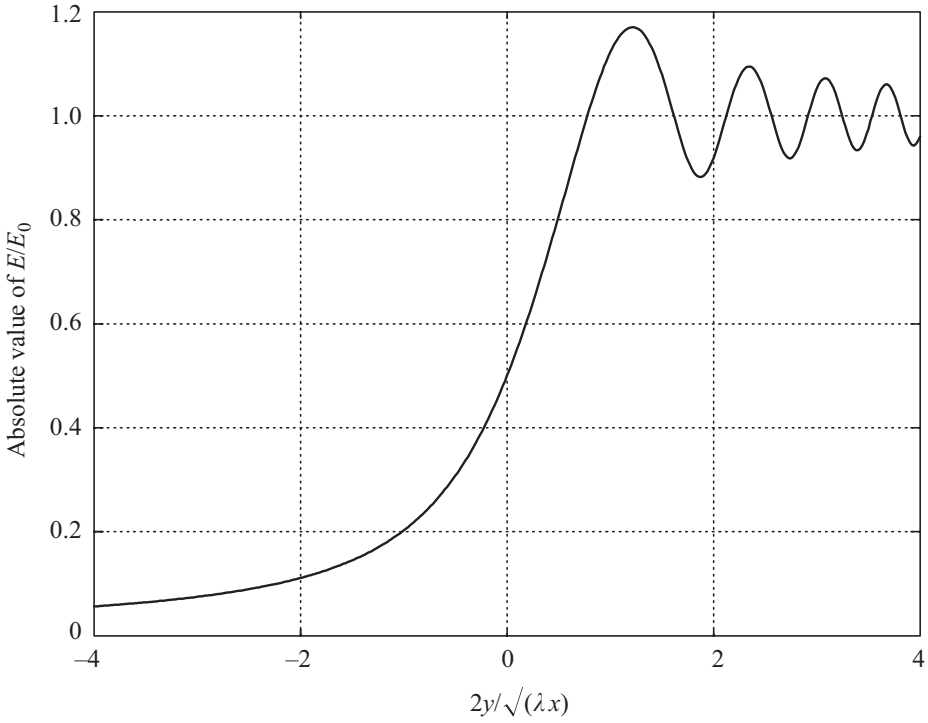


Figure 3.3.2 Absolute value of the total field across the shadow boundary between the shadow and the lit region for plane wave illumination with PO approximation. The abscissa is $2y/\sqrt{(\lambda x)}$*, the argument in the Fresnel integral*

describe the transition zone fields. This deficiency was repaired by Kouyoumjian and Pathak (1974), who kept the transition functions in the UTD, and also allowed for a source different from a plane wave. Since UTD reduces to GTD in the regions beyond the transition regions, the discussion is limited to UTD. The case of a wedge, which is a very useful component in treating hilly terrain or man-made urban structures, is treated first.

3.3.1 UTD diffraction by a wedge

Following Kouyoumjian and Pathak (1974) the wedge is described by two faces, an angle of the wedge, a source and an observation point. The two faces may be hard or soft (in an acoustic manner) with reflection coefficient $+1$ and -1, respectively, and the geometry is given in Figure 3.3.4.

The situation is now obviously much more complicated than in the simple half-plane case of Figure 3.3.1, but the main ingredients are the same. The case of normal incidence on the wedge axis is considered; for more complicated situations the reader is referred to the specialist literature such as McNamara *et al.* (1990). Antennas are placed at A and at B, and which one is radiating and which one is receiving does

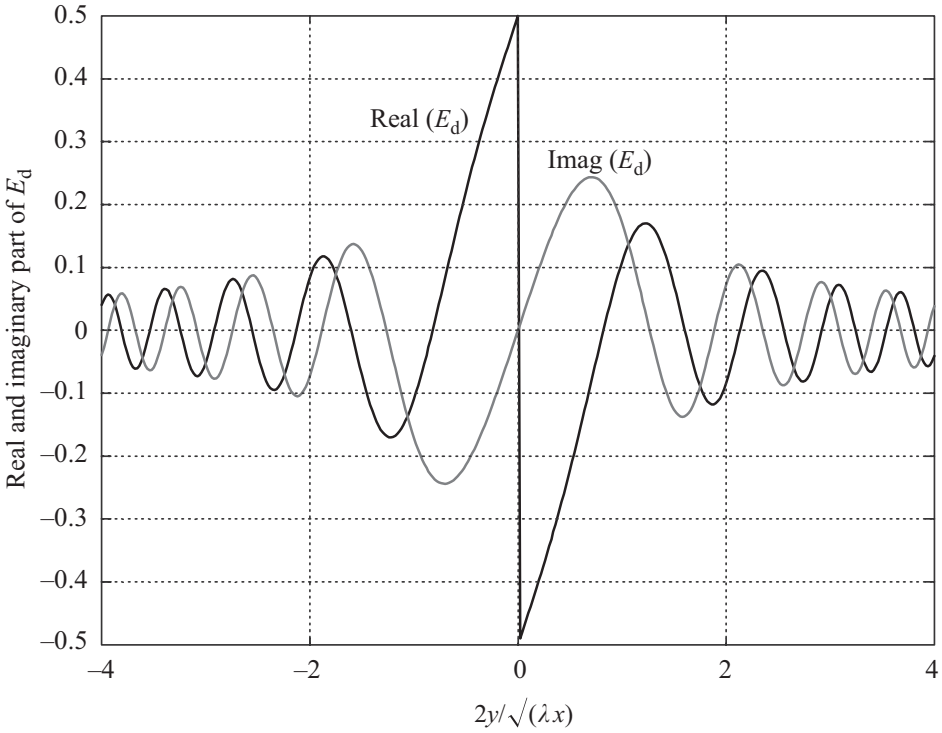


Figure 3.3.3 The real and imaginary part of diffracted field showing the discontinuity over the transition zone and the wave-like character further away

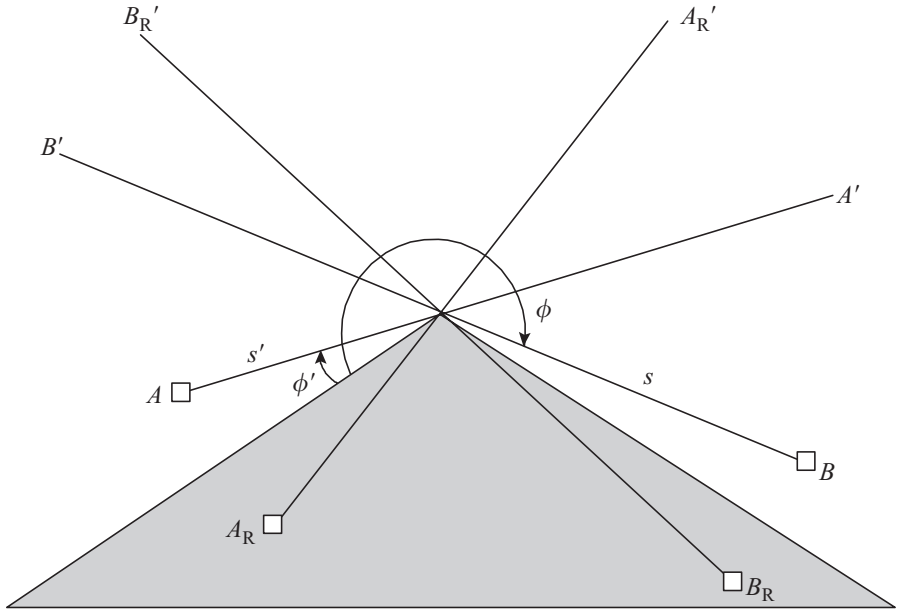


Figure 3.3.4 Geometry of wedge diffraction. The transmitter is at A and the receiver is at B with images A_R and B_R. The lines are shadow or reflection boundaries

not matter, since the solution will satisfy the reciprocity theorem. For ease of discussion, however, consider A as the transmitting antenna and located at a distance of s' from the wedge tip and an angle ϕ' with the left face of the wedge, similarly s and ϕ for B. The exterior angle of the wedge is $n\pi$, where n can be any number between zero and two. The line going through A and the edge, $(A–A')$, is obviously the shadow boundary between the lit and shadow region. The fields from the source are reflected from the boundary with reflection coefficient R_1 and this image source A_R has its own shadow boundary $(A_R–A'_R)$, the reflection shadow boundary. This is easy to understand if we think of the situation in GO terms. When the point B is to the left of the reflection shadow boundary it will see the image point A_R, to the right it will not see A_R, thus $(A_R–A_{R\prime})$ must be a transition boundary for the reflected field. From reciprocity we must have a similar situation with respect to B, so a total of four transition boundaries exist in the most general case. Four transition functions can thus be expected in the solution, as shown below.

For the case of a 3D source (spherical wave incidence) the total field is given by

$$E_{\text{total}} = E_{\text{GO}} + E_{\text{d}} \tag{3.3.5}$$

where E_{GO} is the geometrical optical field comprising the incident and reflected field to the left of the reflection shadow boundary, incident field alone to the left of the lit shadow boundary, and zero after both boundaries. The diffracted field is given by

$$E_{\text{d}} = E_{\text{i}}(0)\ D(\phi, \phi')\ A_{\text{d}}(s)\ \mathrm{e}^{-\mathrm{j}k\,s}, \tag{3.3.6}$$

where

$$A_{\text{d}}(s) = \sqrt{\frac{s'}{s(s+s')}} \tag{3.3.7}$$

$$E_{\text{i}}(0) = E_0 \frac{e^{-\mathrm{j}\,k\,s'}}{s'} \tag{3.3.8}$$

is the so-called spreading factor ensuring the proper decay rates, and $E_{\text{i}}(0)$ is the incident field on the edge. Writing E_{i} as a spherical wave shows that the reciprocity is fulfilled, with symmetry in s and s' (the product of A_{d} and $E_{\text{i}}(0)$ and $e^{-\mathrm{j}\,k\,s}$ is invariant to a change in s and s'). The diffraction coefficient, which should be symmetric in ϕ and ϕ', is given by

$$D\,(\phi, \phi') = \frac{-e^{-\mathrm{j}\pi/4}}{2n\sqrt{2\pi k}} \cdot (D_1 + D_2 + R_1\ D_3 + R_2\ D_4), \tag{3.3.9}$$

where

$$D_1 = \cot\left(\frac{\pi + (\phi - \phi')}{2n}\right) F(kLa^+(\phi - \phi')), \tag{3.3.10}$$

$$D_2 = \cot\left(\frac{\pi - (\phi - \phi')}{2n}\right) F(kLa^-(\phi - \phi')), \tag{3.3.11}$$

$$D_3 = \cot\left(\frac{\pi - (\phi + \phi')}{2n}\right) F(kLa^-(\phi + \phi')), \tag{3.3.12}$$

$$D_4 = \cot\left(\frac{\pi + (\phi + \phi')}{2n}\right) F(kLa^+(\phi + \phi')), \tag{3.3.13}$$

and R_2 is the reflection coefficient of the shadow side of the wedge.

The transition function embedding a Fresnel-like integral is given by

$$F(X) = 2\mathrm{j}\sqrt{X}\,\mathrm{e}^{\mathrm{j}X} \int_{\sqrt{X}}^{\infty} \mathrm{e}^{-\mathrm{j}t^2}\,\mathrm{d}t. \tag{3.3.14}$$

X is zero at the boundaries and increases away from the boundaries. L is a distance factor

$$L = \frac{ss'}{s + s'}, \tag{3.3.15}$$

and the functions $a^{\pm}(\beta)$ are given by

$$\begin{aligned} a^{\pm}(\beta) &= 2\cos^2\left(\frac{2n\pi\ N^{\pm} - \beta}{2}\right), \\ \beta &= \phi \pm \phi', \end{aligned} \tag{3.3.16}$$

where the sign is chosen appropriately, and the integers $N^{\pm}$ are those integers that most nearly satisfy

$$\begin{aligned} 2\pi n N^+ - \beta &= \pi \\ 2\pi n N^- - \beta &= -\pi. \end{aligned} \tag{3.3.17}$$

In practice, the equations are easy to use numerically, but it is not so straightforward to see through the physical interpretation. The equations for D_1 to D_4 represent the transition functions across the four shadow boundaries as discussed above.

As an example, having a closer look at D_2, the transition functions for small and large values of the argument are given by

$$\begin{aligned} F(X) &\approx 1 + \frac{\mathrm{j}}{2X}, \quad X \gg 1, \\ F(X) &\approx (\sqrt{\pi X} - 2X\mathrm{e}^{\mathrm{j}\pi/4})\,\mathrm{e}^{\mathrm{j}((\pi/4)+X)}, \quad X \ll 1. \end{aligned} \tag{3.3.18}$$

The cotangent in D_2 is singular for $\phi - \phi' = \pi$. Setting $\phi - \phi' = \pi - \epsilon$ and finding $N^- = 0$,

$$D_2 = \frac{-e^{-\mathrm{j}\pi/4}}{2n\sqrt{2\pi k}} \cot\left(\frac{\epsilon}{2n}\right) \; F\left(2\,kL\,\sin^2\left(\frac{\epsilon}{2}\right)\right), \tag{3.3.19}$$

which shows that the diffracted field is bounded and discontinuous on the boundary $\epsilon = 0$ as it should be (the singularity of the cot-function is cancelled by the F-function, and cot is antisymmetric where F is symmetric). Note also that the frequency cancels out near the boundary and only geometric parameters remain. The situation is similar to the PO field shown in Figure 3.3.2. Using the expansion for small X, the diffracted field near the boundary, as expected, is

$$\begin{aligned} E_{D_2} &= \frac{\mathrm{e}^{-\mathrm{j}\,k\,(s+s')}}{s+s'} \cdot \left(-\frac{1}{2}\right) \quad \epsilon > 0 \\ E_{D_2} &= \frac{\mathrm{e}^{-\mathrm{j}\,k\,(s+s')}}{s+s'} \cdot \left(+\frac{1}{2}\right) \quad \epsilon < 0 \end{aligned} \tag{3.3.20}$$

Further away from the boundary where the argument of F is larger than one, the diffracted field is

$$E_{D_2} = -E_0 \frac{e^{-\mathrm{j}\,k\,(s+s')}}{\sqrt{k\,(s+s')}} \sqrt{\frac{1}{ss'}} \frac{\mathrm{e}^{-\mathrm{j}\pi/4}}{2n\sqrt{2\pi}} \cot\left(\frac{\epsilon}{2n}\right). \tag{3.3.21}$$

This is the GTD ray for this particular boundary, and it shows the peculiar behaviour of the diffracted fields. Near the edge for $s \ll s'$ (and that may be many wavelengths), the wave behaves like a cylindrical wave with an amplitude decay of $1/\sqrt{\mathrm{s}}$; while for larger distances $s \gg s'$, it decays by $1/s$ as a spherical wave. The other terms (D_1, D_3, D_4) have similar behaviour. Near the shadow boundary it must of course behave as a spherical wave, otherwise it would not be able to cancel the discontinuity of the incident wave for all distances.

An example will illustrate the phenomenon. Consider the case of a sloping hill 200 m from the base station of height 25.5 m, afterwards sloping down again with the same slope (Figure 3.3.5(a)). The base antenna height is 15 m, and the mobile is 2 m above the ground.

Figure 3.3.5 depicts the case of absorbing ground (R = 0) (Figure 3.3.5(b)), perfectly conducting ground (R = +1) for vertical polarisation (Figure 3.3.5(c)), and

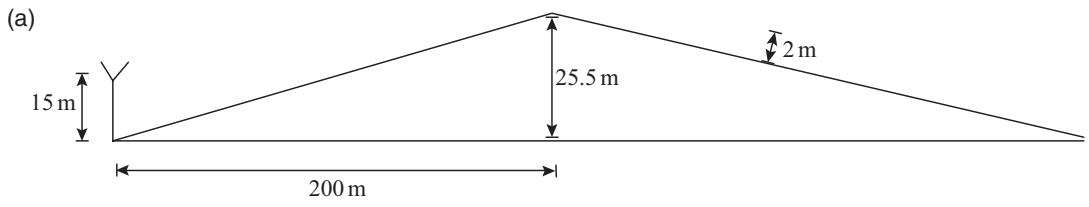


Figure 3.3.5(a) A base station and a mobile on a hill modelled by a wedge

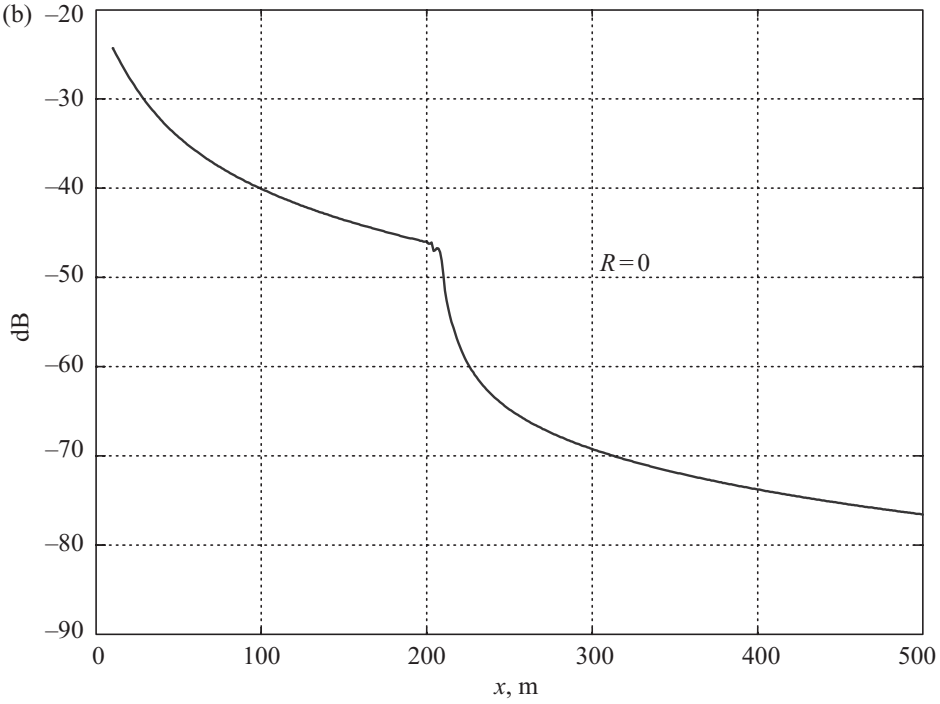


Figure 3.3.5(b) Total electric field over wedge-like hill of height 25.5 m at 900 MHz. Base antenna height 15 m and mobile antenna height 2 m. Perfectly absorbing

perfectly conducting ground for horizontal polarisation (R $= -1$) (Figure 3.3.5(d)). Case (d) is also approximately valid for vertical polarisation for dielectric ground and grazing incidence as was discussed for the smooth and flat ground.

In Figure 3.3.5(b) there is just the incident, free-space field (R $= 0$) before the wedge tip, and after that the diffraction field. We cannot talk about a specific constant diffraction loss, since the loss relative to that of free-space depends on distance, and the diffracted wave decays differently from the free-space wave as discussed above, more like from an effective source sitting at the tip. The case of 1800 MHz will lead to 3 dB greater path loss outside the transition zone.

In Figure 3.3.5(c) the main difference occurs before the wedge tip where the two ray interference dominates. After the tip the decay rate is as in (b) since the two components effectively add. Case (d) is markedly different since now the contribution from the forward scatter and reflection boundary subtract, and the final decay rate tends towards x^{-2}, in contrast to x^{-1} for cases (b) and (c). Note the recovery effect right after the tip, since the reflection boundary is passed first, and the cancellation of the direct and reflected wave is reduced. Note also that in all cases the field has a rapid immediate decay after passing the tip, which may be contributed to the usual 6 dB ($E = 0.5^* E_0$, equation (3.3.20)) loss at the boundary.

While the above discussion concerned a sharp wedge, the real-world situations have rounded surfaces. For very large radii of curvature (say rolling hills), the curves

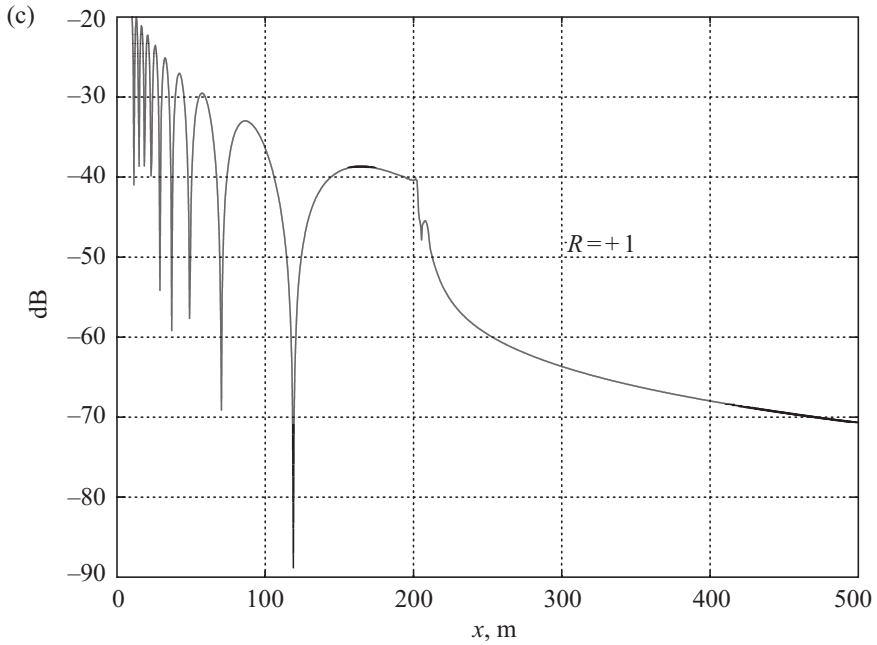


Figure 3.3.5(c) As in (b), but for reflection coefficient of +1

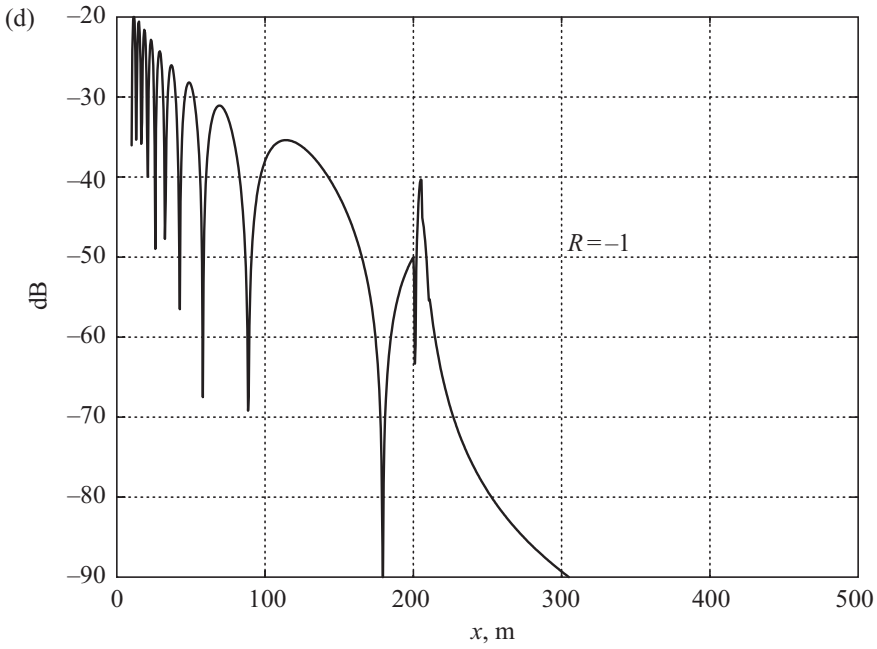


Figure 3.3.5(d) As in (b), but for reflection coefficient of −1

may be approximated by a series of wedges. For a single wedge with a small radius of curvature for the edge, there are expressions available for the diffraction coefficients. However, the principles here still hold, with the main change being a 'blurring' of the transition boundaries, which can be seen from geometric considerations.

In summary, the diffracted fields comprise four terms. These are attributed to the source and receiver positions and their geometric images in the faces of wedges. These contributions can be understood by recalling that the fields are given by a Kirchhoff integration over the plane aperture above the wedge. The aperture has contributions from the source and its image, and these contributions are 'seen' by both the receiver and its image.

3.3.2 Transition zone boundary

Further insight may be obtained by simplifying to the simplest case of the absorbing half-plane (Figure 3.3.6). For this case equation (3.3.9) simplifies, since $n = 2$, $R_1 = R_2 = 0$, and recalling the angle $\epsilon = \pi - (\phi - \phi')$, it is found that

$$D_{\mathrm{a}} = -\frac{\mathrm{e}^{-\mathrm{j}\pi/4}}{2\sqrt{2\pi k}} \frac{F(2kL\sin^2(\epsilon/2))}{\sin(\epsilon/2)} \tag{3.3.22}$$

since the terms D_1 and D_2 in equation (3.3.9) collapse into one. This is a good case for studying the extent of the transition zone. The transition zone boundary is here defined as the curve, where the argument in the transition function equals one, i.e.

$$2\,kL\ \sin^2(\epsilon/2) = 1. \tag{3.3.23}$$

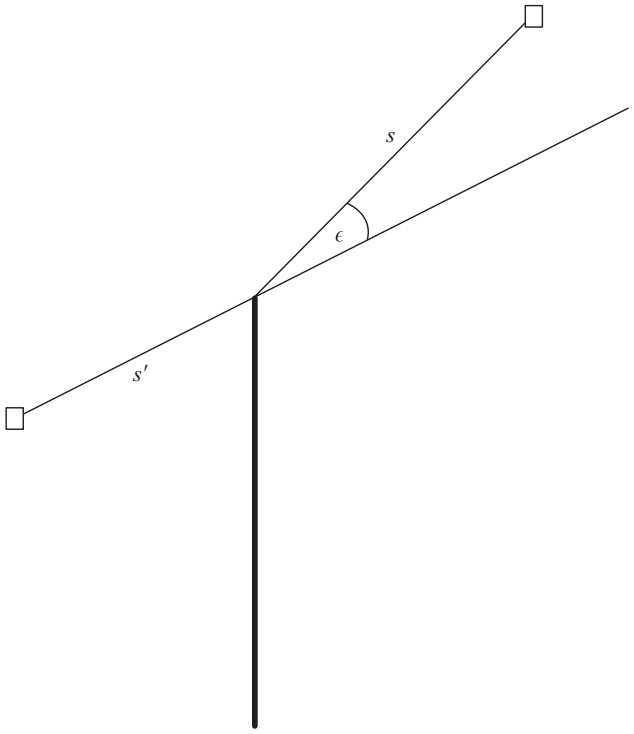


Figure 3.3.6 Absorbing half-plane

The curve may be found to be a hyperbola in the general case with s' finite. The asymptotes are given by

$$\begin{aligned} y &= \pm a(x-b), \\ a &= \sqrt{\frac{2ks'-1}{(ks'-1)^2}} \approx \sqrt{\frac{2}{ks'}}, \\ b &= -\frac{k\,s'^2}{2ks'-1} \approx -\frac{s'}{2}. \end{aligned} \tag{3.3.24}$$

For $s' \gg s$, the curve simplifies to the parabola

$$ky = \sqrt{2kx-1} \approx \sqrt{2kx} \tag{3.3.25}$$

$$y \approx \sqrt{\frac{x}{k}} = \sqrt{\frac{x\,\lambda}{2\pi}} \tag{3.3.26}$$

with focus on the edge. For small wavelength, the transition zone is thus a narrow region around the shadow boundary. The parabola may also be interpreted as a Fresnel zone with the other focus at infinity and with excess distance of $\lambda/2\pi$ (the difference between the direct ray from the edge and one going from an obstacle on the Fresnel zone).

Usual UTD for several obstacles is restricted to the cases where the following edge is outside the transition zones of the previous edges. It may be seen from Figure 3.3.7 that this is a great restriction, since often the height differences for a distance of 500 m would not be less than 10 m (or 5 m at 200 m distance), so there is a need for a more complete theory. As an example of the failure of usual UTD the case of N screens of equal height and equal distance may be taken where the antennas are at the same height and distance. Following the ideas in the previous section there would be a loss of 0.5 at each screen, i.e. exponential decay, $D_c = 0.5^N$. The exact solution gives a combined diffraction coefficient of $1/(N+1)$ (Lee, 1978). By a combined diffraction coefficient is meant the ratio of the total field to the free space field. One approximate solution to this dilemma is to include slope diffraction, as described in the following section.

3.3.3 UTD transition zone diffraction

In this section a strict ray-tracing tool is developed, which allows an approximate, yet accurate, and fast determination of fields diffracted around a multiplicity of obstacles. The main problem with straightforward application of the UTD is that in many cases one edge is in the transition zones from the previous edges. Strictly speaking this forbids application of ray techniques, but in the spirit of UTD the principle of local illumination of an edge should still be valid, so at least to some approximate degree, a solution can still be obtained. As will be shown, the solution is quite accurate in most cases of practical interest. The first key point in the theory is to include slope diffraction (Andersen, 1997), which is usually neglected as a higher-order term in an asymptotic expansion in frequency, but in transition zone diffraction the term has the

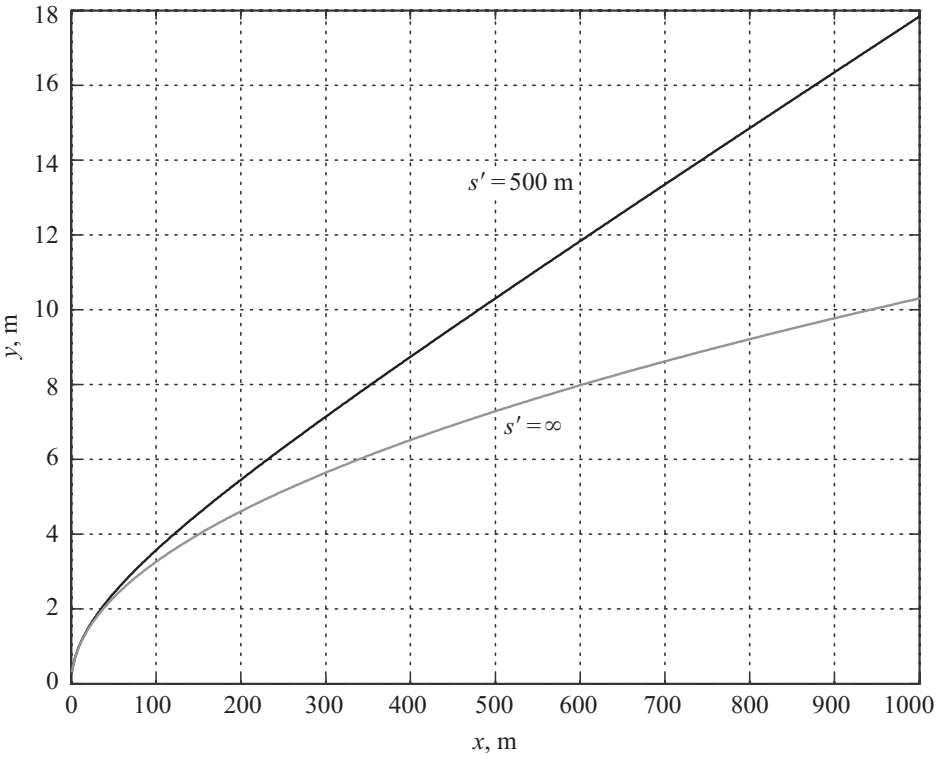


Figure 3.3.7 Transition zone lateral extent for $s' = 500\,m$ (a hyperbola) and for $s' = \infty$ (a parabola). Frequency equals 900 MHz

same order as the ordinary amplitude diffraction terms. The second key element in the method is automatic enforcement of continuity of amplitude and slope at each point. The extension to higher-order derivatives has not been considered. In the following, only absorbing screens will be treated, but extension to the general case of reflecting wedges is not difficult, since all the formulas are available.

The basic UTD theory gives the following equation for the total diffracted field for a simple absorbing half-plane

$$\begin{aligned} E_{\mathrm{d}} &= \left(E_{\mathrm{i}}(0)\ D + \frac{\partial E_{\mathrm{i}}(0)}{\partial n}\ d_{\mathrm{s}} \right)\ A(s)\ \mathrm{e}^{-\mathrm{j}ks} \\ &= (E_{\mathrm{da}} + E_{\mathrm{ds}})\ A(s)\ \mathrm{e}^{-\mathrm{j}ks} \end{aligned} \tag{3.3.27}$$

(cf. equation (3.3.6), but with a slope term added) where the derivative is in the direction normal to the wave direction. The diffraction term is recalled as

$$D(\alpha) = -\frac{\mathrm{e}^{-\mathrm{j}\pi/4}}{2\sqrt{2\pi k}\sin(\alpha/2)}\ F(2kL\sin^2(\alpha/2)). \tag{3.3.28}$$

L is a distance factor to be determined later, α is the angle above the shadow boundary and $A(s)$ is the so-called spreading factor

$$A(s) = \sqrt{\frac{s_0}{s\ (s + s_0)}}, \tag{3.3.29}$$

as described earlier. The notation has been changed slightly in this section, with s_0 corresponding to s' and α corresponding to ϵ in the previous section; d_s is the slope diffraction coefficient and is intimately connected with the diffraction coefficient D,

$$d_s = \frac{1}{jk}\frac{\partial D}{\partial \alpha}. \tag{3.3.30}$$

The transition function is, as above, given by

$$F(X) = 2j\sqrt{X}\, e^{jX} \int_{\sqrt{X}}^{\infty} e^{-ju^2} du. \tag{3.3.31}$$

Since we need derivatives of F in the following, we state without derivation

$$F'(X) = j\ (F(X) - 1) + \frac{F(X)}{2X}. \tag{3.3.32}$$

The slope diffraction coefficient d_s may now be determined from equation (3.3.30). Straightforward application of the formulas above gives

$$d_s = \frac{e^{-j\pi/4}}{\sqrt{2\pi k}}\, L \cos(\alpha/2)(1 - F(X)). \tag{3.3.33}$$

Outside the transition region where X is large the following approximations are valid:

$$F(X) \approx 1 + \frac{j}{2X}, \quad X \gg 1, \tag{3.3.34}$$

$$F'(X) \approx \frac{j}{4X^2}, \quad X \gg 1, \tag{3.3.35}$$

and in the transition zone, where X is small

$$F(X) \approx \sqrt{\pi X}\, e^{j\pi/4}, \quad X \ll 1, \tag{3.3.36}$$

$$F'(X) \approx \frac{e^{j\pi/4}}{2}\sqrt{\frac{\pi}{X}}, \quad X \ll 1. \tag{3.3.37}$$

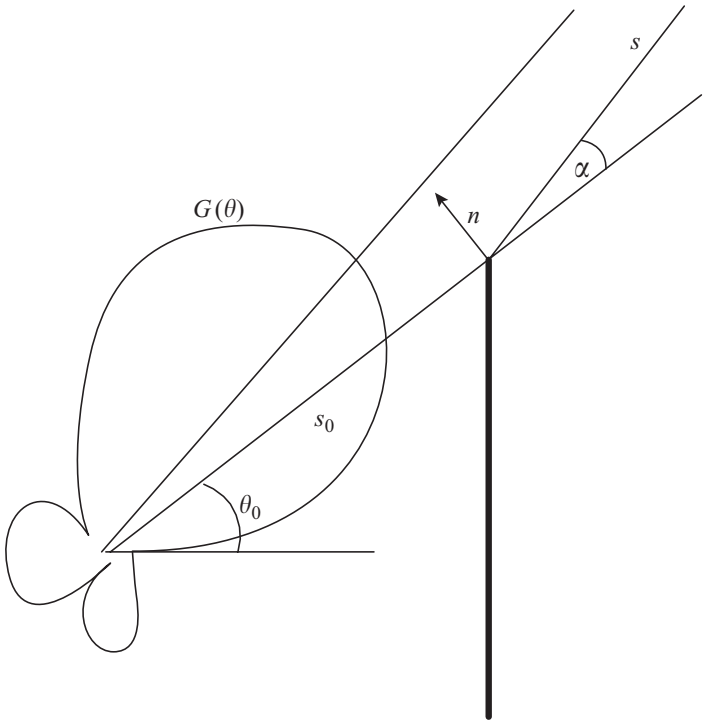


Figure 3.3.8 An antenna with radiation pattern $G(\theta)$ illuminating a screen. The pattern has a non-zero derivative in the direction of the edge

3.3.3.1 An antenna and one screen

Consider the situation with an antenna at a distance s_0 from the edge of a vertical screen and the angle θ_0 between the horizon and direction of incidence (Figure 3.3.8). The antenna is assumed to have the radiation pattern $G(\theta)$. The incident field and its derivative at the edge are given by

$$E_{\mathrm{i}}(0) = \frac{\mathrm{e}^{-\mathrm{j}\,k\,s_0}}{s_0}\, G(\theta_0), \tag{3.3.38}$$

$$\frac{\partial E_{\mathrm{i}}(0)}{\partial n} = \frac{\mathrm{e}^{-\mathrm{j}\,k\,s_0}}{s_0^2}\, G'(\theta_0). \tag{3.3.39}$$

The governing principle for the determination of the L values is the continuity of field and its derivative (slope) along the transition boundary after the edge, along $\alpha = 0$. Considering first the field magnitude, the diffraction coefficient near the boundary may be found as

$$D = -0.5\,\sqrt{L}\,\mathrm{sign}(\alpha). \tag{3.3.40}$$

In order to have continuity of the total field, the discontinuity of the diffracted field should cancel the discontinuity of the incident field, i.e. the diffracted field E_{d} should

be one half of the incident field for all values of s, the distance from the edge,

$$E_{\mathrm{da}}\ A(s)\mathrm{e}^{-\mathrm{j}\,k\,s} = -\frac{\mathrm{e}^{-\mathrm{j}\,k\,s_0}}{s_0}\sqrt{\frac{\sqrt{L}}{2}}\ \sqrt{\frac{s_0}{s\ (s+s_0)}}\ \mathrm{e}^{-\mathrm{j}\,k\,s} = -\frac{1}{2}\frac{\mathrm{e}^{-\mathrm{j}k\,(s+s_0)}}{s+s_0} \tag{3.3.41}$$

from which the value of L may be found as

$$L = \frac{s_0 s}{s_0 + s}, \tag{3.3.42}$$

which is the same as in equation (3.3.15).

A similar argument may be applied to the slope, which should also be continuous after the edge, i.e the slope of the slope diffracted wave should be minus one half of the slope of the incident field along the shadow boundary.

Since it is not certain *a priori* that the L for the slope wave (short notation for the slope diffracted wave) is the same as for the amplitude diffracted wave, it is denoted L_{s}. The slope of the slope wave is

$$\begin{aligned}\frac{\partial d_{\mathrm{s}}}{\partial \alpha} &= \frac{\mathrm{e}^{-\mathrm{j}\pi/4}}{\sqrt{2\pi k}}\ L_s\ (-F'(X))\frac{\mathrm{d}X}{\mathrm{d}\alpha} \\ &= -\frac{1}{2}\ {L_s}^{3/2}\ \mathrm{sign}(\alpha).\end{aligned} \tag{3.3.43}$$

It is noted that the slope wave has discontinuous slope, as it should have to balance out the discontinuity of the slope of the incident plus amplitude diffracted wave when $G'(\theta)$ is non- zero. The amplitude diffracted wave (D) has a discontinuous amplitude but a continuous slope, so the total slope discontinuity comes from the incident field. Going through the same exercise as above to find L_{s}, i.e. matching the slopes, the following equations can be set up

$$\frac{1}{2}\ \frac{\partial E_{\mathrm{i}}(s)}{\partial n} = -\frac{\partial E_{\mathrm{i}}(0)}{\partial n}\ \frac{\partial d_{\mathrm{s}}}{\partial n}\ A(s) \tag{3.3.44}$$

i.e.

$$\frac{1}{2}\ G'(\theta_0)\ \frac{\mathrm{e}^{-\mathrm{j}\,k\,(s+s_0)}}{(s+s_0)^2} = -G'(\theta_0)\ \frac{e^{-\mathrm{j}\,k\,s_0}}{s_0^2}\ \frac{(-1)}{2\,s}\ L_{\mathrm{s}}^{3/2}\ \sqrt{\frac{s_0}{s\ (s+s_0)}}\ e^{-\mathrm{j}\,k\,s} \tag{3.3.45}$$

from which it is found that $L_{\mathrm{s}} = L$ in this particular case.

It is now possible to find the total solution by using equation (3.3.27). Along the shadow line the ratio between the two components (the slope diffracted, and the amplitude diffracted) may be found to be

$$\frac{E_{\mathrm{ds}}}{E_{\mathrm{da}}} = \frac{2\ e^{-\mathrm{j}\frac{\pi}{4}}}{\sqrt{2\pi}}\ \frac{1}{\sqrt{k s_0}}\ \sqrt{\frac{s}{s+s_0}}\ \frac{G'(\theta_0)}{G(\theta_0)}. \tag{3.3.46}$$

When the antenna is many wavelengths from the edge of the single screen, i.e. $k\ s_0 \gg 1$, the slope wave has only marginal significance.

3.3.3.2 Diffraction over two screens

The same method will now be applied to two screens and later by recursion to several screens. For simplicity the antennas and the edges will all have the same height, but this is not a fundamental assumption. It is also assumed that the slope of the incident field is zero (see Figure 3.3.9). The incident wave on the second edge consists now of two components, an amplitude wave and a slope wave. The amplitude wave is the combined incident and amplitude diffracted wave for a position just above the shadow boundary. (It is really not important if a position above or below the boundary is chosen, since continuity is now established.)
With the assumptions above, the total incident field on the second edge is

$$E_2 = E_i + E_1\ D_1\ A_1(s_1) \tag{3.3.47}$$

where the suffix 1 refers to the first edge. E_1 is the incident wave from the source evaluated at edge 1. D_1 is given by equation (3.3.40), and L by equation (3.3.42), so E_2 is

$$\begin{aligned} E_2 &= \frac{1}{s_0}\, 0.5 \sqrt{\frac{s_0 s_1}{s_0 + s_1}} \sqrt{\frac{s_0}{s_1(s_0 + s_1)}} \\ &= \frac{0.5}{s_0 + s_1}, \end{aligned} \tag{3.3.48}$$

a well-known result: the first edge halves the incident field. In equations (3.3.46) and (3.3.47) the phase terms have been suppressed, since they are trivial; the phase simply progresses uniformly with distance.

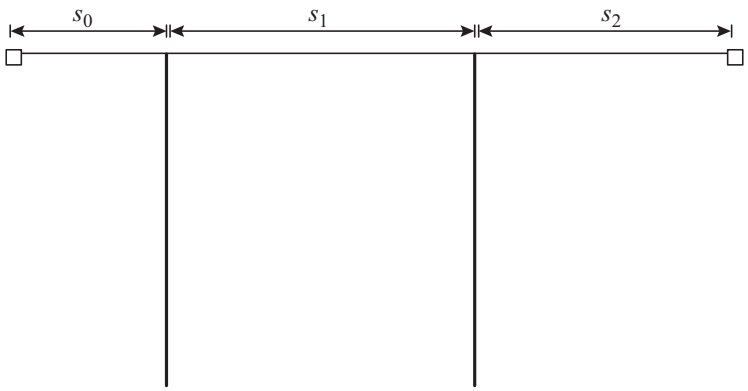


Figure 3.3.9 Diffraction over two screens of zero height

In order to find the slope wave it is necessary to find the slope of the incident field as a function of distance from the first edge,

$$\begin{aligned}\frac{\partial E_2}{\partial n} &= E_1 \frac{\partial D_1}{s\partial\alpha} A_1(s) \\ &= E_1 \frac{1}{s} \mathrm{j}\, kd_{\mathrm{s}}(L_1)\, A_1(s) \\ &= \frac{\mathrm{e}^{\mathrm{j}\pi/4}}{\sqrt{2\pi}} \sqrt{k} \sqrt{\frac{s_0}{s}} \frac{1}{(s_0+s)^{3/2}}\end{aligned} \tag{3.3.49}$$

The value at the edge is found by setting $s = s_1$. As before, the L_{s} value is found by enforcing the L_{s} in the slope wave after the second edge to have such a value that it exactly balances the slope of the E_2 wave, i.e.

$$\frac{1}{2}\frac{\partial E_2(s_1+s_2)}{\partial n} = -\frac{\partial E_2(s_1)}{\partial n}\frac{\partial d_{\mathrm{s}}}{\partial n} A_2(s_2), \tag{3.3.50}$$

i.e.

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{e}^{\mathrm{j}\pi/4}}{\sqrt{2\pi}}\sqrt{\frac{ks_0}{(s_1+s_2)}}\frac{1}{(s_0+s_1+s_2)^{3/2}} \\ &\quad = -\frac{\mathrm{e}^{\mathrm{j}\pi/4}}{\sqrt{2\pi}}\sqrt{\frac{k\, s_0}{s_1}}\frac{1}{(s_0+s_1)^{3/2}}\frac{(-1)}{2\, s_2} L_{\mathrm{s}}^{3/2}\sqrt{\frac{s_0+s_1}{s_2(s_0+s_1+s_2)}},\end{aligned} \tag{3.3.51}$$

which leads to

$$L_{\mathrm{s}} = \left(\frac{s_0+s_1}{s_0+s_1+s_2}\right)^{2/3}\left(\frac{s_1}{s_1+s_2}\right)^{1/3} s_2. \tag{3.3.52}$$

It is now a simple matter to find the field after edge 2,

$$\begin{aligned}E_3 &= E_{\mathrm{i}} + E_2 D_2 A_2 + \frac{\partial E_2}{\partial n} d_{\mathrm{s}2} A_2 \\ &= \frac{0.25}{s_0+s_1+s_2} + \frac{\mathrm{e}^{\mathrm{j}\frac{\pi}{4}}}{\sqrt{2\pi}}\sqrt{k}\sqrt{\frac{s_0}{s_1}}\frac{\mathrm{e}^{-\mathrm{j}\frac{\pi}{4}}}{(s_0+s_1)^{3/2}}\frac{L_{\mathrm{s}}}{\sqrt{2\pi k}} A_2(s),\end{aligned} \tag{3.3.53}$$

where

$$A_2(s_2) = \sqrt{\frac{s_0+s_1}{s_2(s_0+s_1+s_2)}}. \tag{3.3.54}$$

Two things are worth noting. The frequency dependence disappears, since the slope of the incident field increases with frequency at the same rate that the slope coefficient decreases; and the slope wave adds in phase with the amplitude diffracted wave. By normalising the field with the free space field, the result can be expressed as a

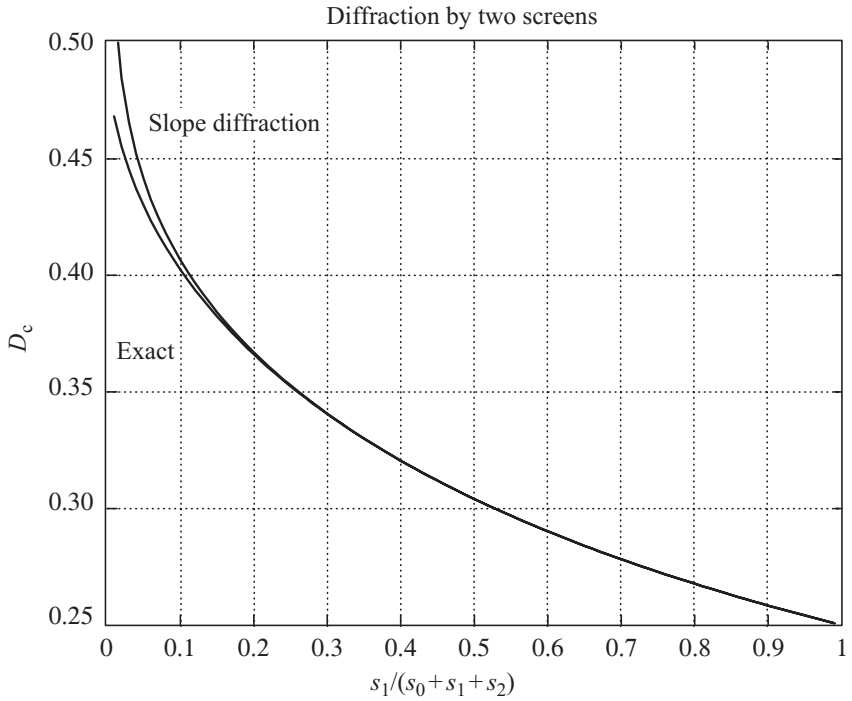


Figure 3.3.10 Comparison between exact and approximate solution where $s_0 = s_2$; s_1 is the distance between the two screens and when it approaches zero, the solution diverges

combined diffraction coefficient D_c, including the spreading and distance factors, as

$$D_c = 0.25 + \frac{1}{2\pi}\sqrt{\frac{s_0}{s_1}}\frac{(s_0+s_1+s_2)}{(s_0+s_1)^{3/2}}L_s A_2$$

$$= 0.25 + \frac{\sqrt{s_0 s_2}}{2\pi}(s_1(s_0+s_1+s_2))^{-1/6}((s_0+s_1)(s_1+s_2))^{-1/3}. \quad (3.3.55)$$

Despite its simplicity this expression is remarkably accurate. For all distances equal ($s_0 = s_1 = s_2$), the result is 0.3335 where the exact result is 1/3; see Figure 3.3.10. However, it is an approximation, and it is noted that the result diverges for s_1 tending to zero, where the exact result is 0.5 (two screens collapsing into one). Note also that reciprocity is satisfied, since interchange of s_0 and s_2 leaves the expression unchanged.

3.3.3.3 Diffraction over *N* screens by recursion

It is now possible to generalise the result above to the general case of N screens by following the following principles. The diffracted field after each screen has an amplitude and slope component, each determined by its own length parameter L and L_s, which are found by enforcing amplitude and slope continuity over the shadow boundary. This process becomes unwieldy analytically but very easy in a numerical,

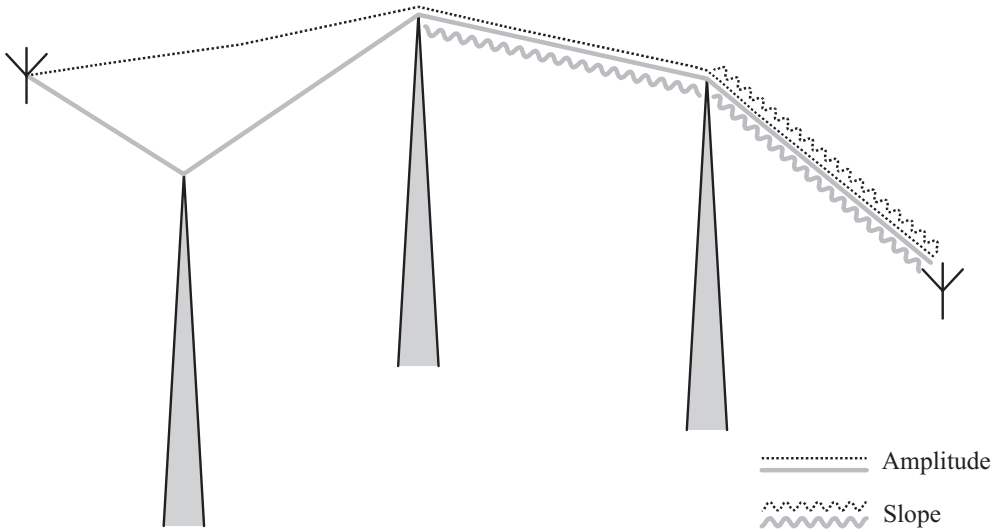


Figure 3.3.11 Diffraction over three screens. Wavy lines are slope waves and straight lines amplitude waves. The direct wave from source to edge 2 generates a slope wave after edge 3. The wave diffracted from edge 1 generates a slope wave after edge 2. There is a total of four waves after edge 3. The slope of the antenna pattern is neglected

recursive algorithm. It may best be understood by following an example for three screens (Figure 3.3.11),

$$E_4(s) = E_{\text{inc},4} + (E_3(0)\ D_3(s) + \frac{\partial E_3(0)}{\partial n} d_3(s)) A_3(s). \tag{3.3.56}$$

The value of the field at the receiving antenna may be found through the basic UTD formula where the two L and L_s should be determined from the value of the amplitude and slope of the incident field ($E_{\text{inc},4}$) through continuity of amplitude and slope along the shadow boundary. The incident field is the field after diffraction by two screens, screen three not present, i.e. this requires two new L-values, determined by the incident field at the second screen, which again is determined by the free space field from the original source. Thus it is seen that the field at a given point after N screens needs information from all previous screens for that point. In this sense, transition zone diffraction has 'memory', in contrast to the usual GTD multiplication of independent factors. The slope diffraction from the antenna pattern is neglected, which means that there has to be one screen diffraction before slope diffraction occurs.

An algorithm according to the principles mentioned has been produced and tested for the canonical case of N screens with equal heights and distances, where the result is known (Lee, 1978; Vogler, 1982). The comparison is shown in Figure 3.3.12.

It is noted that the slope diffraction gives an error of about 1 dB after ten screens. This is inherent in the method, since even after two screens there is an error, and the errors accumulate.

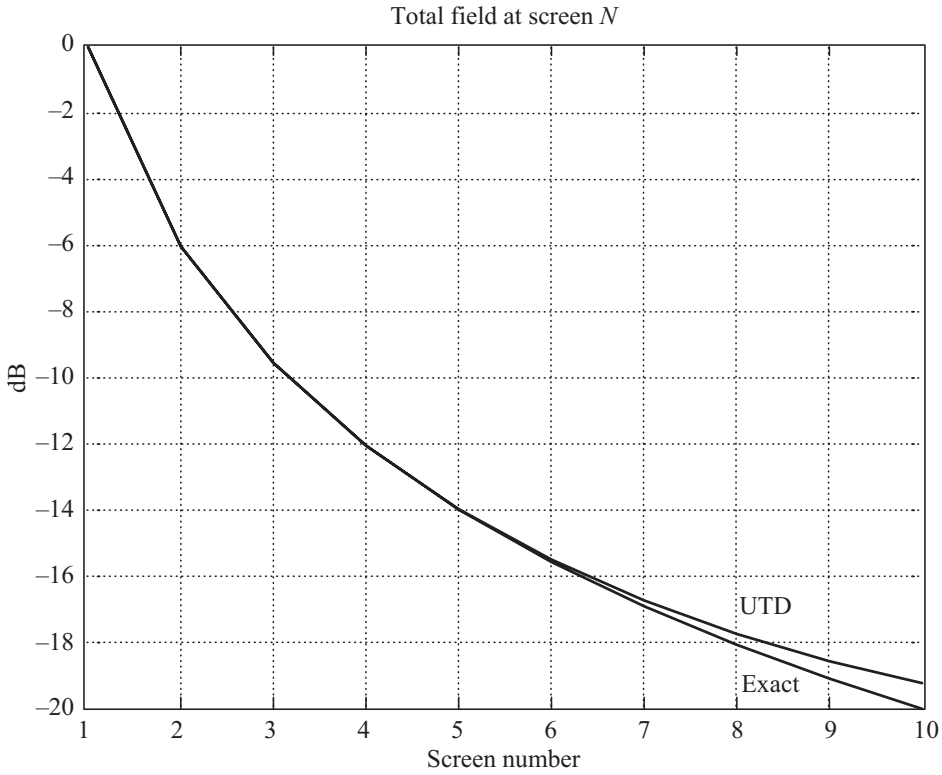


Figure 3.3.12 Field magnitude relative to free space at screen N by UTD slope diffraction

The case of diffraction over multiple screens of arbitrary heights and spacings has been treated within the frame of the UTD. By including slope diffraction and enforcing continuity of amplitude and slope of the total field after each diffraction a solution is obtained, which agrees to a good approximation with the known results for constant spacing. The limitation of the method is that it is not applicable when one spacing becomes very small relative to other spacings. Thus the method cannot predict the collapse of two screens into one.

3.4 Ray tracing and radiosity

The desire for computer modelling of microcells and indoor environments has led to the development of various techniques based on ray tracing. This seems obvious considering the small wavelengths usually considered for mobile communications, and conceptually it is straightforward even in 3D environments including polarisation and diffractions, but in practice the amount of data increases enormously, and some limitations must be considered. We shall briefly discuss three techniques, two for geometrical optics, ray tracing and ray launching, for the specular fields, and one, radiosity, for the diffuse fields.

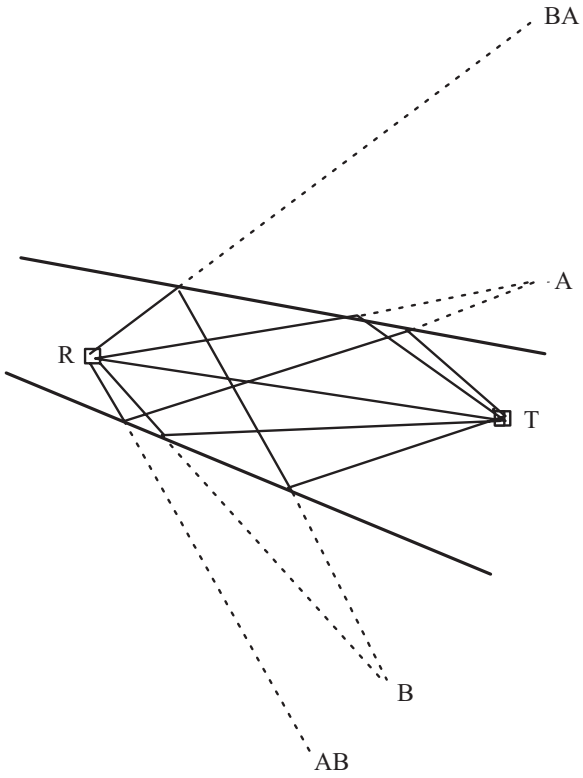


Figure 3.4.1 Ray tracing between two walls using image theory

3.4.1 Ray tracing

Ray tracing as such is an old discipline based on geometrical optics, which means straight ray paths and power decay versus distance as inverse squares. For electromagnetic fields, polarisation may be incorporated. In the present context of modelling houses, street canyons and such, the obstacles have usually been modelled as planar structures with smooth surfaces. This means that the fields reflected from a surface can be replaced by the fields from an effective image source with a proper reflection coefficient as a weight factor. The earlier two-path model is an example of such a development. An example with two walls will illustrate the technique (Figure 3.4.1).

Ray tracing is essentially a point-to-point technique as shown in Figure 3.4.1. The transmitter T has a direct path to the receiver R. Then there is a one bounce reflection off wall A computed by drawing a line from R to point A, the image of T in plane A. In this way the point of reflection is determined, and the angles determine the reflection coefficient. The time delay relative to the direct path is automatically calculated via the path length difference. For a finite size wall there will be points where the image point is not visible, and in a GO model this would give zero field, but in a diffraction model a finite value is found. A rule of thumb would be, that if the Fresnel zone (see Appendix D) surrounding the line between R and A is well within the reflecting plane, then it is not necessary to include diffraction.

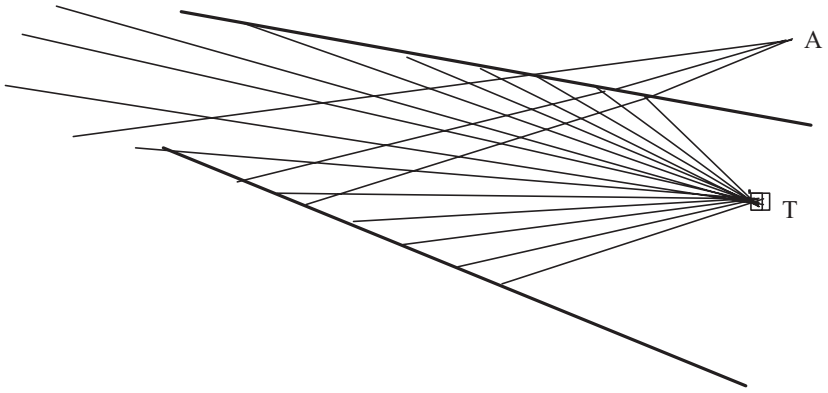


Figure 3.4.2 Ray launching. Rays are launched in a fan of discrete angles

The drawback of the method is that computation time grows with the number of receiving points, so it is not very well suited to coverage calculations. Besides, there is an exponential dependency on the number of interaction points. There are a number of computer programs available dealing with ray tracing, mainly for the indoor or microcell environment.

3.4.2 Ray launching

Ray launching is an alternative technique better suited to calculation of area coverage. The space around the transmit antenna is divided into equiangular regions in 2D or 3D space as needed. Field strength is summed at all potential receiver points and the ray search is stopped when energy falls below a minimum or a given maximum of interactions is reached (Gschwendtner *et al*., 1995). The technique is illustrated in Figure 3.4.2.

3.4.3 Radiosity or diffuse scattering

The main problem with the methods above is that they assume smooth surfaces giving rise to specular reflections. Occasionally, the coherent reflection coefficient is multiplied by the Rayleigh roughness factor, but the diffuse field is often neglected.

Radiosity (Ashdown, 1994) is a name given by the Computer Graphics community to incoherent scattering, and they have used it for creating artificially lit scenes. Some of the same programming tools may be practical for predicting radio wave propagation as well. Essentially, it is just a simple application of the results from Section 3.2.3 on surface roughness and diffuse fields. A further approximation is a simple form of scattering function, like Lambertian scattering.

$$\sigma = \rho \cos(\theta_1) \cos(\theta_2), \tag{3.4.1}$$

where θ_1 and θ_2 (not identical) are the incident and reflected angles with the normal to the surface, and ρ is a power reflection coefficient. The total scattered power is then found by integrating over the total surface, eventually leading to an integral equation

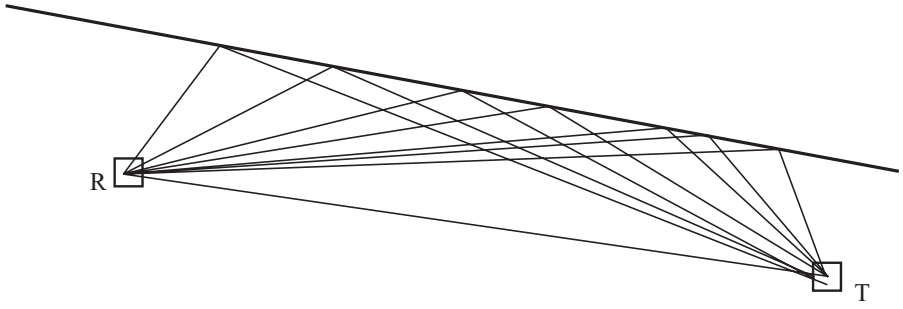


Figure 3.4.3 Illustration of diffuse incoherent scattering from a rough surface. Contributions arrive from all parts of the surface, not just from the specular reflection point

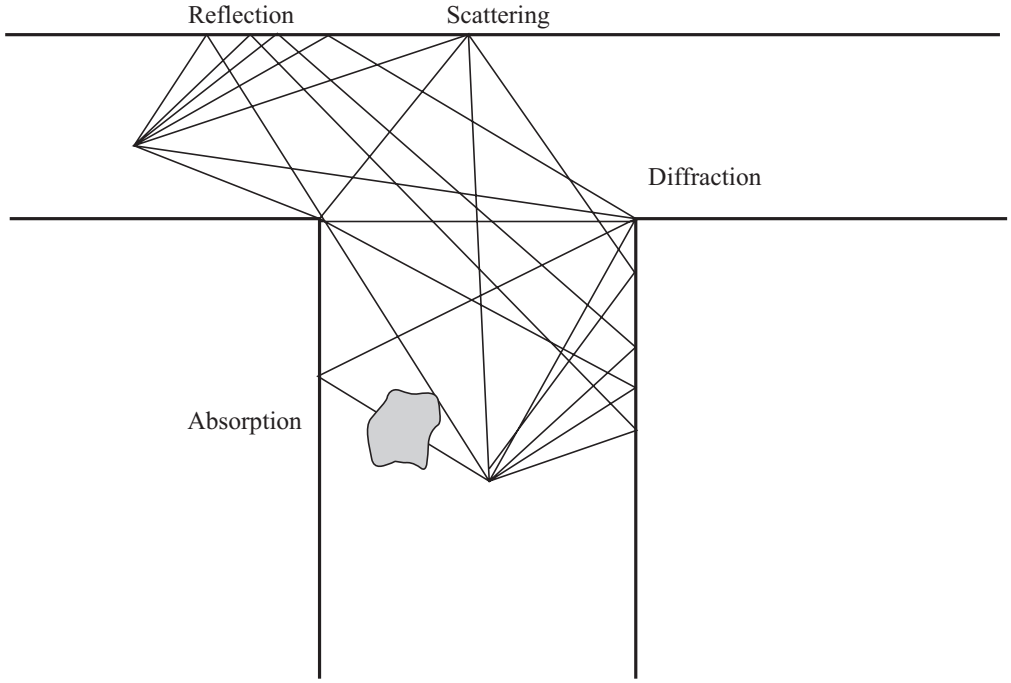


Figure 3.4.4 Ray mechanisms near a street corner

for the environment in question when all multiply scattered contributions are taken into account. The situation may be seen in Figure 3.4.3 for just one surface. Since the new rays originate at the surface and not from an image point, the fall-off with distance is more rapid than for the reflected rays in the previous sections.

3.4.4 An example of a street junction

The techniques or mechanisms of Chapter 3 may be illustrated by considering a simple example, where all the mechanisms come into play, a simple 2D street

junction (Figure 3.4.4). The two positions do not have line-of-sight, and some of the mechanisms seen (the reader is asked to identify for himself which is which) are

scattering
diffraction
scattering–diffraction
scattering–reflection
reflection–diffraction
reflection–reflection
diffraction–diffraction
and so on.

A tree has been added, which gives additional attenuation and scattering, changing over the year. It is no wonder that some people prefer to have more empirical formulas for turning a corner.

3.5 References

ANDERSEN, J. BACH: 'UTD multiple-edge transition zone diffraction', *IEEE Transactions*, July 1997, **AP-45** (7), pp. 1093–1097.

ASHDOWN, I.: *Radiosity: A Programmers Perspective*, Wiley, 1994.

BASS, F.G. and FUKS, I.M.: *Wave Scattering from Statistically Rough Surfaces*, Pergamon Press, 1979.

BECKMANN, P. and SPIZZICHINO, A.: *The Scattering of Electromagnetic Waves from Rough Surfaces*, Pergamon Press, 1963.

BERTONI, H.L. and WALFISCH, J.: 'A theoretical model of Uhf propagation in urban environments', *IEEE Transactions*, 1988, **AP-12**, pp. 1788–1796.

BLAUNSTEIN, N.: *Radio Propagation in Cellular Networks*, Artech House, 2000.

DESANTO, J.A. and BROWN, G.S.: *Progress in Optics*, **XXIII**, Ed. E. Wolf, North Holland, 1986.

GSCHWENDTNER, B.E., WÖLFLE, G., BURK, B., and LANDSTORFER, F.M.: 'Ray tracing vs ray launching in 3D microcell modelling', *Proc. EPMCC'95*, Bologna, 1995, pp. 74–79.

JORDAN, E.C. and BALMAIN, K.G.: *Electromagnetic Waves and Radiating Systems*, Prentice Hall, 1950 (also reprinted 1968).

KOUYOUMJIAN, R.G. and PATHAK, P.H.: 'A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface', *Proceedings of the IEEE*, 1974, **62** (11), pp. 1448–1461.

LEE, S.W.: 'Path integrals for solving some electromagnetic edge diffraction problems', *Journal of Mathematical Physics*, 1978, **19**, pp. 1414–1422.

McNAMARA, D.A. *et al.*: *Introduction to the Uniform Geometrical Theory of Diffraction*, Artech House, 1990.

NORTON, K.A.: 'The propagation of radio waves over the surface of the earth and in the upper atmosphere', *Proc. IRE*, 1937, **24**, p. 1192.

OGILVY, J.A.: *Theory of Wave Scattering from Random Rough Surfaces*, IOP Publishing, 1991.

SOMMERFELD, A.: 'The propagation of waves in wireless telegraphy', *Ann. Physik*, 1909, **28**, p. 665.

TZARAS, C. and SAUNDERS, S.R.: 'An improved heuristic UTD solution for muliple-edge transition zone diffraction', *IEEE Transactions Antennas and Propagation*, December 2001, **49** (2), pp. 1678–1682.

VALENZUELA, G.: Depolarisation of EM waves by slightly rough surfaces, *IEEE Transactions*, 1967, **AP-15**, pp. 552–557.

VOGLER, L.: 'An attenuation function for multiple knife-edge diffraction', *Radio Science*, 1982, **17**, pp. 1541–1546.

Chapter 4

Propagation modelling

A variety of methods have been applied to the problem of path loss in real environments, natural or man-made. A number of methods have been parametric in the sense that a few parameters describe the path loss based on experimental results, thus yielding fast results for network planning. Others have been more based on simple modelling, like a few knife-edges, and then treated by approximate methods. Recently, there has been some interest in a more accurate, deterministic modelling of the environment, especially for indoor and microcell situations. First, the outdoor above-rooftop case will be treated in the urban environment, followed by the rural case and the indoor environment. Additionally, the time-domain response as well as the angular response are treated since they are important for wideband systems and for the use of adaptive antennas.

4.1 Propagation in urban environments above rooftops

4.1.1 Okumura–Hata case for urban environments

The most extensively used propagation data for mobile communications are still Okumura's measurements from Japan in 1968 (Okumura *et al*., 1968). The frequencies ranged from 450 to 1900 MHz with variation of antenna heights and terrain type, and a large amount of data was given as graphs showing average path loss. Statistical distributions were also studied with indication of instantaneous Rayleigh fading, and log-normal fading for the small sector median values. Figure 4.1.1 is taken from Okumura *et al*. (1968), and it shows the distance dependency for the smoothed data.

The scale in Figure 4.1.1 is log–log and the median attenuation in Okumura's notation is the field strength relative to free space; the polarisation is vertical. Note that up to 10–15 km a straight line with slope $n = 0.5$ is a good model, and after 35 km another straight line with slope $n = 2.3$ is valid. Comparing with the power slopes normally used in the literature

$$P \approx \frac{1}{d^{\alpha}}, \tag{4.1.1}$$

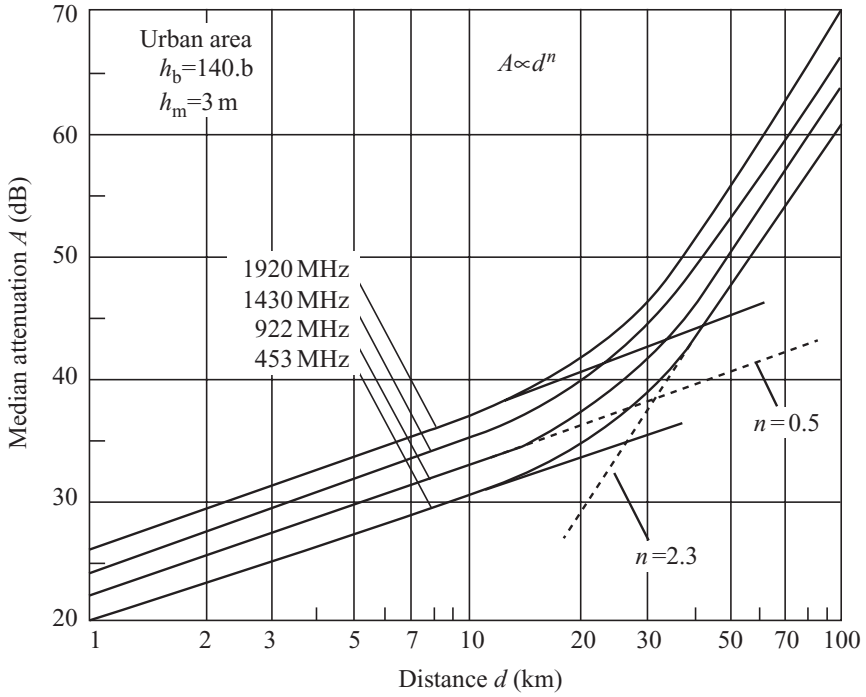


Figure 4.1.1 Example of distance dependence of median attenuation in urban area (figure 11 from Okumura et al., 1968); the base station height is 140 m and the mobile antenna height is 3m

where $\alpha = 2 + 2n$, so the two slopes correspond to $\alpha = 3$ and 6.6 respectively. The values are for base antenna height of 140 m. The power-law dependencies will be discussed later in a theoretical context. The Okumura curves are not so easy to use in a planning tool, so Hata (1980) derived some parametric formulas, which are in general use. It should be noted that Hata made some simplifications by neglecting the second slope, limiting the range to 20 km. The path loss in Hata's notation reads

$$L_p(\text{dB}) = 69.5 + 26.16 \log f_c - 13.82 \log h_b - a(h_m) + (44.9 - 6.55 \log h_b) \log R, \quad (4.1.2)$$

where carrier frequency f_c is in MHz, base station height h_b is in m and range R is in km. The correction factor for mobile antenna height h_m is

$$a(h_m) = (1.1 \log f_c - 0.7)h_m - (1.56 \log f_c - 0.8) \quad (4.1.3)$$

with h_m in m (Figure 4.1.2). The values are for a medium–small city, and there are a number of other correction factors for other environments. It is interesting to note that the value of α (equation (4.1.1)) is independent of frequency and range in Hata's approximation, only dependent on base antenna height. For later comparison, the values are shown in Figure 5.1.1. The antenna heights are measured from an average ground level, and the average building heights in Japan are quoted as around 5 m (Okumura *et al*., 1968).

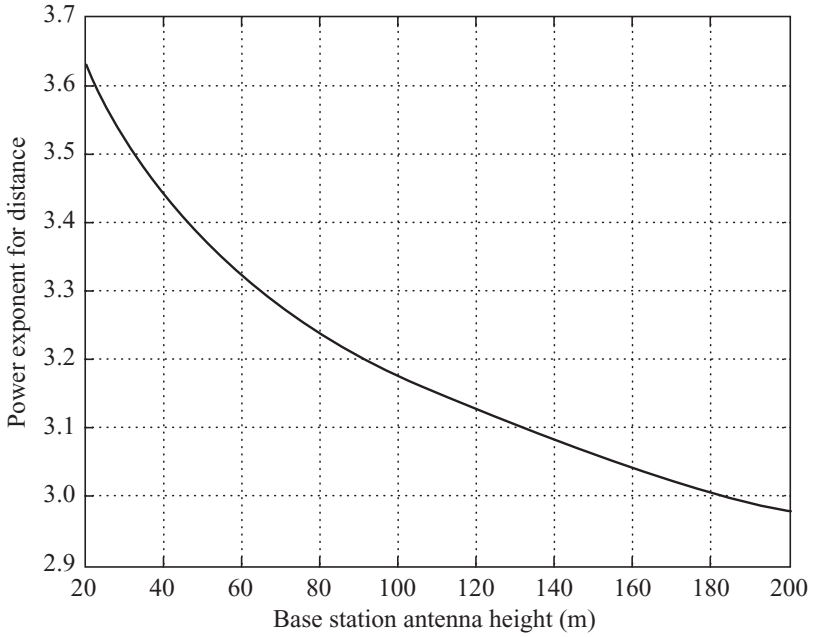


Figure 4.1.2 Power exponent for distance in Hata's approximation

The fact that the exponent is independent of frequency hints that the propagation process is not based on diffraction, but rather an incoherent process of scattering over the many buildings. It is also remarkable that the exponent is the same for all the environments, including open undulating terrain. From Section 3.3 it was noted that along-rooftop propagation, modelled as a number of screens, gave an exponent of $\alpha = 4$ independent of frequency, but for antenna heights much above the screens the result is frequency dependent.

4.1.2 2D diffraction models

Some simplifications must be made in order for the urban propagation models to be reasonably simple and amenable to calculations. The Hata models above reflect the reality (at least in Japanese cities), but they do not give any information on the basic physical mechanisms involved, and how they depend on the details of the city. A popular simplification of suburbia (or at least outside the high-rise core of modern cities) is a row of equidistant screens of equal height. Though such regularity may only be found in certain environments, it is nevertheless a good model for explaining many factors. As discussed in Section 3.3 on diffraction, this multiplicity of screens of equal heights is a difficult electromagnet problem, only solved exactly (Lee, 1978) for the situation where the antennas are at the same height as the edges. The problem of antennas above rooftops was studied by Walfisch and Bertoni (1988) by direct numerical evaluation of the Kirchhoff–Huygens integral. The results were later extended by Xia and Bertoni (1992) to include cylindrical waves and using a faster iterative way of calculating the integrals. The results may be summarised as follows. The excess

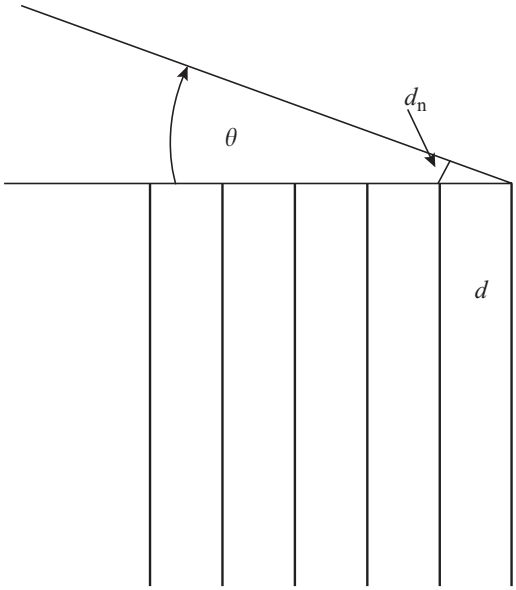


Figure 4.1.3 Geometry of the Walfisch–Bertoni model for urban buildings as an array of screens

attenuation (for fields relative to free space) depends on the local angle of incidence (θ in Figure 4.1.3), the spacing and the wavelength through one parameter, g, raised to a power n, where n is close to unity. The excess attenuation is

$$g^n = \left(\sqrt{d/\lambda}\,\frac{h_b}{R}\right)^n, \tag{4.1.4}$$

where n was found to be 0.9 in Walfisch and Bertoni (1988), and approached 1 in Xia and Bertoni (1992) for long distances; d is the spacing between the screens. The result is an α between 3.8 and 4 for the power range exponent, somewhat higher than the values given by Hata. That the factor g is the relevant factor is not surprising considering the situation (Figure 4.1.3).

If the incident wave is a plane wave and the array of screens infinite, then for a given angle θ the field must have a certain value, which is also approached when the number of screens is sufficiently large that the end of the array is outside the first Fresnel zone (see Appendix D). Walfisch and Bertoni call this value the settled field. The diffraction must depend on the distance from the line of sight to the next edge, d_n, and the ratio between the distance and the distance to the Fresnel zone (or transition region), d_{Fr}, is

$$\frac{d_n}{d_{Fr}} = \frac{d\sin\theta}{\sqrt{d\lambda/2\pi}} - \sqrt{\frac{d}{\lambda}\frac{h}{r}}, \tag{4.1.5}$$

which equals the factor g above. When g is near 1 only a single edge is involved, but normally in urban communications g is quite small.

Another comparison we can make is the dependency on the base antenna height. Rearranging terms in equation (4.1.2) we find for the Hata model that the base station

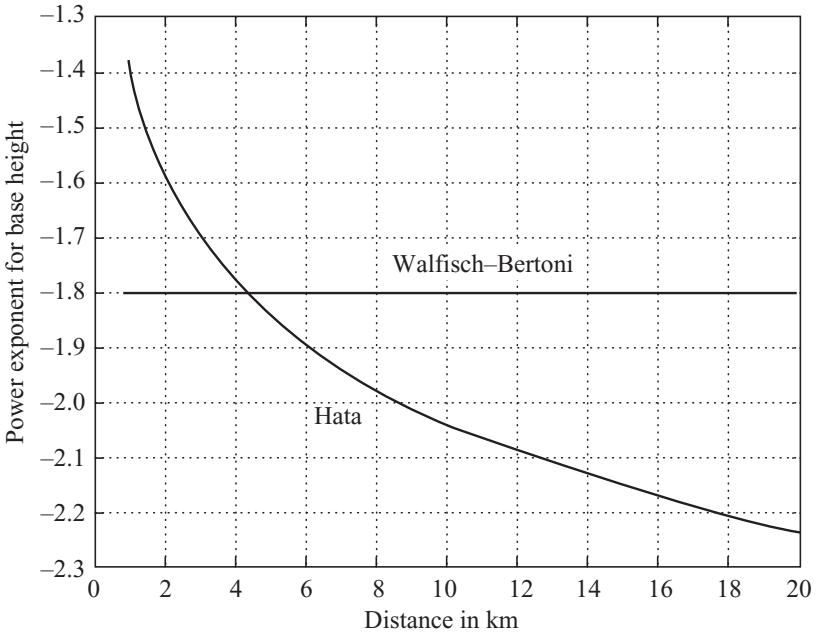


Figure 4.1.4 Power exponent for dependence of base antenna height for the Hata model. The Walfisch–Bertoni model gives −1.8

height power law exponent is

$$\alpha_{\mathrm{h}} = (1.382 + 0.655 \log (R)), \tag{4.1.6}$$

i.e. range dependent, where the Walfisch model gives $\alpha_{\mathrm{t}} = -1.8$ independent of range. Figure 4.1.4 shows the Hata result.

It seems that multiscreen diffraction can explain some of the features of the experimental results, but not all. The multiscreen model has generated considerable interest, although the main deficiency is the regularity of the structure in contrast to the more random structure of many cities.

Another way of solving the multiscreen problem is by using UTD (Section 3.4) especially involving slope diffraction for the cases of interest here. The main advantage is that the screen positions may be arbitrary (with the exception of them being very close), and that it adds additional physical insight. In Figure 4.1.5 the case of an elevated antenna is shown with the slope of $\alpha = 3.8$ as with the Walfisch–Bertoni model. The antenna has a height of 2 m above the equally spaced screens. The straight line is the constant slope, while the other curve is the UTD solution on top of the screens. The base station is elevated 2 m above the screen level (rooftops).

Street crossings and propagation along streets with buildings are covered by Blaunstein (1999).

The calculations for Figure 4.1.5 were related to field strengths at the edge of the screen level, i.e. at rooftop level. In order to get down to street level another diffraction is needed. It is noted that the slope field is an important part of the total field when at

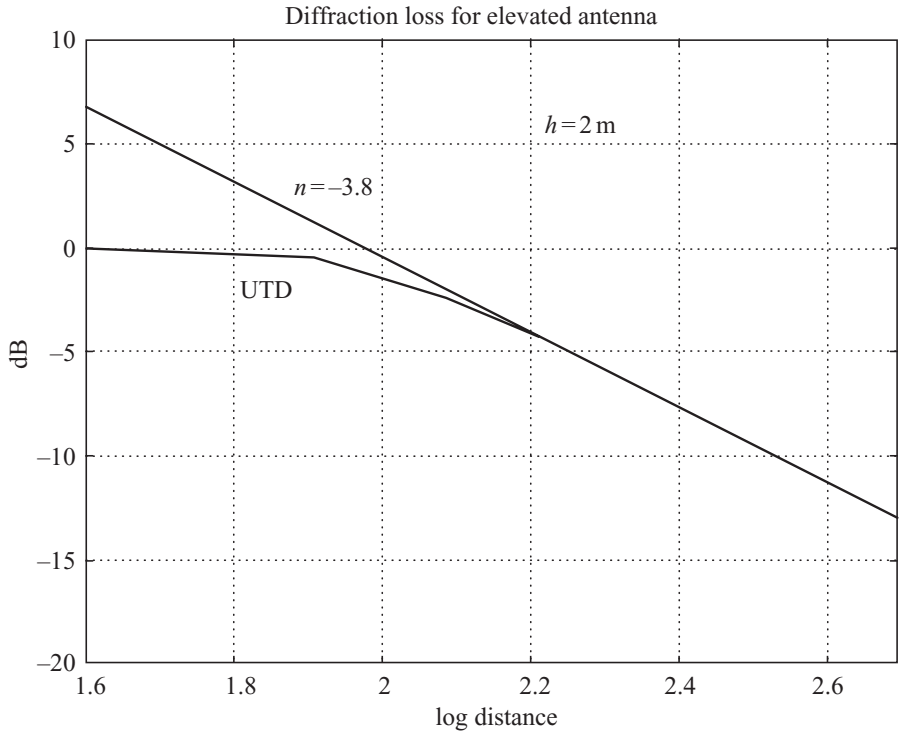


Figure 4.1.5 UTD diffraction over a regular array of screens using slope diffraction

rooftop level, but that it decays much faster away from the transition region, so that it is negligible at street level (Figure 4.1.6).

4.1.3 Bridged knife edges

It was shown in Section 3.4 for the wedge diffraction that the result depended quite heavily on the reflecting properties of the surfaces, especially when the wedge angle is near π. It can therefore be expected that the multiscreen model will fail when there is some material between the edges. The comparison has been shown by Whitteker (1990, 1993), who has also indicated how the reflecting surface can be replaced by another knife-edge using Babinet's principle. In fact, a piecewise linear model of the contour can be built up using those principles. The result of filling out the space with linear sections with reflection coefficient 1 is illustrated in Figure 4.1.7 (Whitteker, 1993), where differences of several dB relative to the knife edge solution are apparent.

4.1.4 COST231 model

A European working group has developed a model for urban environments, which incorporates many of the features above (COST231, 1999). It is also called COST231–Walfisch–Ikegami model (COST-WI). The Ikegami part is related to the diffraction and scattering down to street level. The model is stated with reference to the notation in Figure 4.1.8.

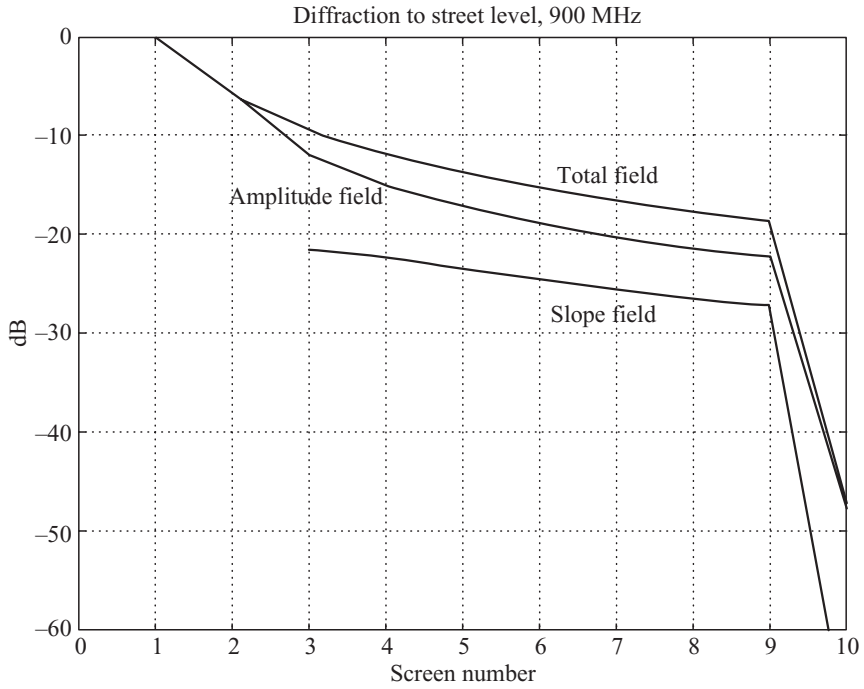


Figure 4.1.6 Propagation along nine screens and then diffraction down to street level

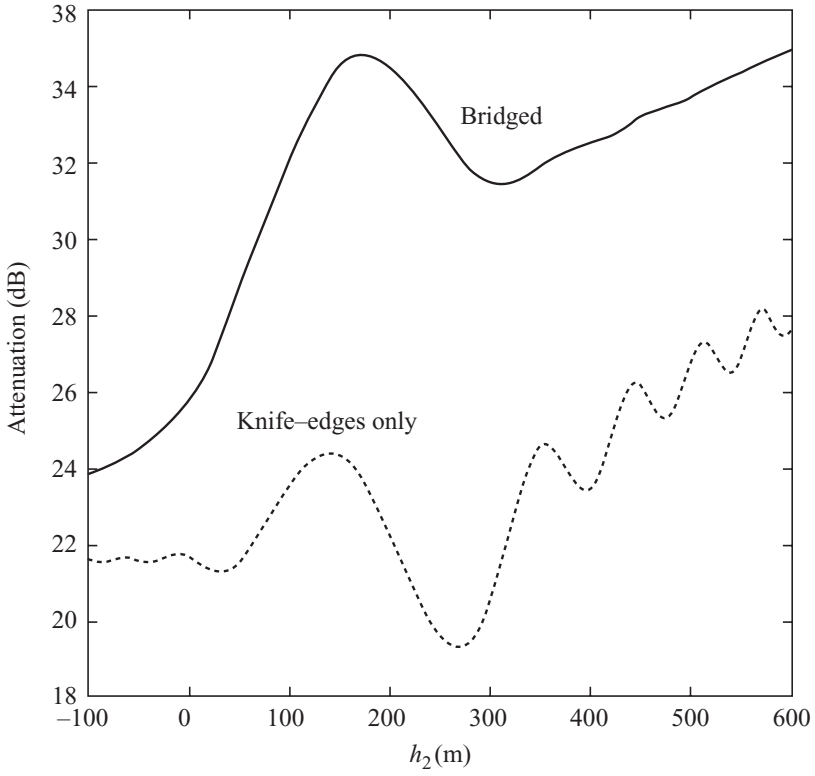


Figure 4.1.7 The attenuation due to three knife-edges, with and without reflecting bridges between the knife-edges. The height of the middle knife-edge varies (Whitteker, 1993)

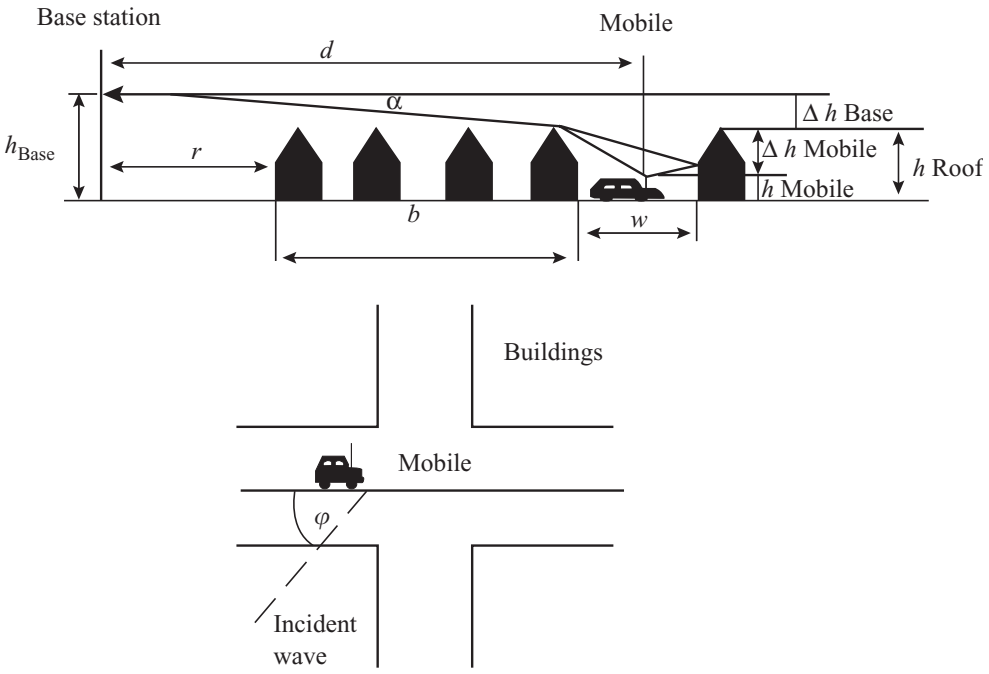


Figure 4.1.8 Geometry for COST-WI model with definition of parameters

Only the non-line-of-sight model is given. The basic path loss is given by three terms:

$$L_{\mathrm{b}} = L_0 + L_{\mathrm{rts}} + L_{\mathrm{msd}}, \tag{4.1.7}$$

where L_0 is the free-space path loss, L_{rts} is the roof-to-street loss, and L_{msd} is the multi-diffraction loss.

$$L_{\mathrm{rts}} = -8.8 + 10\log(f) + 20\log(\Delta h_{\mathrm{mobile}}) - 10\log(w) + L_{\mathrm{ori}} \tag{4.1.8}$$

with f in MHz and distances in m, $\Delta h_{\mathrm{mobile}}$ is the height of the mobile below the rooftop, and L_{ori} is a street orientation function

$$L_{\mathrm{rts}} = \begin{cases} -10 + 0.354\phi & 0 \le \phi < 35^\circ \\ 2.5 + 0.075(\phi - 35) & 35^\circ \le \phi < 55^\circ \\ 4.0 - 0.114(\phi - 55) & 55^\circ \le \phi < 90^\circ. \end{cases} \tag{4.1.9}$$

It should be noted that the Ikegami model assumes $w \gg h$, which is not always the case, and the model overestimates the loss. A UTD diffraction from rooftop to street level is more accurate, the difference being close to 3 dB when $w > h$. The multi-diffraction is given by

$$L_{\mathrm{msd}} = L_{\mathrm{bsh}} + k_{\mathrm{a}} + k_{\mathrm{d}}\log(d(\mathrm{km})) + k_{\mathrm{f}}\log(f) - 9\log(b), \tag{4.1.10}$$

where

$$L_{\mathrm{bsh}} = \begin{cases} -18\log(1 + \Delta h_{\mathrm{Base}}) & h_{\mathrm{Base}} > h_{\mathrm{Roof}} \\ 0 & h_{\mathrm{Base}} \le h_{\mathrm{Roof}} \end{cases}$$

$$k_a = \begin{cases} 54 & h_{Base} > h_{Roof} \\ 54 - 0.8\Delta h_{Base} & d \geq 0.5\,\text{km},\ h_{Base} \leq h_{Roof} \\ 54 - 0.8\Delta h_{Base} d(\text{km})/0.5 & d < 0.5\,\text{km},\ h_{Base} \leq k_{Roof} \end{cases} \tag{4.1.11}$$

$$k_d = \begin{cases} 18 & h_{Base} > h_{Roof} \\ 18 - 15\Delta h_{Base}/h_{Roof} & h_{Base} \leq h_{Roof} \end{cases}$$

$$k_f = -4 + \begin{cases} 0.7(f/925 - 1) & \text{medium-sized city} \\ 1.5(f/925 - 1) & \text{metropolitan centre} \end{cases}$$

$$\Delta h_{Base} = h_{Base} - h_{Roof}.$$

4.1.5 A comparison of methods

The multi-obstacle problem may be attacked by several different methods, most of them relying on the two-dimensional model where lateral contributions are neglected. In several cases when the Fresnel–Kirchhoff approximation is used, the slopes are assumed to be small, so the paraxial approximation can be made, and this can be justified. The main problem with the Vogler-type solution is the repeated integrations of Fresnel-type integrals going from one aperture to the next. Looking at the problem as a strict numerical problem, it may be eased by applications of the Fast Fourier Technique, since a convolution-type integral is involved. This type of solution grows only linearly with distance, and quite a rough sampling distance can be used. In this respect the method has similar features to a numerical solution of the differential equations directly in the parabolic approximation (Levy, 1990). Both techniques may be quite fast. The ray methods, augmented with slope diffraction in transition regions, have the advantage of not being dependent on the paraxial approximation. In the transition regions the ray methods have the tendency to grow faster than linearly with distance, because the incident fields must be back-tracked to the source. The filling in of media between the screens is not a problem in the ray methods, since it corresponds to using wedges instead of half-planes. The ray methods have the additional advantage of giving some physical insight into the problems. For network planning in mobile communications there is a compromise between accuracy and speed. Standard deviation of errors of 6–8 dB is usually sufficient, so there is no need to spend a lot of computer time on more details than necessary. Variations of the Hata type modelling will probably be sufficient in most cases.

4.2 Propagation in rural and hilly terrain

The situation over open land is not too different from the urban situation treated in Section 4.1. Historically, the environment was simplified to a number of knife-edges, where the two-edge case was first treated by Millington *et al*. (1962). Other techniques developed later are those of Deygout (1966, 1991) and Giovaneli (1984). The Deygout approach is that of successive evaluations of most important obstacles,

using single-edge diffraction for each case in turn. In case two edges are too close in heights, a Millington correction factor for two edges is used. Seen in a ray perspective it corresponds to a GTD model with an approximate consideration of the transition zone, but it lacks the rigour of UTD slope diffraction and it does not satisfy reciprocity.

4.2.1 Open undulating terrain

Okumura has a correction factor to the urban case for open terrain, and it has been modelled by Hata to be subtracted from the general path loss model of Section 4.1.1. It turns out that the open land has a loss, which is typically 25–30 dB less than that of the urban case. This is usually called a clutter correction factor for a particular type of environment. It is noteworthy that the general decay with distance is unchanged $(\alpha - 4)$, and similar to the two-path model after the breakpoint. The scattering mechanisms may, however, be quite different. The earlier discussions of the two-path model gave a frequency-independent field strength over flat ground. The total frequency dependence of the excess path loss for the Okumura–Hata model in the open terrain is only 1–2 dB over a wide frequency range, so it is, in effect, frequency independent. However, having experimental data agree with a theory does not prove the theory; there may be more theories giving the same dependence. In fact, rough surface scattering is also close to being frequency independent.

The Hata data cannot be expected to be valid in all environments. A large number of data have been accumulated for suburban environments by Erceg *et al*. (1998), including hilly and heavily wooded terrain, where Hata underestimates the path loss. The main contribution is that the Hata parameters themselves are stochastic variables, such that the modified model looks like

$$PL = A + 10\gamma \log(d/d_0) + s; \quad d \leq d_0 \tag{4.2.1}$$

where $A = 20\log(4\pi d_0/\gamma)$, $d_0 = 100$ m, and $\gamma = (a - bh_b + c/h_b) + x\sigma_\gamma$, where σ_γ is the standard deviation of γ, and x is a zero-mean Gaussian variable of unit standard deviation. Similarly $s = y\sigma$ is the shadow fading component with σ as the standard deviation and y like a zero-mean Gaussian variable with unit standard deviation; σ itself is a random variable with mean μ_σ and standard deviation σ_0. The parameters follow from Table 4.1.

The Hata theory gives an average decay without accounting for terrestrial undulations. This tends to give rather large standard deviations of errors, so there is an interest in trying to include the undulations. This can be done with the theories described in the previous section. Here a closer look will be given to an exact integral equation formulation.

4.2.2 An integral equation formulation

All the above methods had some sort of theoretical limitation, which makes it worthwhile to study a simulation technique of a greater accuracy. If the technique can be justified by experiment, then it is much simpler to study variation of parameters than to do additional experiments. Such a method can be derived through the use of

Table 4.1 Parameters for the general suburban model (Erceg et al., 1998), equation 4.2.1

Model parameter	Hilly/Moderate to heavy tree density	Hilly/Light tree density or Flat/Moderate to heavy tree density	Flat/Light tree density
a	4.6	4.0	3.6
b (m^{-1})	0.0075	0.0065	0.0050
c	12.6	17.1	20.0
σ_γ	0.57	0.75	0.59
μ_σ	10.6	9.6	8.2
σ_σ	2.3	3.0	1.6

integral equation methods, although some approximations must be made in order to have a reasonable computation time for distances of thousands of wavelengths. In the following, the method of Hviid *et al.* (1995) is described, where the boundary condition is that of a perfect magnetic conductor, which was seen earlier to be a good approximation for both polarisations, finite-electrical conductivity and small grazing angles. Introducing equivalent surface current density

$$\boldsymbol{M} = -\hat{\boldsymbol{n}} \times \boldsymbol{E} \tag{4.2.2}$$

$$\boldsymbol{M} = T\boldsymbol{M_i} + \hat{\boldsymbol{n}} \times \frac{T}{4\pi} \int_s \boldsymbol{M} \times \nabla' \phi \, \mathrm{d}s', \tag{4.2.3}$$

where ϕ is the free-space Green's function (in 3D), T is 2 if the point is on the surface and 1 elsewhere, $\hat{n}$ is the outward normal to the surface, $\boldsymbol{M}_i$ corresponds to the field from the source in free space, and the remaining term represents the scattered field from the surface. This could in principle be solved numerically by matrix methods, but this is not necessary in this case, since back-scattering often may be neglected for a relatively smooth surface, concentrating on the forward scattered field. Also recall that there is no interest in predicting the fast fading, only the forward slow fading. It may be shown that contributions outside the two antennas can be neglected, which turns the integral equation into a simple integral. The method is illustrated by Figure 4.2.1, showing that all multiple forward scattering terms are included, but not the backward ones. All reflections, diffractions and shadowing are included.

One final assumption is the two-dimensional character of the surface with no variations transverse to the direction of propagation. This turns the integral into a line integral and discretising, the following algorithm results

$$M_n = T M_{n,i} + \frac{T}{4\pi} \sum_{m=0}^{n-1} M_m f(n, m) \nabla l_m, \tag{4.2.4}$$

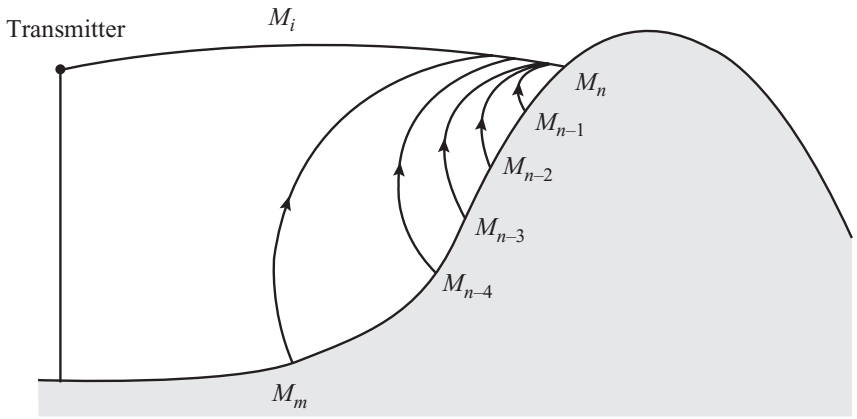


Figure 4.2.1 Illustration of the multiple scattering algorithm

where

$$
\begin{aligned}
f(n,m) &= (\hat{n}\cdot\hat{r}_2)\sqrt{\frac{R_1}{\lambda R_2(R_1+R_2)}}\,\mathrm{e}^{-\mathrm{j}(kR_2-\pi/4)} \\
R_1 &= \sqrt{s_m^2+(\tau_m-h_1)^2} \\
R_2 &= \sqrt{(x_n-x_m)^2+(h_2-\tau_m)^2}.
\end{aligned}
\tag{4.2.5}
$$

The vector r_0 is a unit vector from the source point m to the observation point n. In order to calculate the field above the surface the surface current is first determined by having $T = 2$ and the points on the surface. After that h_2 (the receiving antenna height) may be chosen arbitrarily.

The method includes all the forward reflections and diffractions. The disadvantage is that the step size is less than a wavelength, and that the computation time increases quadratically with distance, since the integration must start from the beginning each time. However, there are ways of speeding up the algorithm by simplifying the far field from a block of data (Brennan and Cullen, 1998).

The method has been compared with experimental results obtained in Denmark (Hviid *et al*., 1995), and in Figure 4.2.2(a) and (b) are shown the results for 144 MHz and 970 MHz respectively. In Figure 4.2.3 the standard deviation of the error is shown.

It is apparent from the results that the deviations around the mean slope are highly correlated with the profile with a positive correlation coefficient (in general the power increases when going uphill). This may be utilised in a fast modelling routine where the Hata model is supplemented with the height variations. An example is given in Figure 4.2.4, where an artificially generated Gaussian surface is illuminated at 150 MHz from a base station at 30 m height. The standard deviation of the surface is 5.9 m, and the smallest spatial wavelength is 100 m. The result is a slope for the distance of 43 (dB decade^{-1}) and an offset with respect to free space of 19.7 dB. The power has approximately a lognormal distribution around the slope with a standard deviation of 12 dB, so it makes sense to use a transformation factor from m to dB, in

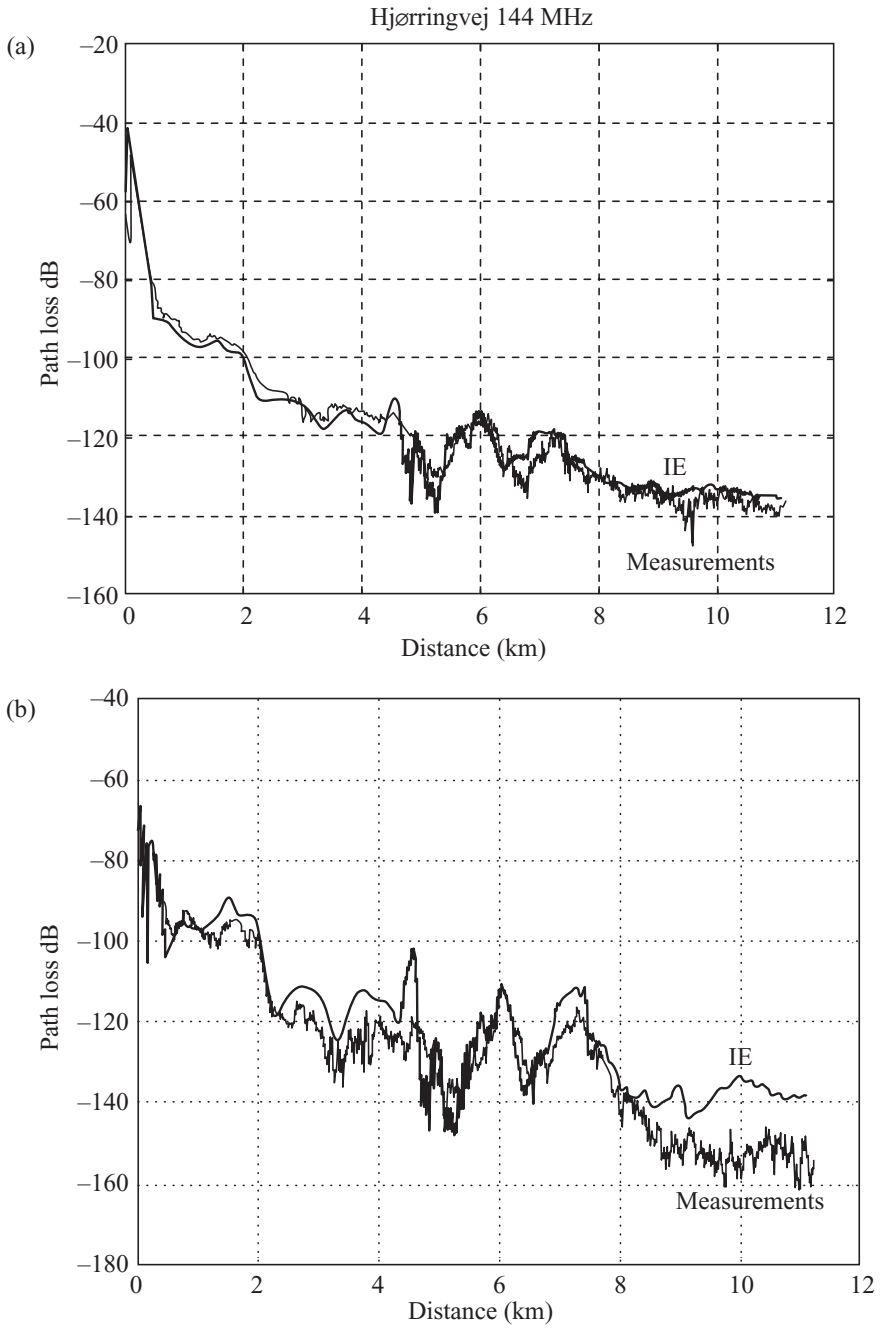


Figure 4.2.2 Experimental path 'loss' along undulating road at (a) 144 MHz and (b) 970 MHz. Measured results are compared with the integral equation method

this case a factor of two. The profile needs to be shifted slightly back (here 10 m) in order to minimise the error which is only 5 dB. So in this particular case the Hata error is reduced by 7 dB with a minimum of calculation. However, many more analyses need to be done before all the parameters are known.

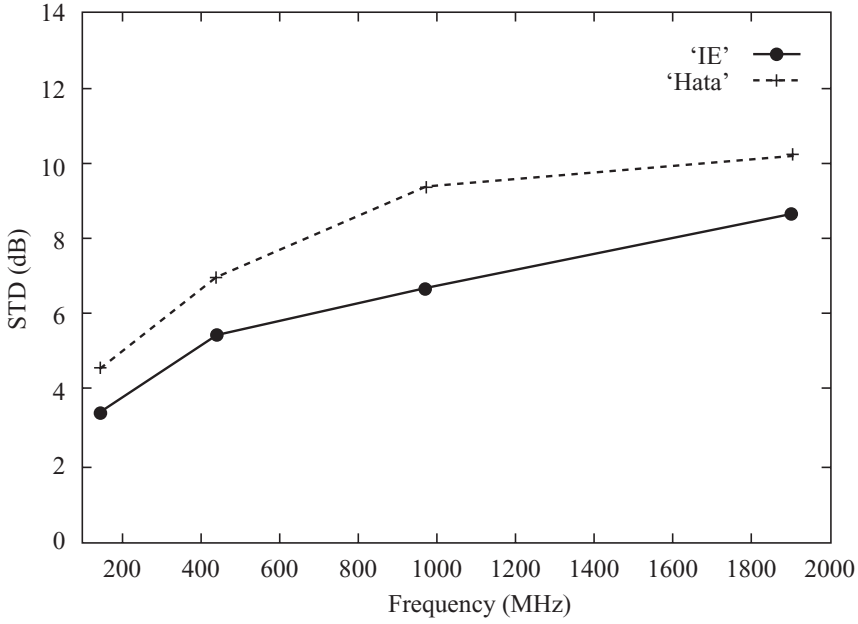


Figure 4.2.3 Standard deviation of dB error for two models compared with experiment. The methods are integral equations (lower curve) and the Hata model (upper curve)

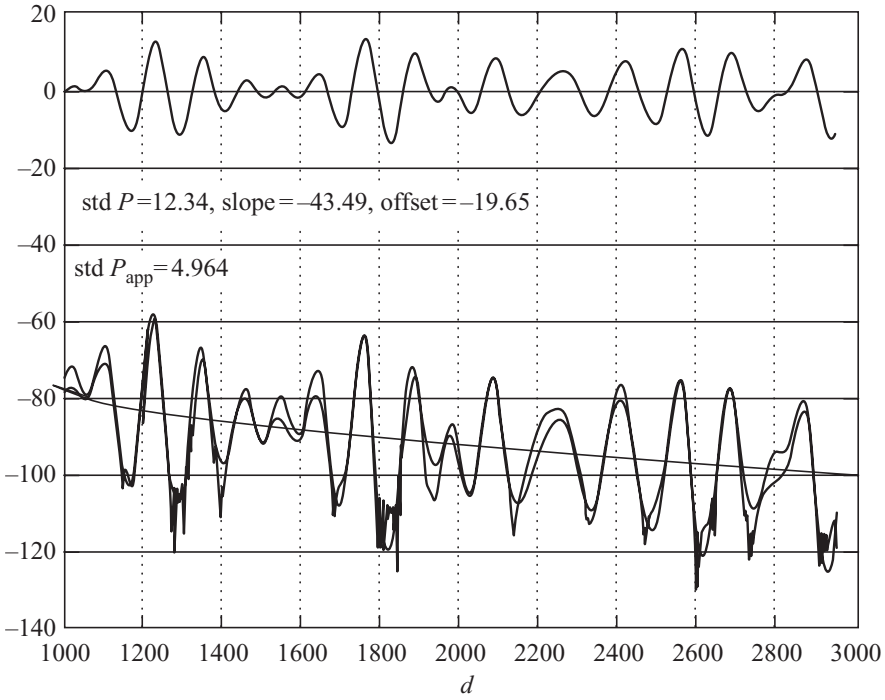


Figure 4.2.4 Utilising the correlation between the profile and the power in dB reduces the std of error from 12 to 5 dB. The smooth curve is the approximation, the other curve is calculated by means of the integral equation method

4.2.3 Combining terrain and building effects, the parabolic method

In many areas buildings are sprawling over irregular terrain and there is a need to take both into account. This is possible in an approximate way by using a wide angle version of the parabolic wave equation (Janaswamy and Andersen, 2000). Briefly speaking the parabolic equation method develops an approximation to the wave equation,

$$\frac{\partial^2 U(x,z)}{\partial x^2} + \frac{\partial^2 U(x,z)}{\partial z^2} + k_0^2 U(x,z) = 0, \tag{4.2.6}$$

where x is the horizontal coordinate and z is the vertical coordinate for a 2D situation. Pulling out the fast phase variation by introducing ψ,

$$U = \psi \frac{1}{\sqrt{x}} \mathrm{e}^{-\mathrm{j}kz}. \tag{4.2.7}$$

The following approximate equation for ψ may be derived

$$\frac{\partial \psi}{\partial x} + \frac{\mathrm{j}}{2k_0} \frac{\partial^2 \psi}{\partial z^2} = 0, \tag{4.2.8}$$

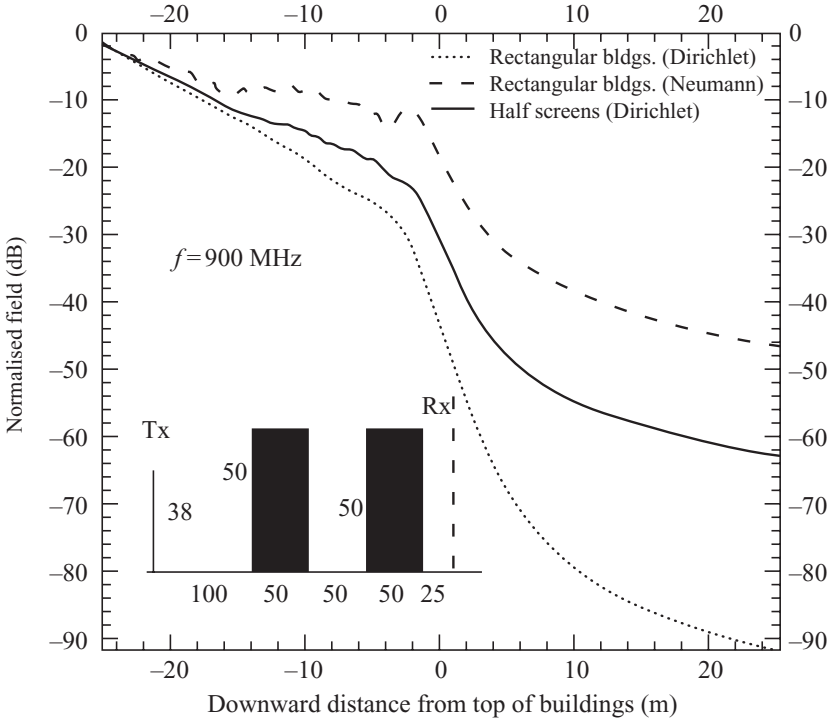


Figure 4.2.5 Field strength normalised to free space using the parabolic equation method to calculate the combined effect of buildings and terrain. Note the large dependency on the boundary condition on the top of the buildings (Janaswamy and Andersen, 2000)

where only the forward running part has been used and an almost horizontal propagation path has been assumed. The advantage is that now rather large steps may be taken in the x-direction, more determined by the environment than by the wavelength, with great savings in computer time. In Janaswamy and Andersen (2000) a wide-angle approximation is being used, showing that buildings may be covered as well if they are approximated by absorbing screens on the vertical sides and by proper boundary conditions on the top sides. An example is given in Figure 4.2.5. It is noteworthy that it makes a great difference whether the top of the buildings has a reflection coefficient of $+1$ or -1.

4.3 3D effects

All the methods above were of a 2D character, like a vertical plane cutting through the landscape or the city. It is known experimentally from antenna measurements that the field from a mobile does not come from just one direction but from a range of angles, which may vary from a few degrees to about 20°. Ray tracing is a theoretical possibility, but some limitations are necessary in order to limit practical computer times. It is, however, becoming common to have access to urban databases describing in detail the buildings and their heights allowing the use of 3D methods. A 3D approach using UTD is described in Kim *et al.* (1999) with std errors less than 8 dB and mean errors less than 7 dB. One common method is to divide the problem into two 2D problems, one with a vertical plane approach where above rooftop diffractions are included, and one with a horizontal plane giving diffraction and reflections around buildings (Kürner *et al.*, 1993).

Kürner (1999) has shown that it is of most importance near the base station, such as the first 500 m, to take into account 3D effects, and using the database to include scattering from those buildings which can be seen from both the mobile and the base, some improvement in the path loss can be obtained.

A stochastic theory taking into account the stochastic nature of urban environments has been introduced by Blaunstein (1999), and this automatically introduces 3D effects. We shall briefly state the result for power for some simple cases. As described in Section 3.2, the random case gives two components, coherent and incoherent. For the coherent case Blaunstein (1999) obtains

$$P_{\text{coh}} = \exp(-\lambda_0 d)\frac{(\sin(kh_1h_2/d)^2)}{4\pi^2 d^2}, \tag{4.3.1}$$

where

$$\lambda_0 = 2\,L\nu/\pi. \tag{4.3.2}$$

L is the average length of buildings, and ν is the density of buildings (number per unit area). The theory behind the formula relies on the probability of visibility between

two points. The sine-function is of course related to the two-path model above ground with a reflection coefficient of 1. The exponential decay may be explained by the smaller probability of line-of-sight when the distance increases. For a typical urban environment of $L = 30\,\text{m}$ and $\nu = 20\,\text{km}^2$, the 1/e distance is 260 m.

The incoherent part is caused by the scattering from the sides of buildings and the formula includes double scattering and double diffraction above rooftops

$$P_{\text{incoh}} = \frac{\Gamma^2 \lambda^3 l^3 \nu}{24\pi^2[\lambda^2 + (2\pi l_\nu \lambda_0 (h - z_1))^2][\lambda^2 + (2\pi l_\nu \lambda_0 (h - z_2))^3] d^2}, \tag{4.3.3}$$

where l_ν is the correlation distance along the surface of a scatterer, h is the building heights, $z_{1,2}$ represents the heights of mobile and base, and Γ is the power reflection coefficient. The total average power is the sum of the coherent and the incoherent power. It is interesting that in equation (4.3.3) the frequency dependence of the power can vary from f^{-3} to f^{+1}. The formulas above correspond to $1/d^2$ for free space so the usual f^{-2} of Friis' transmission formula is not included.

4.4 Indoor modelling

The indoor environment shows many of the same characteristics as the outdoor one with parametric, deterministic and stochastic models. The indoor environment is partly exposed to the normal mobile services through penetration from the outdoors (see later), but it also carries services of its own such as wireless local area networks (WLAN) at frequencies like 2.4, 5 and 17 GHz. Buildings are complicated objects in the sense that they change, doors are closed, furniture is there, people move around, and since considerable power may leave the building through windows and be scattered from the surrounding environment, the power is difficult to control. We shall anyway briefly review the most common models.

4.4.1 Parametric models

In the COST231 (1999) a number of models have been investigated along the lines of a general Hata-like power law such as:

4.4.1.1 Model 1

$$L = L_1 + 10\,n\,\log(d), \tag{4.4.1}$$

where L_1 is the path loss at a distance of 1 m, d is in m, and n is a parameter to be determined. It is difficult to attach any special physical meaning to such a model, it just has to fit to the experiments.

4.4.1.2 Model 2

Model 2 has a direct physical content being based on floor and wall attenuations (Keenan and Motley, 1990),

$$L = L_0 + L_c + \sum_{i=1}^{1} k_{wi} L_{wi} + k_f^{\left(\frac{k_f+2}{k_f+1} - b\right)} L_f, \tag{4.4.2}$$

where L_0 is the free-space loss, L_c is a constant, k_{wi} is the number of penetrated walls of type i of wall loss L_i, L_f the loss between k_f adjacent floors and b an empirical factor. The strange function of k_f expresses the observation that the total floor loss is a non-linear function of the number of floors. The various parameters are determined through a number of measurements at 1800 MHz.

Two wall types are considered:

- a light wall L_{w1}, such as plaster board, light concrete;
- a heavy wall L_{w2}, a thick wall (>10 cm) of concrete or brick.

4.4.1.3 Model 3

Model 3 expresses exponential attenuation

$$L = L_0 + \alpha d. \tag{4.4.3}$$

Model parameters may be found in Table 4.2. Comparisons with measurements show standard deviations of error around 10 dB.

It is of interest to compare with other results from other buildings at other frequencies. Devasirvatham *et al.* (1991) performed measurements in a large building (New York Stock Exchange) at four different frequencies, 0.85, 1.9, 4.0, 5.8 GHz. The path loss was matched to model 3 with the following values of α: 0.54, 0.49, 0.62,

Table 4.2 Model parameters for the three models based on measurements in a number of buildings by ten different institutions at 1800 MHz

	L_1	n	L_{w1} dB	L_{w2} dB	L_f dB	b	α dB m^{-1}
Dense							
One floor	33.3	4.0	3.4	6.9	18.3	0.46	0.62
Two floors	21.9	5.2					
Three floors	44.9	5.4					2.8
Open	42.7	1.9	3.4	6.9	18.3	0.46	0.22
Large	37.5	2.0	3.4	6.9	18.3	0.46	
Corridor	29.2	1.4	3.4	6.9	18.3	0.46	

0.55 dB m^{-1}, respectively. This is in good agreement with Table 4.2, indicating that the additional path loss on top of free-space path loss is fairly independent of carrier frequency. Janssen *et al*. (1996) have measured at three different frequencies and have used model 1 with an n of 3.3, 3.8, 4.5 at $f = 2.4$, 4.75 and 11.5 GHz, respectively, for obstructed paths. For unobstructed paths the values of n were close to two.

The fact that basic path loss is independent of frequency over a wide frequency range apart from an antenna factor is interesting in the context of adaptive antennas (Chapter 9), since effectively the basic transmission loss may decrease with carrier frequency by using an adaptive aperture instead of a fixed gain antenna.

4.5 Time domain features

The basic features of time domain modelling are described in Chapter 5, with definition of parameters like excess delay and delay spread. The wideband behaviour of mobile channels is becoming more and more important with the increasing bandwidths becoming available. We shall go through the same environments as above and emphasise the physical mechanisms.

4.5.1 Rural environments

Ray tracing is a good tool for analysing time behaviour since the delay is directly proportional to the length of the path using the speed of light in vacuum. It is also clear that the power-delay spectrum (or power-delay profile) depends on the decay of the energy towards and away from the scatterer. In this section we shall assume that the energy is propagating according to a Hata-model $P = d^{-\alpha}$, and that there is a uniform distribution of scatterers over the area, created by small hills and such.

Introducing elliptical coordinates (Andersen and Eggers, 1992)

$$\begin{aligned} x &= f(\tau/\tau_0)\cos\varphi \\ y &= f\sqrt{(\tau/\tau_0)^2 - 1}\sin\varphi, \end{aligned} \tag{4.5.1}$$

where $2f$ is the focal distance, φ is an angular parameter, and (τ/τ_0) is a parameter characterising the individual ellipse. The ellipse given by (τ/τ_0) corresponds to a relative delay

$$\frac{\tau}{L/c} = \frac{\tau}{\tau_0}, \tag{4.5.2}$$

where $L = 2f$ is the horizontal distance between the transmitter and receiver (Figure 4.5.1). With the assumption that the power propagates with an exponent α the scattered power arriving at the receiver is given by

$$\begin{aligned} \mathrm{d}P &= SE_1^{-\alpha}R_2^{-\alpha}\,\mathrm{d}x\,\mathrm{d}y \\ &= Sf^{2-2\alpha}\frac{((\tau/\tau_0)^2 - \cos^2\varphi)^{1-\alpha}}{\sqrt{(\tau/\tau_0)^2 - 1}}\,\mathrm{d}(\tau/\tau_0)\,\mathrm{d}\varphi \end{aligned} \tag{4.5.3}$$

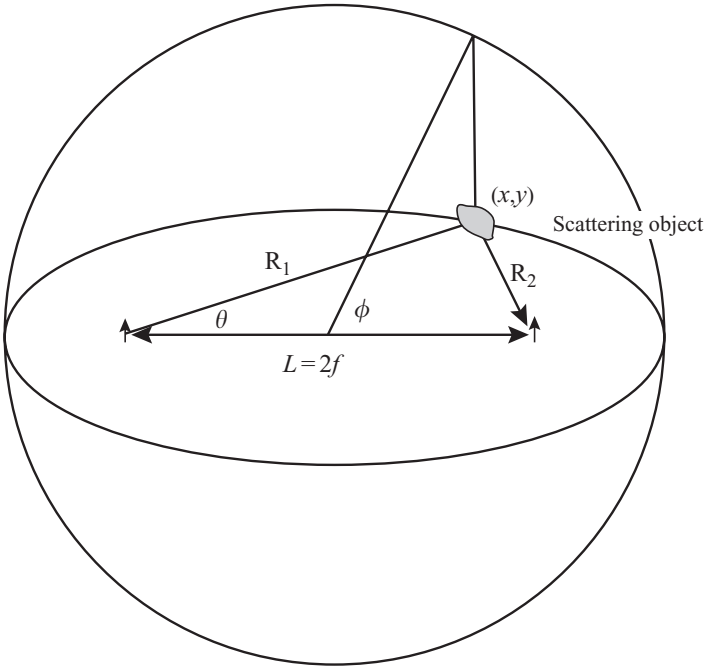


Figure 4.5.1 Scattering over a rough hilly surface with a uniform distribution of scatterers of the same scattering cross section. Scatterers lying on an ellipse will have the same time delay

and the total power at one delay is found by integrating over φ, $P(\tau) = \int_0^{2\alpha} \mathrm{d}P \cdot S$ is proportional to the scattering cross section. A numerical integration is shown in Figure 4.5.2, which has several interesting features. First, the time delay function diverges near the origin, the model no longer being valid when τ approaches unity, since the field there is more coherent, and usually dominated by the direct wave corresponding to a certain K-factor in the Rice distribution. This should be added to the incoherent power. The fact that the scattering cross section is independent of angle is also an approximation (and not in agreement with the rough surface treatment in Section 3.2). Nevertheless, the shape of the delay function is interesting because it can be considered as power law dependency in contrast to the usual exponential decay assumed.

The asymptotic behaviour of the tail for large $\tau \gg \tau_0$ is

$$P(\tau) = \tau^{1-2\alpha} \quad \tau \gg \tau_0 \tag{4.5.4}$$

so the slower the power decay vs distance, the longer the tail. In fact for $\alpha = 2$ the decay is so slow that the delay spread is infinite. It is also clear from the formulation that the delay spread s will grow linearly with τ_0, that is with distance. Figure 4.5.3 shows experimental results for a slightly rolling countryside with good agreement with the general shape of the delay curve. In other cases the delay curve will be dominated by discrete scatterers, buildings, mountain sides and such.

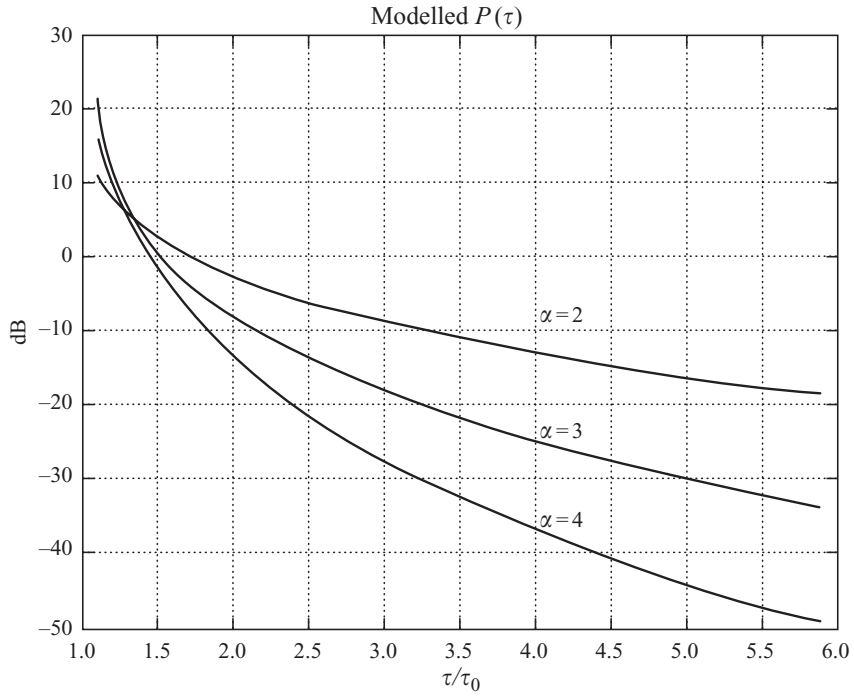


Figure 4.5.2 Power-delay profiles for a uniformly distributed scattering function assuming a power law of propagation with exponent α

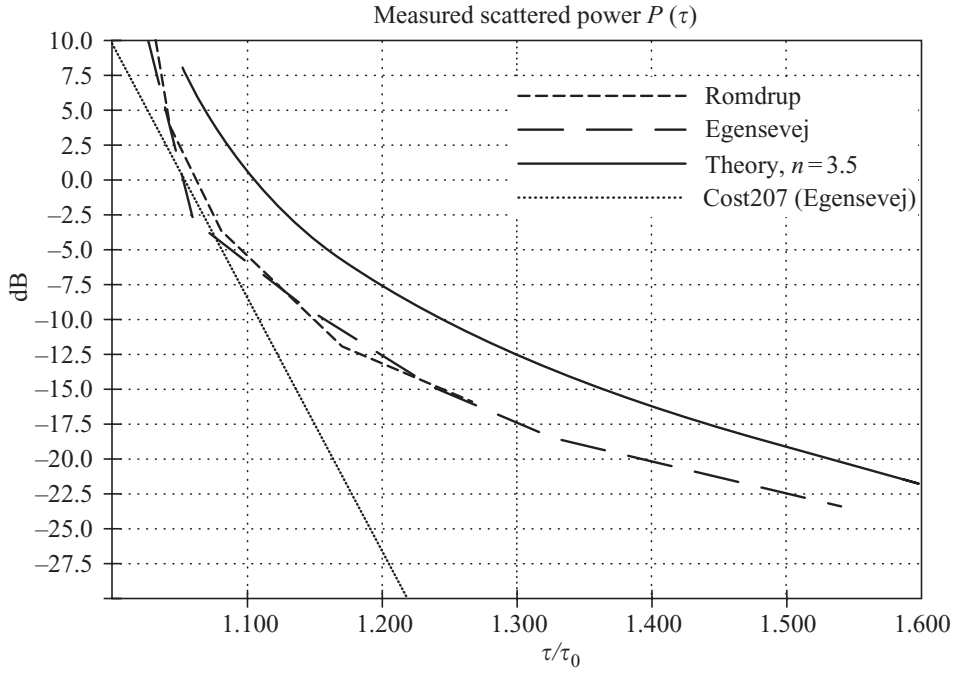


Figure 4.5.3 Power-delay profiles for a rural environment showing good agreement with the theory. The vertical displacement is arbitrary. The straight line is the COST207 exponential model

4.5.2 Urban environments

In an urban environment it may be expected that multiple reflections or multiple scatterings will be more dominant in the rural areas. This would have an impact on the shape of the power-delay profile, since multiple reflections with a certain fraction of power being lost at each impact would lead to a model with exponential decay,

$$P(\tau) = \mathrm{e}^{-\tau/s}, \quad \tau > 0, \tag{4.5.5}$$

where s is the delay spread. In urban areas the delay spread is a function of the antenna height and typically is between 0.1 and 3 μs, with larger spreads the lower the antenna above the rooftops (Pedersen *et al.*, 2000). In certain cases with distinct separate reflection areas the profile may be subdivided in several subgroups where each has an exponential decay. Examples are the well-known COST207 models (Figure 4.5.4).

The value of such standardised models lies in their common use in communication simulations. This leads us to a discrete model, the so-called Turin model (Turin

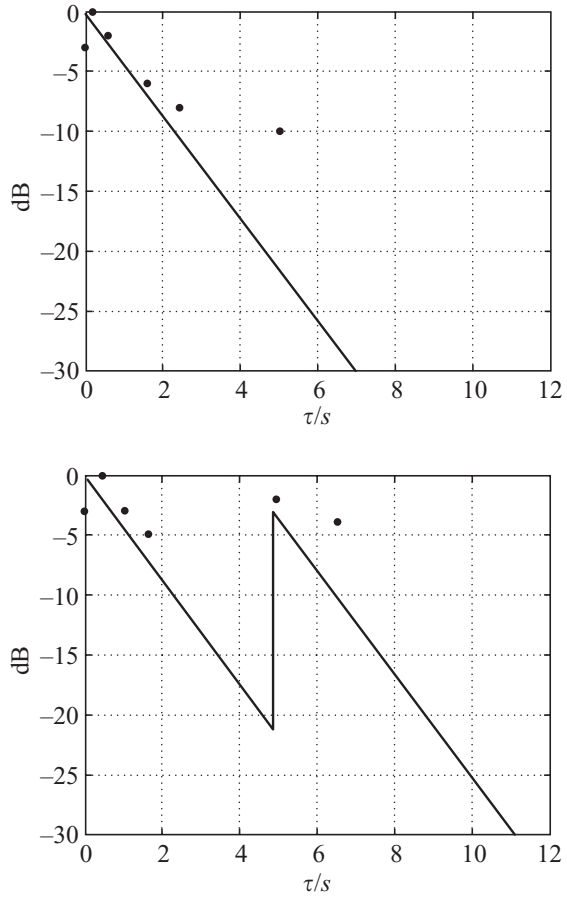


Figure 4.5.4 COST207 profiles for urban environments, typical urban (TU) (upper) and bad urban (BU) (lower). Dots represent taps of equivalent FIR filter

et al., 1972), where the individual rays or clusters of rays are seen as the basis for a propagation model.

The channel is then modelled as a number of discrete rays or multipath components

$$h(\tau) = \sum_{k=1}^{N} a_k \delta(\tau - \tau_k)\, \mathrm{e}^{j\theta_k}, \tag{4.5.6}$$

where N is the number of multipath components, a_k is the amplitude, τ_k is the delay, and θ_k is the phase. This has proved to be a good model for simulations of communication systems, since the statistics of the individual parameters can also be found. This is for a stationary channel; for a moving terminal of course the Doppler spectrum should be included. It should be remembered that the number of paths is a function of the sensitivity of the measurement system and of the bandwidth of the system. From a theoretical point of view a complex environment must give an infinite number of rays, but a given system with a given sensitivity and bandwidth can only resolve a finite number. The unresolved components in the frequency or delay domain will usually be fading with some statistics to be determined from measurements.

Channel models are often given in terms of FIR filters with constant spacing between the delay taps. For the COST207 models above, the discrete tap values for a six-tap reduced filter for the typical urban case are indicated in Figure 4.5.4. They are chosen such that the delay spread of the continuous and the discrete model is the same. The COST models are useful for GSM bandwidths or smaller, while they are not adequate for the UMTS bandwidths of several MHz used in a CDMA mode, where the strength and position of the individual Rake fingers are important for a good model.

4.5.3 Indoor environments

As for the outdoor models, in the wideband description of the indoor channel it has been found convenient to have a more detailed model than the first and second moments of the power delay profile (Hashemi, 1993). The Turin model was applied to two indoor environments, and the parameters of equation (4.5.6) (Hashemi, 1993) were found with a resolution of 5 ns. The number of multipath components N was found to increase with the sensitivity, as would be expected, but also with range. This would also be in accordance with the ray-tracing models since more and more scatterers would be encountered. The arrival times did not satisfy a simple Poisson model, but rather a modified version. In Saleh and Valenzuela (1987) a cluster model was advanced, where each cluster followed a Poisson model, as did the individual contributions within a cluster.

Indoor environments range from office buildings with partitioned offices to rather large factory environments. For the latter, Rappaport *et al*. (1991) have performed and analysed a number of measurements at 1.3 GHz. One interesting result was that the individual multipath components were lognormal fading instead of the usual Rayleigh. In some unpublished work from Aalborg University the power-delay function for

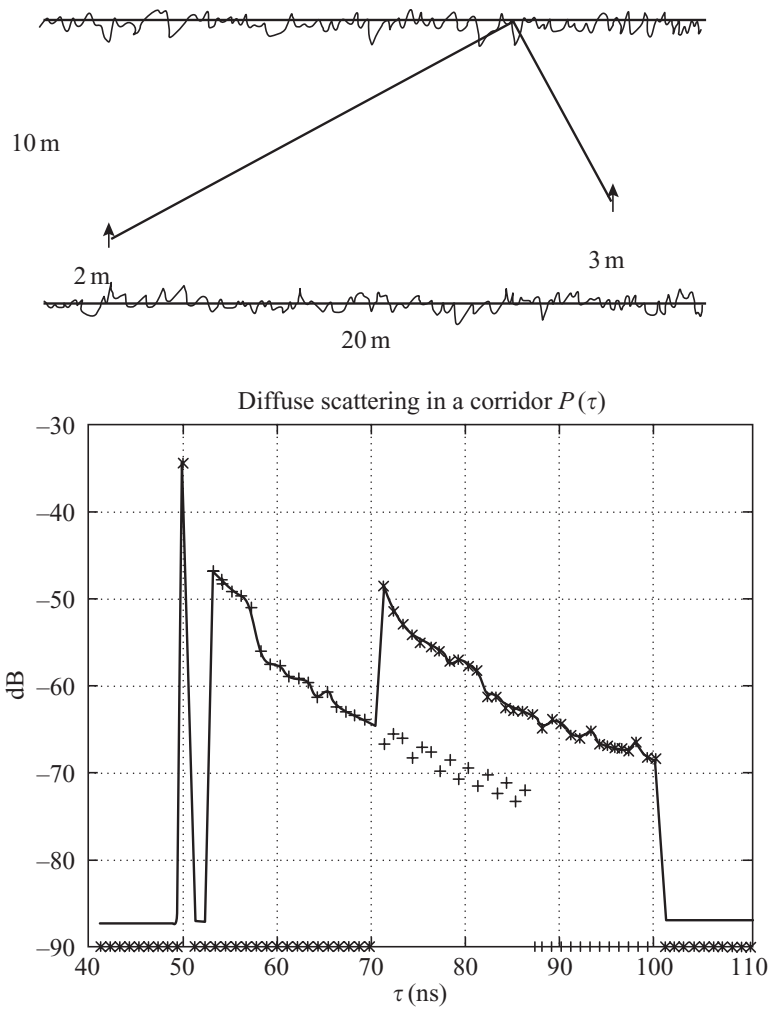


Figure 4.5.5 Diffuse scattering from two walls. Each wall contributes a continuous delay profile

the continuous model was shown to follow the power-law model of the rural area, Section 4.5.1, for a rather large storage hall.

A simple model will illustrate the problem. Consider a simplified model of rough scattering from two sides in a corridor: length 20 m, width 10 m, source 2 m from lower wall, receiver 15 m along the corridor and 3 m from the wall (Figure 4.5.5).

First, after 50 ns, the direct ray arrives. Then, the first reflections arrive from the bottom side since it is closest. The first arrival is at 53 ns, where the sum of the distances are minimum, i.e. at the reflection point of the specular ray. Smaller contributions from the rest of that wall continue up to 88 ns. At 72 ns comes the first contribution from the top, and after that they both contribute. An interesting observation is that from an experimental point of view it is difficult to distinguish between the specular reflected and the diffuse reflected field. The main difference is the width of the impulse. The impulse function is a combination of exponential decay

due to multiple scattering (not included in Figure 4.5.5) and power-law decay due to the range dependence. The results in Figure 4.5.5 are with an infinite bandwidth, where in practice it would be folded with a finite bandwidth filter. A bandwidth of about 1 GHz would be required to see the details above.

4.6 Angular domain features

The angular domain has been introduced recently for base stations to improve the interference situation (Chapter 7). The angular distribution of energy from a mobile user as seen from a base station is important for the proper understanding of the channel and for determination of the correlation properties of the antenna elements. Furthermore, a joint determination of the temporal as well as the spatial components of the signal adds greatly to this understanding.

4.6.1 Rural environments

In the same manner as for the time domain, the uniform distribution of scatterers will be studied (Figure 4.5.1). This model is assumed to be valid for a rural terrain. From the elliptical curve valid for one value of τ/τ_0 the value of θ may be found as a function of ϕ, where θ is the azimuth angle measured from the focus to the specific scattering point (Figure 4.5.1), and the power density may then be plotted as a function of θ (Figure 4.6.1).

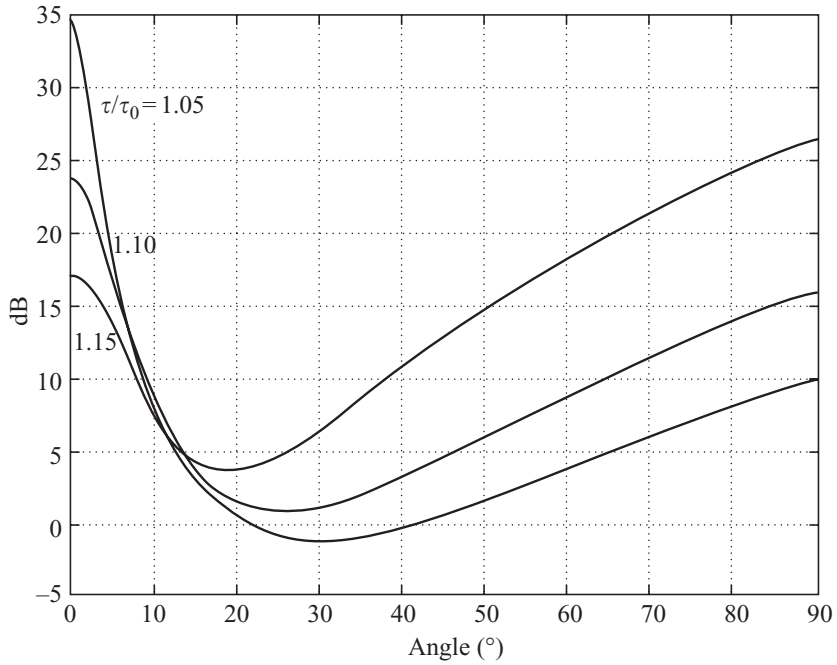


Figure 4.6.1 Angular distribution of power for different arrival times $\alpha = 4$

Recall that τ_0 is the time of arrival of the direct path so $\tau/\tau_0 = 1.05$ means that part of the energy which arrives with an excess delay of 5 per cent. For this early part the distribution is fairly narrow around the main direction, while later arrivals are fairly broad approaching a constant, since the energy will be arriving with equal probability from all directions. It is interesting to observe that the level at the base station at $\theta = 90°$ is quite high, which could be the case for an omnidirectional antenna. It should be emphasised that the direct path is not included. Also, no experimental evidence is available for the moment.

It is convenient to define an angular spread s_θ similar to the temporal spread of the power angular function

$$S_\theta^2 = \frac{\int P(\theta)\theta^2 \, \mathrm{d}\theta}{\int P(\theta) \, \mathrm{d}\theta} - \overline{\theta}^2. \tag{4.6.1}$$

The angular spread is an important parameter for understanding array properties (Chapter 9). It is clear from the derivation that the distance between the antennas does not enter the shape of the angular pattern, so the angular spread is independent of range for a uniform distribution of scatterers, neglecting again the direct path. This is in contrast to the temporal spread which increases linearly with range. In reality the situation is more complicated, since the angular spread may decrease or increase with distance depending on the environment (Figure 4.6.2). The results are based on measurements in various rural environments (Pedersen *et al*., 1998).

The results may possibly be explained by the effect of the line-of-sight path and a non-uniform distribution of scatterers.

The angular spread depends on the 'Rice'-factor of the situation considered as the ratio of the direct path versus the sum of all the scattered powers (Figure 4.6.3). In the simulation it is around 0 dB near the scatterers, but higher for smaller distances, so the angular spread increases when moving towards the scatterers. Note that the peak of the curve occurs before the centre of the 'city'.

4.6.2 Urban environments

The urban environment displays a different characteristic from the rural one. Instead of the long-range effect of scatterers, multiple reflections near the mobile and over rooftop diffractions seem to dominate, and the delay response is decoupled from the azimuth response. As shown in Pedersen *et al*. (2000) the power-delay spectrum has an exponential shape independent of the angle and the azimuthal spectrum as a Laplacian shape independent of delay, i.e.

$$P(\varphi, \tau) = P_\lambda(\varphi) P_\mathrm{D}(\tau) = \mathrm{e}^{-\tau/s} \mathrm{e}^{-\sqrt{2}|\varphi|/\sigma}. \tag{4.6.2}$$

Note that this is markedly different from the rural case (equation (4.5.3)). The above is a somewhat simplified, though common, case. Examples from Stockholm, Sweden, and Aarhus, Denmark, are shown in Figure 4.6.4 for azimuthal variations. The base antennas are above rooftops.

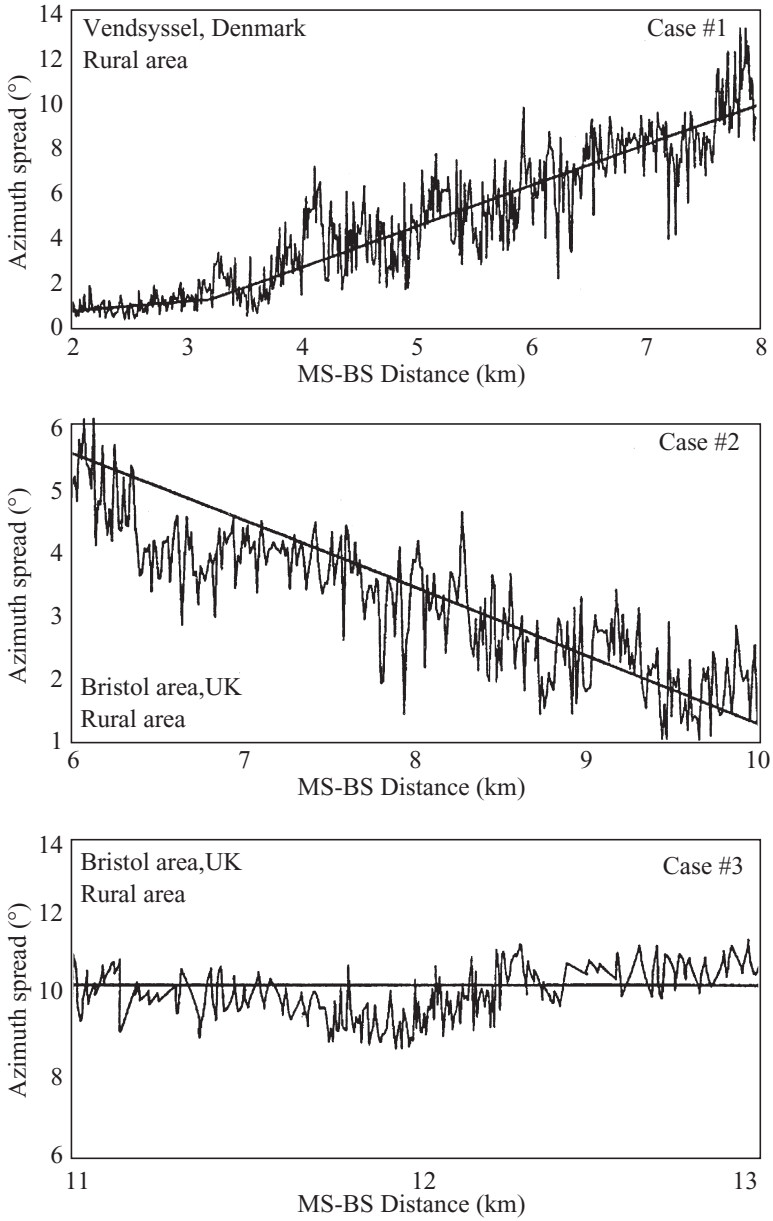


Figure 4.6.2 Angular spread versus distance for three different environments (Pedersen et al., 1998)

The exponential delay function is easily understood through multiple reflections; a certain fraction of the energy is lost at each impact. The Laplacian azimuth function, which is to be found in the indoor case as well, does not have a similar simple interpretation. The fact that it occurs in many diverse situations should imply some basic mechanism. In Section 6.3 it is shown that a simple model using the phase distribution over a small array as an indication of the direction of arrival leads to an

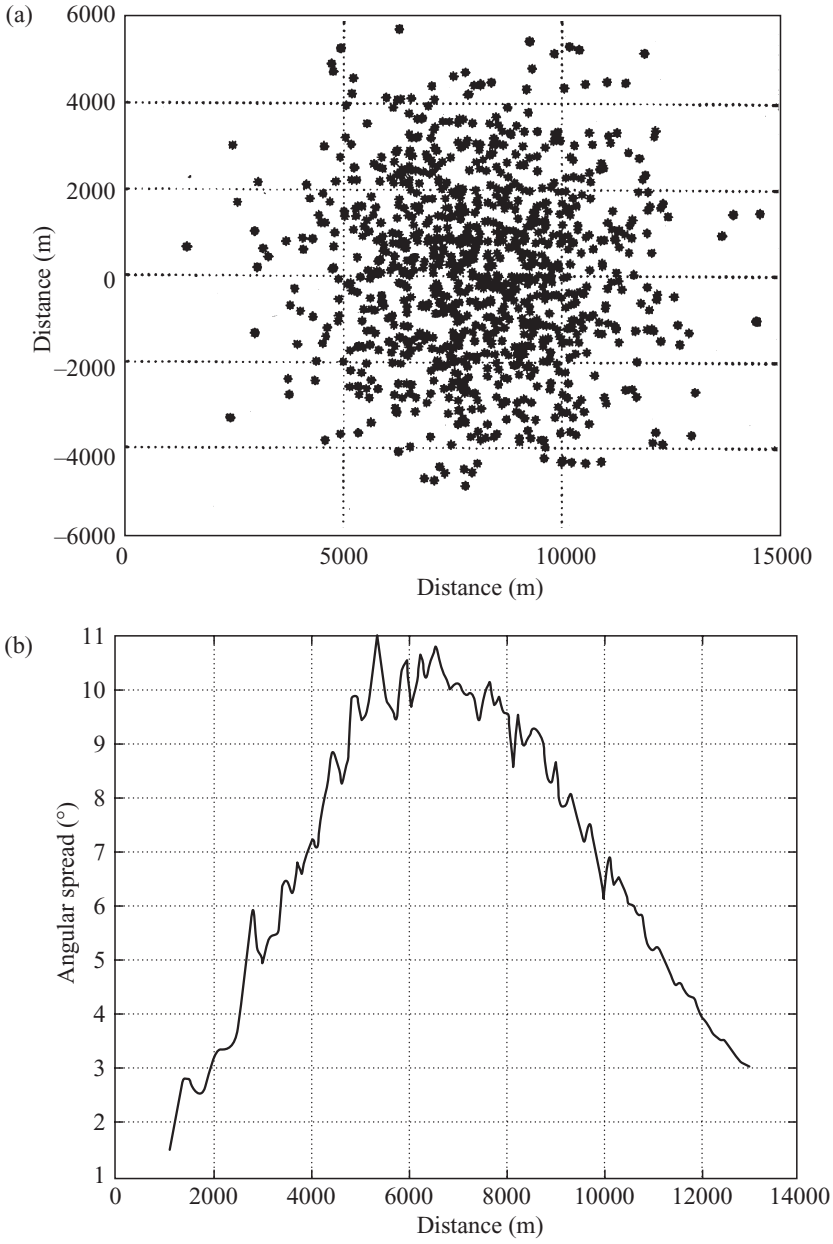


Figure 4.6.3 Variation of angular spread through a collection of scatterers. (a) Distribution of 1000 scatterers around (8000 m). (b) Average of 20 realisations of angular spread versus distance

angular distribution as

$$P(\varphi) = \frac{c}{\left(\sin^2 \varphi + s^2_{\sin(\varphi)}\right)}, \tag{4.6.3}$$

where $s_{\sin \varphi}$ is the spread of the sine of the angle. This function also agrees well with the experimental results.

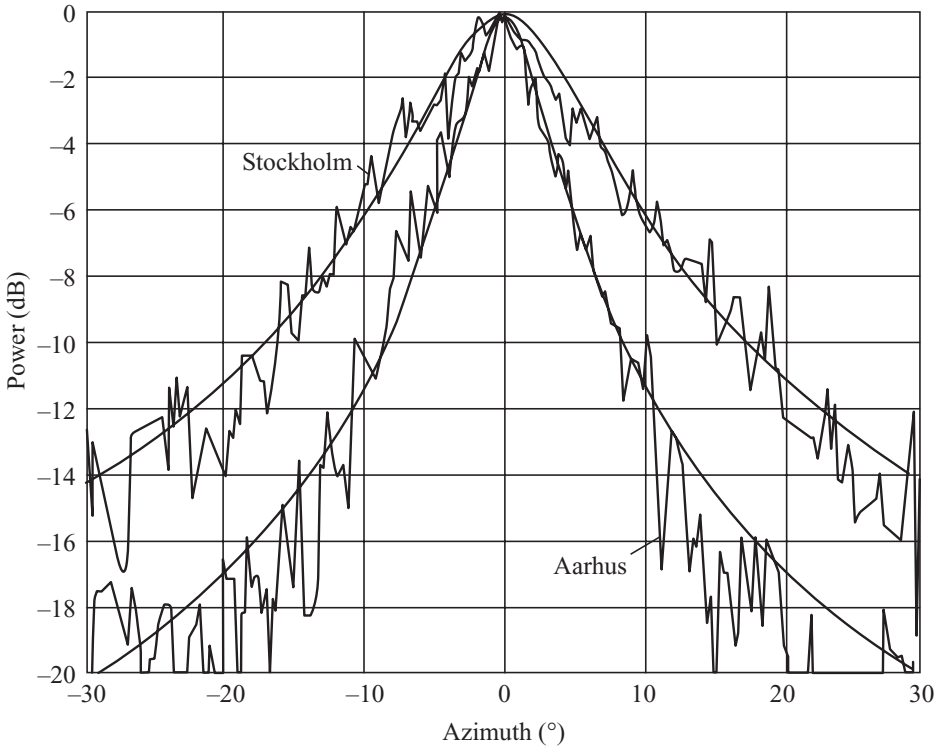


Figure 4.6.4 Power azimuth spectrum from experiments in Stockholm and Aarhus compared with equation (4.6.2) (Andersen and Pedersen, 2001)

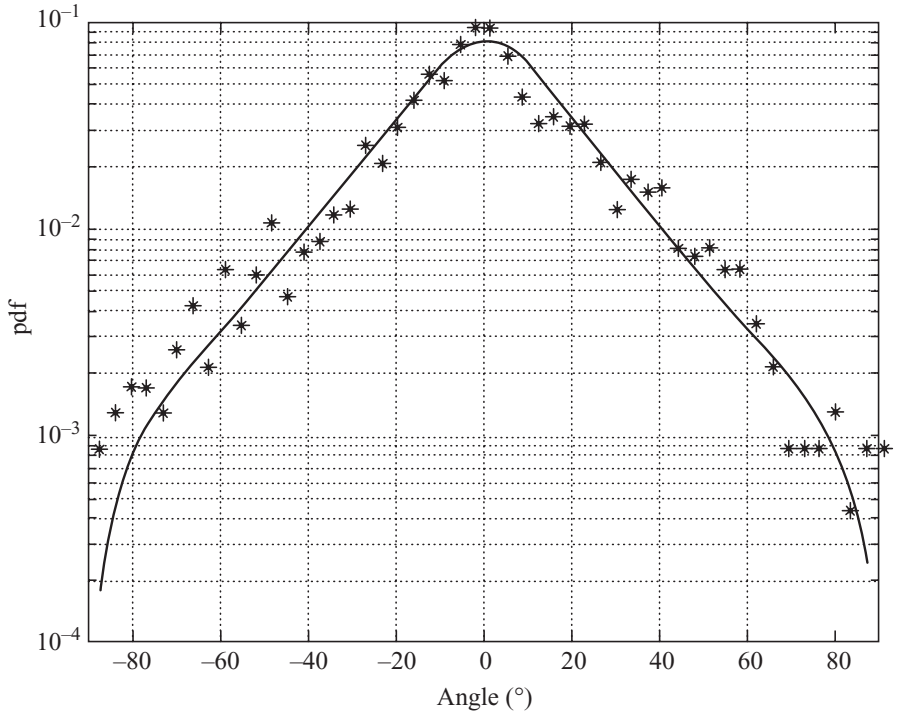


Figure 4.6.5 Probability density of angle-of-arrival in Clyde Building (Spencer et al., 1997) compared with theory, $s_u = 0.42$, $\phi_s = 24.8°$

4.6.3 Indoor environments

4.6.3.1 Experimental examples of angular power for indoors environment

Indoor measurements for two different buildings have been reported by Spencer *et al*. (1997). The scanning was done mechanically with a 6° horn at 7 GHz over a 360° range. The pdf of the observed angular power is shown for the forward range of 180° and for the Clyde Building in Figure 4.6.5. Only the forward range is shown since the back side is supposed to be reflections from a back wall. The smooth curve is the theoretical result with a best visual fit (see Section 6.3). The angular spreads are 24.8° and 17.5°, respectively. The agreement is reasonable. The antenna beamwidth is less than the angular spread, so we should expect the observed distribution to be dependent on the true (incident power) distribution to some extent. The double exponential (Laplacian) chosen in Spencer *et al*. (1997) is also an excellent fit to the data, but it lacks a theoretical explanation.

4.7 References

ANDERSEN, J. BACH and EGGERS, P.: 'A heuristic model of power delay profiles in landmobile communications', *Proc. URSI Symposium on Electromagnetic Theory*, Sydney, Australia, August 17–20, 1992, pp. 55–57.

ANDERSEN, J.B. and PEDERSEN, K.I.: 'Angle-of-arrival statistics for low resolution antennas', *IEEE Trans. Antennas and Propagation*, 2002, **50** (3), pp. 391–395.

BLAUNSTEIN, N.: *Radio Propagation in Cellular Networks*, Artech House, 1999.

BRENNAN, C. and CULLEN, L.J.: 'Application of the fast far-field approximation to the computation of UHR pathloss over irregular terrain', *IEEE Trans. Antennas and Propagation*, 1998, **43** (1), pp. 881–889.

COST231: 'Digital mobile radio towards future generation systems', final report, European Commission, 1999.

DEVASIRVATHAM, D.M.J., BANERJEE, C., MURRAY, R.R., and RAPPAPORT, D.A.: 'Four-frequency radiowave propagation measurements of the indoor environment in a large metropolitan commercial building', *Globecom '91*, December 1991, pp. 1281–1286.

DEYGOUT, J.: 'Multiple knife edge diffraction of microwaves', *IEEE Trans. Ant. Prop.*, August 1966, **AP–39**, pp. 1256–1258.

DEYGOUT, J.: '*Correction factor for multiple knife-edge diffraction*, *IEEE Trans. Ant. Prop.*, August 1991, **AP-39**, pp. 1256–1258.

ERCEG, V., *et al*.: 'An empirically-based path loss model for wireless channels in suburban environments', *Globecom '98*, 1998, **2**, pp. 922–927.

GIOVANELI, C.L.: 'An analysis of simplified solutions for multiple knife edge diffraction', *IEEE Trans. Ant. Prop.*, March 1984, **IP 32** (3), pp. 297–301.

HASHEMI, H.: 'The indoor radio propagation channel', *Proc. IEEE*, July 1993, **81** (7), pp. 943–967.

HATA, M.: 'Empirical formula for propagation loss in land mobile radio services', *IEEE Transact. Vehicular Technology*, August 1980, **VT-29** (3), pp. 317–325.

HVIID, J.T., BACH ANDERSEN, J., TOFTGÅRD, J., and BØJER, H.: 'Terrain-based propagation model for rural area – an integral equation approach', *IEEE Trans. Ant. Prop.*, January 1995, **43** (1), pp. 41–46.

JANASWAMY, R. and ANDERSEN, J.B.: 'Path loss predictions in urban areas with irregular terrain topography', *Wireless Personal Communications*, 2000, **12**, pp. 255–268.

JANSSEN, G.J.M., STIGTER, P.A., and PRASAD, R.: 'Wideband indoor channel measurements and BER analysis of frequency selective multipath channels at 2.4, 4.75 and 11.5 GHz', *IEEE Transactions on Communications*, October 1996, **44** (10), pp. 1272–1288.

KEENAN, J.M. and MOTLEY, A.J.: 'Radio coverage in buildings', *British Telecom Technology Journal*, January 1990, **8** (1), pp. 19–24.

KIM, S.O., *et al.*: 'Radio propagation measurements and prediction using three-dimensional ray tracing in urban environments at 908 MHz and 1.9 GHz, *IEEE Trans. Veh. Techn.*, May 1999, **48** (3), pp. 931–945.

KÜRNER, T., CICHON, D.J., and WIESBECK, W.: 'Concepts and results for 3D digital terrain based wave propagation models – an overview', *IEEE Journal on Selected Areas in Communications*, 1993, **11** (7), pp. 1002–1012.

KÜRNER, T.: 'A run-time efficient 3D propagation model for urban areas including vegetation and terrain effects', *Proc. VTC99*, Houston, 1999.

LEE, S.W.: 'Path integrals for solving some electromagnetic edge diffraction problems', *J. Math. Phys.*, 1978, **19**, pp. 1414–1422.

LEVY, M.: 'Parabolic equation modelling of propagation over irregular terrain', *Elec. Lett.*, 1990, **26** (15), pp. 1153–1155.

MILLINGTON, G., *et al.*: 'Double knife edge diffraction in field strength prediction', *IEE Monograph* 507E, 1962, pp. 419–429.

OKUMURA, Y., *et al.*: 'Field strength and its variability in VHF and UHF land-mobile radio service', *Rev. of the Electrical Communication Laboratory*, September 1968, **16** (9–10), pp. 825–873.

PEDERSEN, K.I., MOGENSEN, P.E., and FLEURY, B.H.: 'Spatial channel characteristics in outdoor environments and their impact on BS antenna system performance', *Proc. Veh. Tech. Confer.*, VTC '98, Ottawa, 1998, pp. 719–724.

PEDERSEN, K.I., MOGENSEN, P.E., and FLEURY, B.H.: 'A stochastic model of the temporal and azimuthal dispersion seen at the base station in outdoor propagation environments', *Trans. Veh. Techn.*, 2000, **49** (2), pp. 437–447.

RAPPAPORT, T.S., SIEDEL, S.Y., and TAKAMIZAWA, K.: 'Statistical channel impulse response models for factory and open plan building radio communication system design', *IEEE Trans. Communications*, May 1991, **39** (5), pp. 794–807.

SALEH, A.M. and VALENZUELA, R.: 'A statistical model for indoor multipath propagation', *IEEE J. Selected Areas Communications*, 1987, **SAC-5**, pp. 128–137.

SAUNDERS, S.R. and BONAR, F.R.: 'Explicit multiple building diffraction attenuation function for mobile radio wave propagation', *Electronics Letters*, 1991, **27** (14), pp. 1276–1277.

SPENCER, W., RICE, M., JEFFS, B., and JENSEN, M.: 'A statistical model for angle of arrival in indoor multipath propagation', *Proc. VTC '97*, May 1997, pp. 1415–1419.

TURIN, G.L., *et al.*: 'A statistical model of urban multipath propagation', *IEEE Trans Veh. Technology*, February 1972, **21**, pp. 1–9.

WALFISCH, J. and BERTONI, H.L.: 'A theoretical model of UHF propagation in urban environments', *IEEE Trans. Ant. Prop.*, December 1988, **36** (12), pp. 1788–1796.

WHITTEKER, J.H.: 'Fresnel Kirchhoff theory applied to terrain diffraction problems', *Radio Science*, September–October 1990, **25** (5), pp. 837–851.

WHITTEKER, J.H.: 'A series solution for diffraction over terrain modelled as multiple bridged knife edges', *Radio Science*, July–August 1993, **28** (4), pp. 487–50.

XIA, H.A. and BERTONI, H.L.: 'Diffraction of cylindrical and plane waves by an array of absorbing half-screens', *IEEE Trans. Ant. Prop.*, February 1992, **41** (2), pp. 170–177.

Chapter 5

Short-term channel behaviour from the two-path model

5.1 Introduction

The waves incident on a mobile receiving antenna sometimes arrive via a line-of-sight from the transmitter, but in a general case they emanate mostly from scatterers. The scatterers can be viewed as electromagnetic sources. The *scenario* comprises these sources and their geometric relation to the receiving antenna.

Precise scenarios are unique to a physical environment and the location of the receiver; however, it is useful if scenarios can be modelled simplistically. Such models are developed in order to help understand the basic processes of the channel, and ultimately to help define the design requirements of the antennas and signal processing required for a communications link.

For fixed pattern antenna design, the receiving pattern should be matched as well as practicable to a time- (or location-) averaged scenario. For diversity antenna design, a similar match is required for the composite pattern, or field-of-view, of the available patterns of an adaptive antenna. In digital communications, this match is not necessarily based solely on gathering the maximum signal power, but rather on achieving the lowest bit-error rate in the context of existing system constraints. There are few experimental data on average scenarios, but models can be postulated based on the propagation and scattering principles. Some parameters can be estimated from measurements.

In this chapter, scenario and propagation channel models are developed. The deterministic two-path model is used extensively as a model since it offers a simple tutorial and can produce representations of most of the behaviour from real-world channels. The deterministic model is developed into stochastic forms, and finally into averaged forms. During the development, the distortion of a signal in the channel is addressed and the antenna interfacing principles are also discussed.

5.2 Static terminal in a static two-source scenario

The two-path model is considered here in the context of a scenario, i.e. as a two-source model. While the model is for elucidating the behaviour of mobile channels, a physical situation could also exist, for example, with a direct path from the transmitter with a single ground bounce, or else as a dominant and a secondary wave from independent principal scatterers. Figure 5.2.1 depicts the two-source, or two-path, model.

5.2.1 Impulse response

The impulse response of the two-path model can be shifted in delay time and scaled to be written as

$$h^{(2)}(\tau) = \delta(\tau) + a_2 \mathrm{e}^{\mathrm{j}\alpha_2}\delta(\tau - \tau_2), \tag{5.2.1}$$

such that one of the waves always has unity amplitude, zero phase, and zero delay. The complex amplitude of the wave arriving from the secondary scatterer is $a_2\mathrm{e}^{\mathrm{j}\alpha_2}$ and τ_2 is the delay with respect to the wave from the first scatterer. In the general case, τ_2 cannot be related to the difference in the physical distances to the scatterers since previous propagation paths are not included (except through the value of τ_2) in the model. No information regarding the direction of arrival of the waves is available from the impulse response. Angle-of-arrival information requires some sampling in the spatial domain, either from additional, spaced antennas or from movement of the receiver as shown in Chapter 2. It is also discussed further below. The delay τ_2 is defined as always positive. The transfer function has symmetry about $a_2 = 1$ in the sense that when $a_2 > 1$, the impulse response can be re-scaled as

$$h^{(2)}(\tau) = \frac{1}{a_2}\,\mathrm{e}^{-\mathrm{j}\alpha_2}\delta(\tau - \tau_2) + \delta(\tau), \quad a_2 > 1. \tag{5.2.2}$$

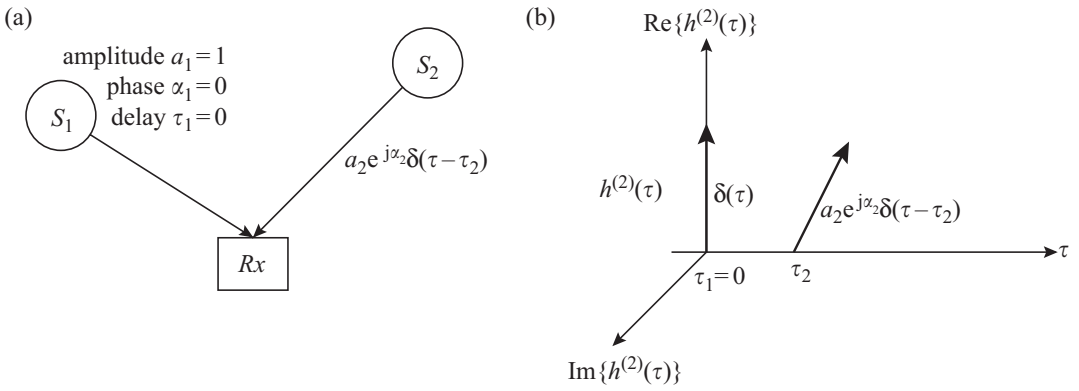


Figure 5.2.1 (a) A two-source scenario model in which two waves are incident at the receiver. The ratio of the wave amplitudes is a_2, and the time-of-arrival difference is τ_2 s where τ_2 is positive. (b) The impulse response of the mobile channel, where $\tau_2 > 0$.

This scaling maintains unit amplitude for the first ($\tau = 0$) term and the symmetry is convenient for checking some of the two-path results.

The *power delay profile* is simply

$$P(\tau) = \left\langle |h(\tau)|^2 \right\rangle = |h(\tau)|^2 \tag{5.2.3}$$

since no averaging is necessary in the static scenario, and so for the deterministic two-path model, this is

$$P^{(2)}(\tau) = \delta(\tau) + a_2^2 \delta(\tau - \tau_2). \tag{5.2.4}$$

It is emphasised that the power delay profile here is the same as the instantaneous power profile in the static channel. In real-world situations, this is not normally the case.

5.2.2 Delay time moments

The *mean excess delay*, or just the *mean delay time*, is the first moment with respect to the power delay profile $P(\tau)$, which for the static scenario can be written

$$\langle \tau \rangle = \frac{\int \tau P(\tau)\, \mathrm{d}\tau}{\int P(\tau)\, \mathrm{d}\tau} = \frac{\int \tau |h(\tau)|^2\, \mathrm{d}\tau}{\int |h(\tau)|^2\, \mathrm{d}\tau}. \tag{5.2.5}$$

The word 'excess' is included to differentiate $\langle \tau \rangle$ from the time-of-flight delay. For the two-path case the mean excess delay is simply the energy weighted delays of the impulses,

$$\langle \tau \rangle^{(2)} = \frac{a_2^2 \tau_2}{1 + a_2^2}. \tag{5.2.6}$$

In order to calculate the variance[1] the *mean square excess delay* (mean square time) is required. The formula again needs no averaging here, and is

$$\langle \tau^2 \rangle = \frac{\int \tau^2 P(\tau)\, \mathrm{d}\tau}{\int P(\tau)\, \mathrm{d}\tau} = \frac{\int \tau^2 |h(\tau)|^2\, \mathrm{d}\tau}{\int |h(\tau)|^2\, \mathrm{d}\tau} \tag{5.2.7}$$

which gives for the two-path case,

$$\left\langle \tau^2 \right\rangle^{(2)} = \frac{a_2^2 \tau_2^2}{1 + a_2^2}. \tag{5.2.8}$$

The standard deviation of the power delay profile is the *delay spread*,

$$s = \sigma_\tau = \sqrt{\langle \tau^2 \rangle - \langle \tau \rangle^2}, \tag{5.2.9}$$

[1] The terms variance and standard deviation, etc., are applied here despite the situation being deterministic.

which for the two-path case is

$$s^{(2)} = \frac{a_2\tau_2}{1+a_2^2}. \tag{5.2.10}$$

This can be referred to as the *instantaneous delay spread*, since the power delay profile has not been (and has not needed to be) averaged for the static scenario. In practice, $P(\tau)$ needs to be found by averaging, before finding the moments. The delay spread is a measure of the time dispersion of the channel, or how spread-out the impulse response is. The two-path mean excess delay time $\langle\tau\rangle^{(2)}$ and delay spread $s^{(2)}$ both give a proportional measure of the spread, namely $s^{(2)} = \langle\tau\rangle^{(2)}/a_2 = \langle\tau^2\rangle/a_2\tau_2$, a situation which does not hold in the general multipath case. Cox (1972) provided some of the first measured delay spread characteristics of the mobile channel.

5.2.3 Transfer function

The transfer function is the usual Fourier transform of the impulse response

$$H(\omega) = \int_{-\infty}^{\infty} h(\tau)\,\mathrm{e}^{-\mathrm{j}\omega\tau}\,\mathrm{d}\tau \quad \Leftrightarrow \quad h(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)\,\mathrm{e}^{\mathrm{j}\omega\tau}\,\mathrm{d}\omega. \tag{5.2.11}$$

For the two-path model (cf. Lee, 1983; Vaughan, 1986; Andersen *et al.*, 1988), this is

$$H^{(2)}(\omega) = 1 + a_2\mathrm{e}^{\mathrm{j}\alpha_2}\,\mathrm{e}^{-\mathrm{j}\omega\tau_2} \tag{5.2.12}$$

and so the power transfer function is

$$|H^{(2)}(\omega)|^2 = 1 + a_2^2 + 2a_2\cos(\omega\tau_2 - \alpha_2). \tag{5.2.13}$$

This has (periodic) maxima and minima, which is the two-path model version of the frequency-selective fading referred to in Chapter 1. The form of the magnitude again has symmetry about $a_2 = 1$ in the sense that a_2 can be replaced by $1/a_2$, as with the impulse response. However, the form of the phase of the transfer function changes when a_2 changes from being less than one to greater than one. Unless otherwise stated, it is assumed that $a_2 \leq 1$, i.e. the minimum-phase case.

5.2.3.1 Magnitude

The magnitude of the transfer function, $|H(\omega)|$, has minima at $\omega\tau_2 = (2n-1)\pi + \alpha_2$ (where $n = 0, \pm1, \ldots$), i.e. spaced by $\omega = 2\pi/\tau_2$, independent of a_2. The mean value of the power transfer function is $1 + a_2^2$, the minima are $(1 - a_2)^2$, the maxima are $(1 + a_2)^2$, and the standard deviation is $\sqrt{2}a_2$. Figure 5.2.2 gives the fade depth in dB relative to the mean and the maxima, in terms of the parameter a_2.

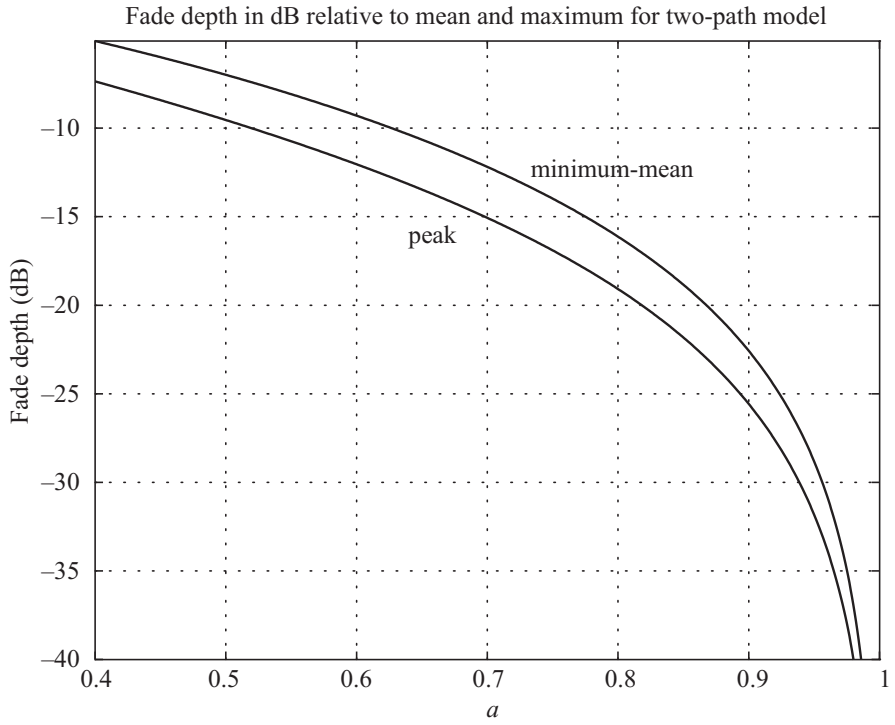


Figure 5.2.2 The dynamic range of the envelope of the two-path model, depicted as the fade depth relative to the peak power and the mean power

5.2.3.2 Phase

Useful expressions are

$$\begin{aligned}
\mathrm{Re}\,\{H^{(2)}(\omega)\} &= 1 + a_2\cos(\omega\tau_2 - \alpha_2)\\
\mathrm{Im}\,\{H^{(2)}(\omega)\} &= -a_2\sin(\omega\tau_2 - \alpha_2)\\
\mathrm{Re}\,\{H^{(2)'}(\omega)\} &= 1 - a_2\tau_2\sin(\omega\tau_2 - \alpha_2)\\
\mathrm{Im}\,\{H^{(2)'}(\omega)\} &= -a_2\tau_2\cos(\omega\tau_2 - \alpha_2)
\end{aligned} \tag{5.2.14}$$

where the prime denotes differentiation with respect to the argument.

The phase of the transfer function is

$$\phi(\omega) = \tan^{-1}\left(\frac{\mathrm{Im}\,\{H(\omega)\}}{\mathrm{Re}\,\{H(\omega)\}}\right), \tag{5.2.15}$$

which for the two-path model is (cf. Andersen *et al.*, 1988)

$$\phi^{(2)}(\omega) = \tan^{-1}\left(\frac{-a_2\sin(\omega\tau_2 - \alpha_2)}{1 + a_2\cos(\omega\tau_2 - \alpha_2)}\right). \tag{5.2.16}$$

The magnitude and phase of $H^{(2)}(\omega)$ are given in Figure 5.2.3. Setting the arguments to $(\omega\tau_2 - \alpha)/\pi$, the transfer function minima are at $\pm 1, \pm 3, \ldots$, etc. The phase of

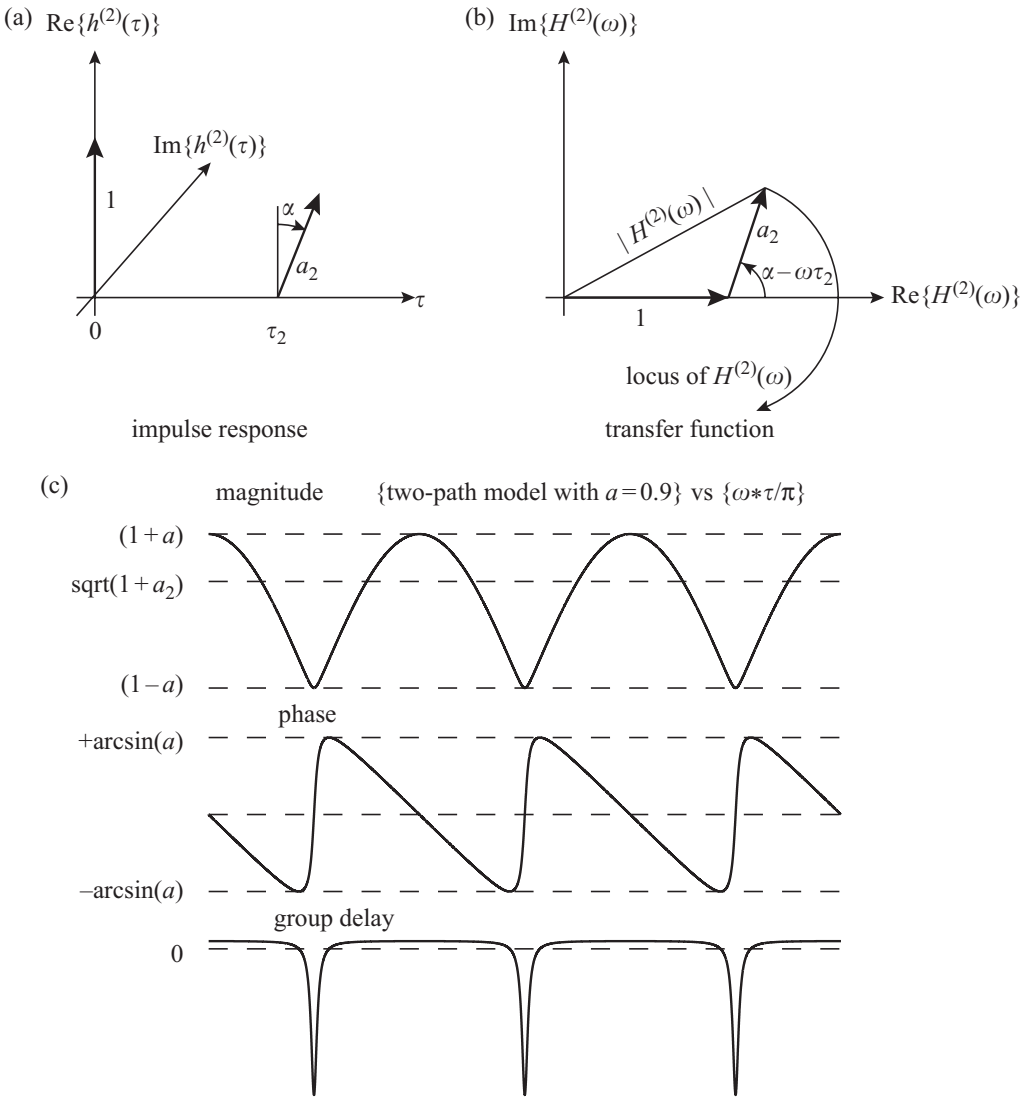


Figure 5.2.3 *(a) The impulse response and (b) phasor diagram part of the periodic transfer function (c), with group delay, for a static receiver in a two-source scenario. The figures are for* $a_2 < 1$

the two-path model oscillates between $\pm \sin^{-1} a_2$ with rapid changes occurring when the transfer function magnitude is at a minimum. For $a_2 \leq 1$, the minimum phase condition, the phase remains within the $\pm\pi/2$ range about 0.

5.2.4 Group delay

The group delay τ_g is defined as

$$\tau_g(\omega) = -\phi'(\omega) \tag{5.2.17}$$

and so it is strictly a narrowband parameter, although further below a mean group delay is taken as an integral over a narrow bandwidth. For a signal with linear phase

and constant magnitude across its bandwidth, the group delay can be interpreted as the time delay imposed on the envelope of the signal as a result of it passing through a channel. This can be seen from a simple two-term Taylor expansion of the phase of the transfer function about ω_0, which for the static scenario, is

$$\begin{aligned} H(\omega) = |H(\omega)|\,\mathrm{e}^{\mathrm{j}\phi(\omega)} &\approx |H(\omega)|\,\mathrm{e}^{\mathrm{j}(\phi(\omega_0)-(\omega-\omega_0)\tau_\mathrm{g}(\omega_0))} \\ &= |H(\omega)|\,\mathrm{e}^{\mathrm{j}(\phi(\omega_0)+\omega_0\tau_\mathrm{g}(\omega_0))}\,\mathrm{e}^{-\mathrm{j}\omega\tau_\mathrm{g}(\omega_0)}. \end{aligned} \tag{5.2.18}$$

Fourier transformation makes explicit the time delay τ_g imposed on the impulse response (time shift property of the Fourier transform). In general, the presence of magnitude slope will also contribute to the delay; this is elaborated further below when the mean of the group delay over a finite bandwidth is discussed.

From the definitions in equations (5.2.15) and (5.2.17) the group delay is (Andersen *et al.*, 1988)

$$\tau_\mathrm{g}(\omega) = \frac{\mathrm{Im}\{H(\omega)\}\,\mathrm{Re}\{H'(\omega)\} - \mathrm{Re}\{H(\omega)\}\,\mathrm{Im}\{H'(\omega)\}}{|H(\omega)|^2}. \tag{5.2.19}$$

For the two-path model of equation (5.2.12), this is (for the minimum phase case)

$$\tau_\mathrm{g}^{(2)}(\omega) = \frac{a_2\tau_2(a_2 + \cos(\omega\tau_2 - \alpha_2))}{1 + a_2^2 + 2a_2\cos(\omega\tau_2 - \alpha_2)}. \tag{5.2.20}$$

The group delay is also periodic in frequency with the minima being narrower than the maxima. Note that the minima of the group delay, occurring where $(\omega\tau - \alpha_2) = \pi$,

$$\tau_{\mathrm{g\ min}}^{(2)} = \frac{a_2\tau_2}{a_2 - 1}, \tag{5.2.21}$$

are negative in Figure 5.2.3(c) since $a_2 < 1$. The maxima are

$$\tau_{\mathrm{g\ max}}^{(2)} = \frac{a_2\tau_2}{a_2 + 1}. \tag{5.2.22}$$

For this case it can be seen from Figure 5.2.3(c) that the maxima of the magnitude of the group delay occur (by definition) where the phase is changing the quickest, i.e. at the minima of $|H^{(2)}(\omega)|^2$. The non-minimum phase case is discussed further below.

The rate of change of group delay with frequency, i.e. $|\phi''(\omega)|$, over a bandwidth, gives a measure of the phase dispersion, which is discussed in detail in Section 5.3. It is noteworthy for communications usage that where the channel magnitude is relatively large and linear, the phase also tends to be linear and the local (in frequency) phase and amplitude dispersion are relatively low. Conversely, at low energy levels, the local dispersion is high. High dispersion and low signal-to-noise ratio are directly related for the two-source model. The same trend is found in real-world, many-path, channels, where, however, high dispersion levels can sometimes accompany high signal-to-noise ratios.

5.2.5 Mean group delay

A communications channel must have finite bandwidth. The group delay can be averaged over frequency to give a mean group delay (Andersen *et al.*, 1988)

$$\langle \tau_{\mathrm{g}} \rangle_{\omega} = \frac{\int \tau_{\mathrm{g}}(\omega) |H(\omega)|^2 \, \mathrm{d}\omega}{\int |H(\omega)|^2 \, \mathrm{d}\omega} = \langle \tau \rangle_{\tau} \tag{5.2.23}$$

and this is mathematically the same as the mean delay, denoted $\langle \tau_g \rangle_{\omega} = \langle \tau \rangle_{\tau}$ (Appendix E). However, interpreting the mean group delay, and indeed the group delay itself, as an envelope delay in fading channels requires care in some situations, as shown below.

5.2.6 Non-minimum phase case

Figure 5.2.4 depicts the transfer function for the two-path case when $a_2 > 1$. The phase becomes monotonic with corresponding positive-only behaviour for the group delay, an inverted form of that for $a_2 < 1$. The fading channel is thus different for the minimum-phase and non-minimum phase cases, i.e. these are two types of fade which are possible in the two-path channel.

5.2.7 Effect of finite pulse width on delay spread

In the preceding treatment, the impulse response was described by delta functions, strictly mathematical expressions that correspond to an infinite bandwidth. To get a feel for the effect of a finite bandwidth, the delta functions are replaced by rectangular pulses, whose separation τ_2 is maintained, but whose width, i.e. duration, is varied. For this section, it is convenient to denote $a = a_2 \, \mathrm{e}^{\mathrm{j}\alpha_2}$. The rectangular form represents an effective duration, T, so the influence of the bandwidth size can be investigated. The bandwidth is approximately the inverse of the pulse duration,

$$B \approx \frac{1}{T}. \tag{5.2.24}$$

The formulation brings in the effect of the relative phase of the two pulses, which is assumed constant over the pulse durations, and illustrates below the difference between the instantaneous group delay and the averaged (over the relative phase) group delay.

5.2.7.1 Impulse response

The impulse response is

$$h(\tau) = r_T \left(\tau - \frac{T}{2} \right) + a \, r_T \left(\tau - \left(\tau_2 + \frac{T}{2} \right) \right), \tag{5.2.25}$$

where T is the duration of the pulses and the pulse separation is τ_2, a is the two-path parameter, which in this section is complex, and $a_2 = |a| \leq 1$; and $r_T(\tau - \tau_0)$ is the

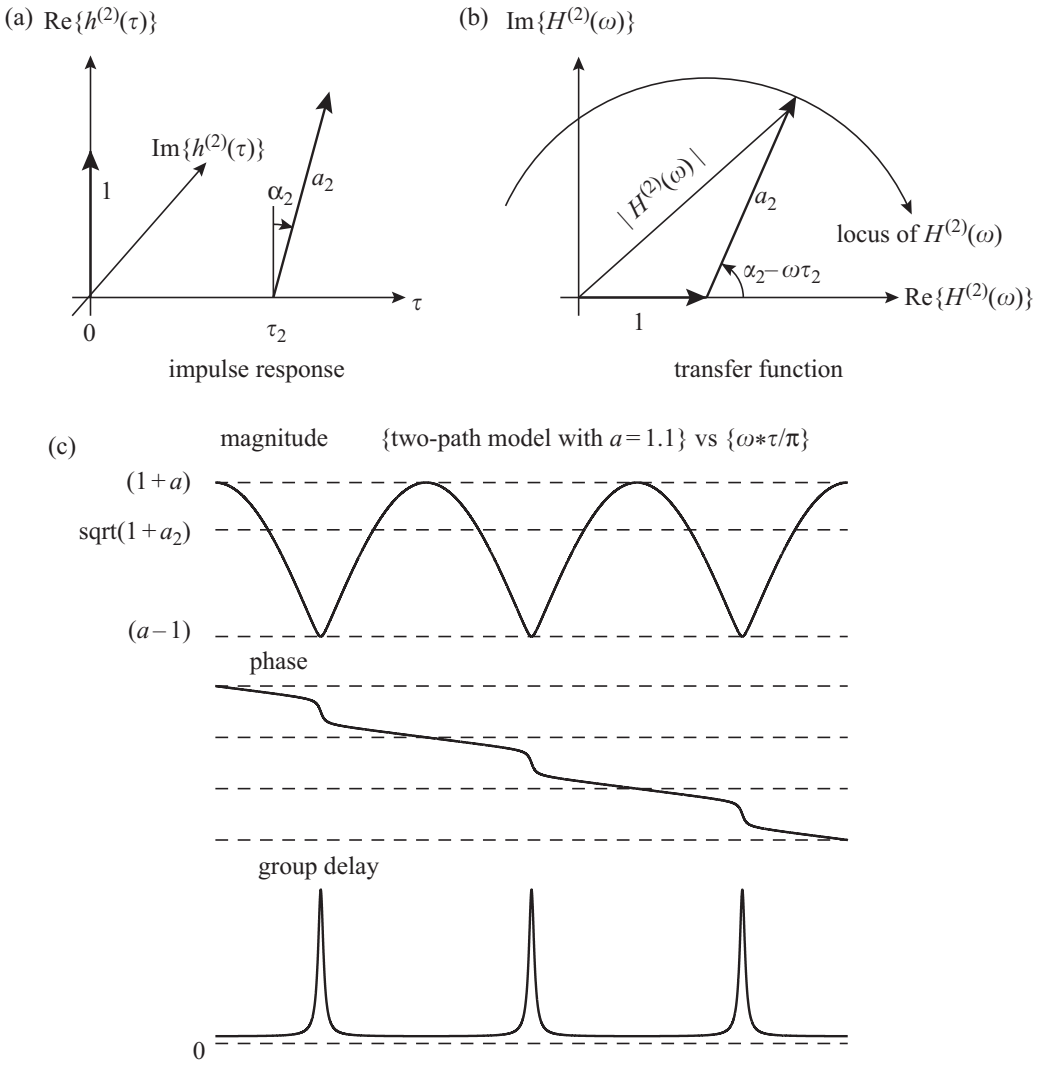


Figure 5.2.4 *(a) The impulse response, (b) phasor diagram, and (c) the transfer function for the two-source model under non-minimum phase conditions, i.e. when* $a_2 > 1$

real rectangle function of width T and centred at $\tau = \tau_0$. The impulse response $h(\tau)$ is arranged so that it occupies positive time only. Once the bandwidth has decreased to the point where the pulses overlap, which is when $T > \tau_2$, the impulse response can be expressed as three rectangular regions,

$$h(\tau) = r_{\tau_2}\left(\tau - \frac{\tau_2}{2}\right) + (1+a)r_{T-\tau_2}\left(\tau - \frac{T+\tau_2}{2}\right) + a\, r_{\tau_2}\left(\tau - \frac{2T+\tau_2}{2}\right), \quad \tau_2 < T. \tag{5.2.26}$$

The resulting time functions are depicted in Figure 5.2.5. The power profile for the static scenario follows as

$$\begin{aligned} P(\tau) &= |h(\tau)|^2 \\ &= r_T\left(\tau - \frac{T}{2}\right) + |a|^2 r_T\left(\tau - \left(\tau_2 + \frac{T}{2}\right)\right), \quad \tau_2 \geq T, \end{aligned} \tag{5.2.27}$$

$$\begin{aligned} P(\tau) = r_{\tau_2}\left(\tau - \frac{\tau_2}{2}\right) &+ |1+a|^2 r_{T-\tau_2}\left(\tau - \frac{T+\tau_2}{2}\right) \\ &+ |a|^2 r_{\tau_2}\left(\tau - \frac{\tau_2 + 2T}{2}\right), \quad \tau_2 < T, \end{aligned} \tag{5.2.28}$$

which is also shown in Figure 5.2.5 for the overlapping pulses.

The central pulse of the power delay profile, which is $|1+a|^2$, can be larger or smaller than the two-component pulses, depending on the phase relation between the components. For co-phased pulses, the central region is large, and will act to reduce the delay spread of the channel. For a quadrature phase relation, the effect is less pronounced, and for pulses in phase opposition the central region is small, depending on $|a|$, and the delay spread is increased.

The energy of the impulse response is

$$\begin{aligned} \int |h(\tau)|^2 \, \mathrm{d}\tau &= \int_0^T \mathrm{d}\tau + |a|^2 \int_{\tau_2}^{\tau_2+T} \mathrm{d}\tau, \quad \tau_2 \geq T, \\ &= T\left(1 + |a|^2\right), \end{aligned} \tag{5.2.29}$$

for non-overlapping pulses, and

$$\begin{aligned} \int |h(\tau)|^2 \, \mathrm{d}\tau &= \int_0^{\tau_2} \mathrm{d}\tau + |1+a|^2 \int_{\tau_2}^{T} \mathrm{d}\tau + |a|^2 \int_{T}^{T+\tau_2} \mathrm{d}\tau, \quad \tau_2 < T, \\ &= \tau_2 + |1+a|^2 (T - \tau_2) + |a|^2 \tau_2 \end{aligned} \tag{5.2.30}$$

for overlapping pulses. From these equations and the pdf, the first and second moments are found and these provide the instantaneous delay spread, which is given in Figure 5.2.6. The delay spread is here normalised by the pulse separation τ_2, and the abscissa is the normalised pulse duration, i.e. T/τ_2. This is to be read as the spacing τ_2 fixed, with the pulse duration increasing (i.e. bandwidth decreasing) along the abscissa.

5.2.7.2 Discussion

For $T/\tau_2 < 1$, the pulses are separate. When $T/\tau_2 > 1$, the pulses overlap, and the interference between the pulses, which depends on their phase difference, comes into effect. When the bandwidth approaches infinity, the pulses become infinitesimally thin, $T/\tau_2 \to 0$, and the delay spread is given by the formulas for the delta function formulation above. $|a| = 0.4$ is given for co-phased and quadrature relations, and

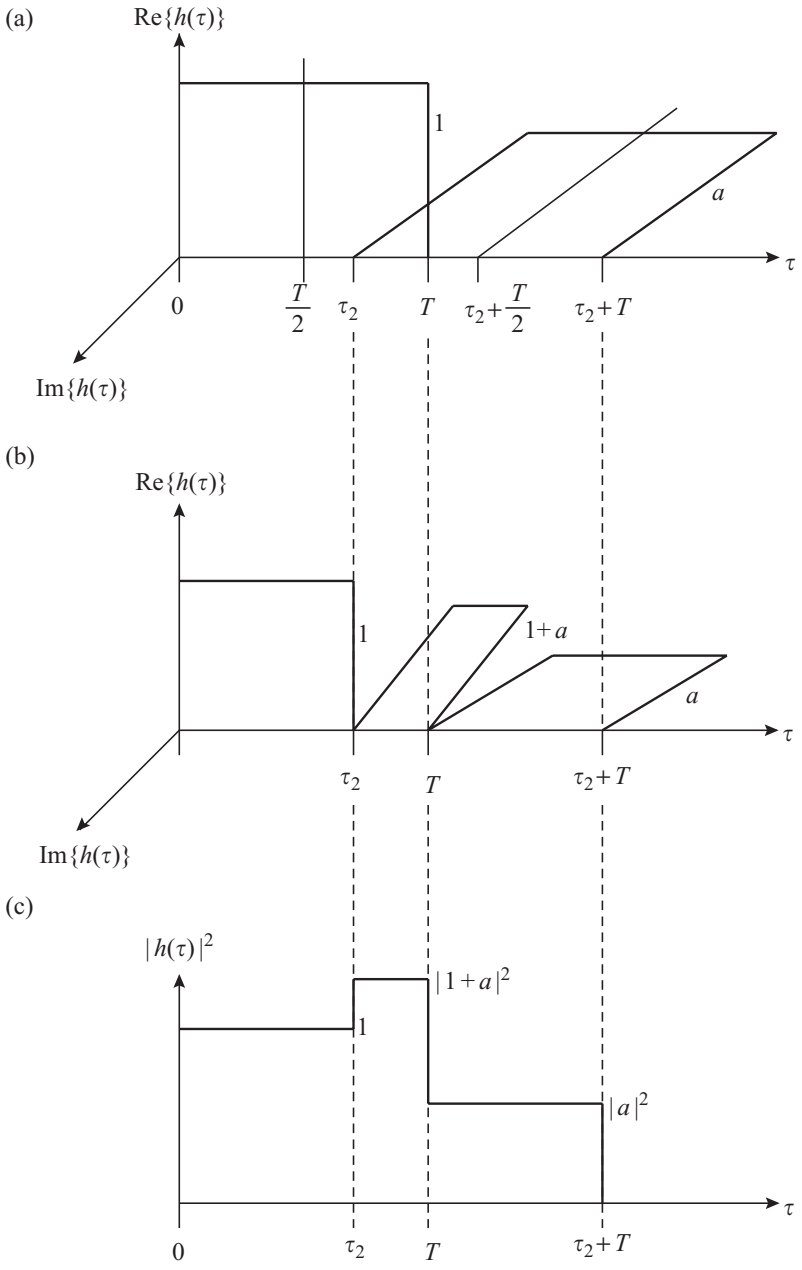


Figure 5.2.5 *Impulse response quantities for a finite bandwidth channel. The pulses have width T, are separated by τ_2, and are arranged so that the response starts at $\tau = 0$: (a) is the impulse response showing the two component pulses, (b) gives an interpretation for when the pulses overlap, and (c) is the power profile for the overlapping pulses*

$|a| = 1$ for in-phase, quadrature, and phase-opposition. Once the pulses are overlapping, the change in delay spread is seen to vary widely, with this variation increasing for increasing $|a|$. For the co-phased case, $a = 1$, the delay spread is seen to momentarily decrease before it (ultimately) returns to a linear trajectory. For the $a = \pm\mathrm{j}$ cases, the curve slope is continuous; and for the phase opposition case, the delay

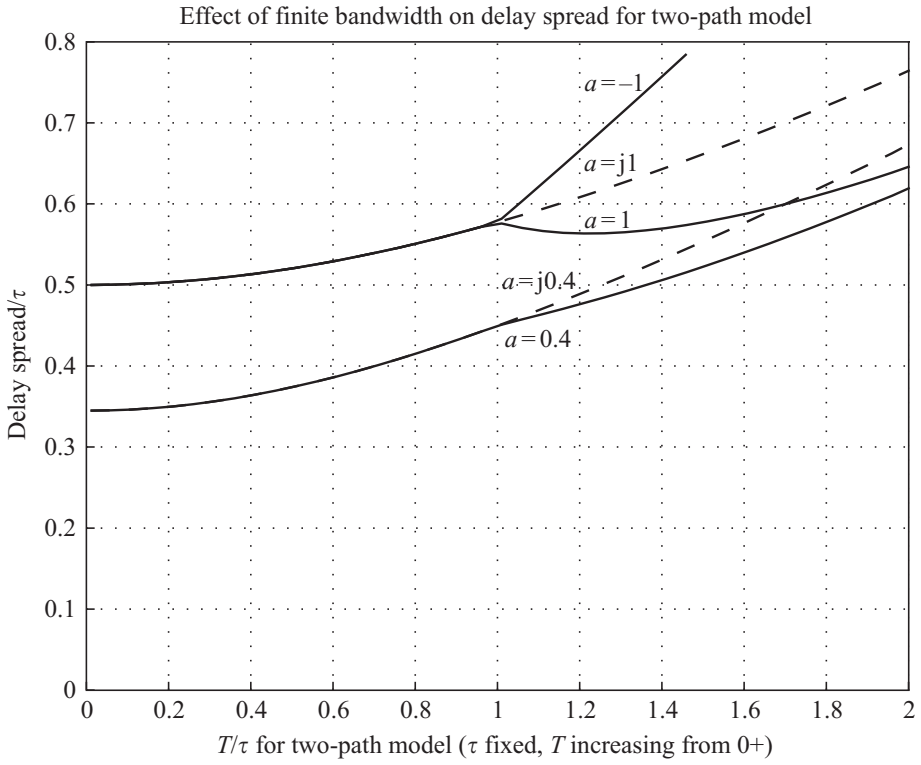


Figure 5.2.6 *The increasing delay spread for decreasing bandwidth in the two-path model, with τ_2 and a complex with $|a| \leq 1$. The delay spread is here normalised by the pulse separation τ_2, plotted against the normalised pulse width, T/τ_2. The bandwidth is in the order of $1/T$ so the abscissa is approximated by $1/(B\tau_2)$. As T/τ_2 increases, the pulses go from separated ($T/\tau_2 < 1$) to overlapping, coinciding as $T/\tau_2 \to \infty$. The instantaneous delay spread in the overlapping region depends on the phase difference of the pulses. The averaged (over phase difference) delay spread follows the dotted lines*

spread increases linearly and more quickly than the other cases. Here, the central rectangle of the power delay profile is zero, leaving a 'hole' in the impulse response, with the corresponding increase in the delay spread.

If the power delay profile is calculated by averaging over all possible phase differences between the two paths, then

$$\begin{aligned} P(\tau) &= \langle |h(\tau)|^2 \rangle_{\angle a} \\ &= r_{\tau_2}\left(\tau - \frac{\tau_2}{2}\right) + \left(1 + |a|^2\right) r_{T-\tau_2}\left(\tau - \frac{T+\tau_2}{2}\right) \\ &\quad + |a|^2 r_{\tau_2}\left(\tau - \frac{\tau + 2T}{2}\right), \quad \tau_2 < T, \end{aligned} \tag{5.2.31}$$

which is the same result as the instantaneous delay spread for a being purely imaginary. The (averaged) delay spread thus follows the continuous trajectory (dashed)

in Figure 5.2.6. The delay spread and instantaneous delay spread are the same for non-overlapping pulses.

5.2.8 Probability functions

The statistical approach is required when there are too many paths to determine, which is usually the case in mobile communications. Here, the pdf of the power in the static two-path model offers a feel for the approach and its interpretation of results. In this section, the complex delay for the second pulse is again denoted $a_2\,\mathrm{e}^{\mathrm{j}\alpha_2}$.

The static two-path transfer function, having fixed source amplitudes, takes on all of its possible values as the phase difference between the sources is varied. The phase difference can be varied directly by changing the phase of one of the sources, or else by changing the mobile position or the frequency. In the context of the static scenario, the frequency is varied. For equi-likely frequencies in the two-path model, the pdf is simply

$$p_\omega(\omega) = \frac{\tau_2}{\pi}, \quad n\frac{\pi}{\tau_2} \le \omega \le (2n+1)\frac{\pi}{\tau_2}, \quad n = 0, \pm1, \pm2, \pm3, \ldots, \tag{5.2.32}$$

since the frequency selectivity of the transfer function is periodic. For the case $a_2 < 1$ and equi-likely frequencies, the pdf for the power is, from transformation of p_ω,

$$\begin{aligned} p_{|H^{(2)}(\omega)|^2}\left(|H|^2\right) &= p_\omega(\omega)\left|\frac{\partial\omega}{\partial|H^{(2)}|^2}\right| \\ &= \frac{1}{\pi\sqrt{(2a_2)^2 - \left(|H|^2 - (1+a_2^2)\right)^2}} \\ &= \frac{1}{\pi\sqrt{\sigma_2^2 - \left(|H|^2 - \mu_2\right)^2}}, \end{aligned} \tag{5.2.33}$$

where μ_2 and σ_2^2 denote the mean and variance of the power in the two-path channel. The cumulative density function (cdf), or cumulative probability function, follows as

$$\begin{aligned} P_{|H|^2}\left(|H^{(2)}(\omega)|^2 \le |H_0|^2\right) &= 1 - \frac{1}{\pi}\cos^{-1}\left(\frac{|H_0|^2 - (1+a_2^2)}{2a_2}\right) \\ &= 1 - \frac{1}{\pi}\cos^{-1}\left(\frac{|H_0|^2 - \mu_2}{\sigma_2}\right). \end{aligned} \tag{5.2.34}$$

This probability approach is an alternative to the deterministic equation (5.2.13) for characterising the two-path channel, and is needed when a deterministic form such as that of the two-path model is not applicable. The pdf and cdf are given in Figure 5.2.7. The pdf is centred at the mean, $1 + a_2^2$, and is confined to its limits, i.e. between $(1 - a_2)^2$ and $(1 + a_2)^2$. At these limits, the pdf $p_{|H(\omega)|^2}$ goes to infinity. This does not cause interpretation problems, however, since the probability of the power being

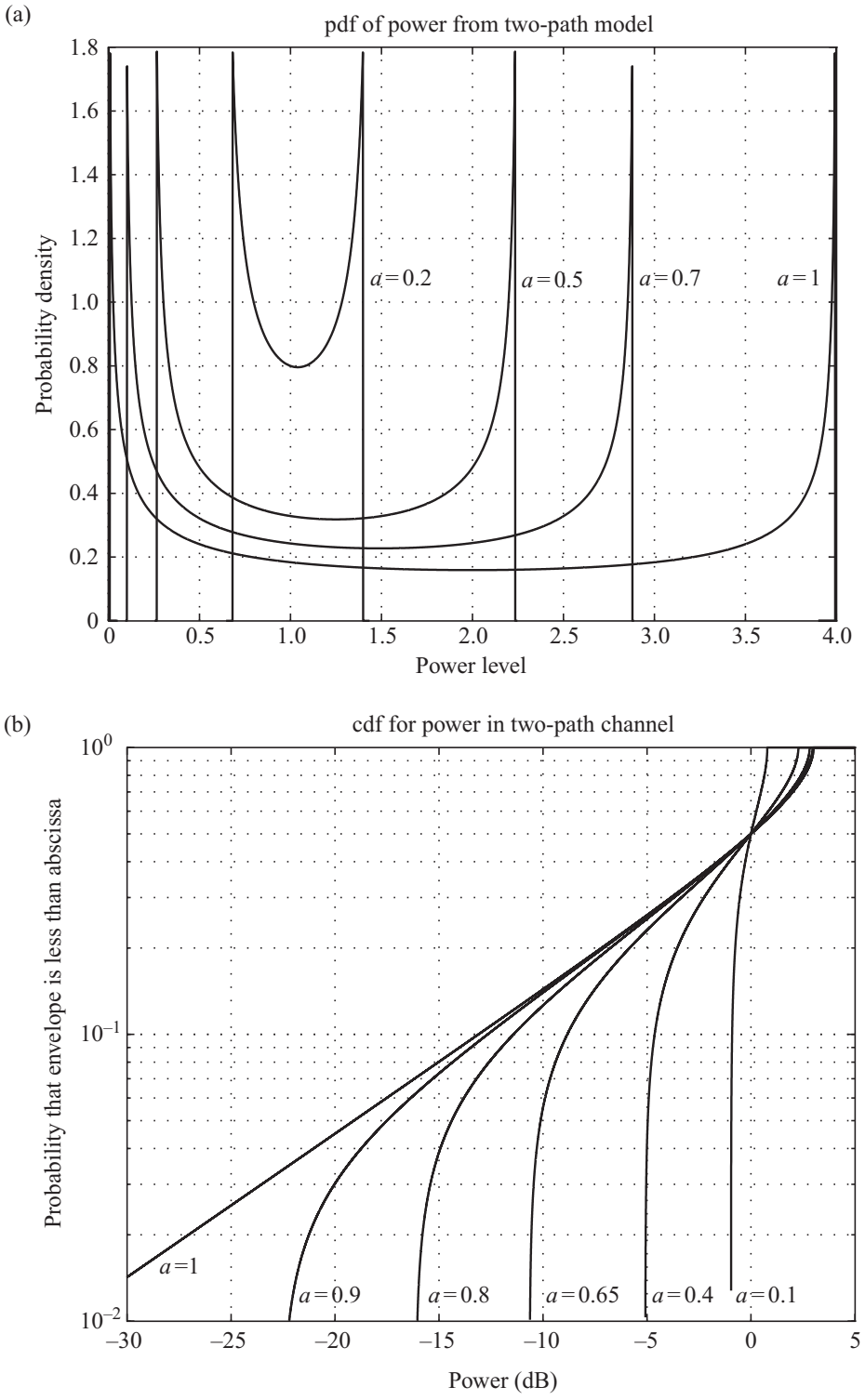


Figure 5.2.7 *(a) The pdf and (b) the cdf for the power of the two-path channel, for various values of the two-path parameter a_2. The channel is periodic in frequency, and so these statistics are independent of the two-path delay time parameter τ_2. A uniform distribution in frequency is assumed. The curves all have their mean values, $1 + a_2^2$ (which is also the median value), at 0 dB*

at these limits is infinitesimal. For example, for $a_2 = 1$, the fades go to zero in the transfer function. In the cdf, the interpretation is that there is an infinitesimally small probability of the power being zero, i.e.

$$P\left(|H(\omega)|^2 \leq |H_0|^2\right) \to 0 \quad \text{as} \quad |H_0|^2 \to 0, \quad a_2 = 1. \tag{5.2.35}$$

A similar situation holds for the power approaching its maximum value $(1 + a_2)^2$,

$$P\left(|H(\omega)|^2 \leq |H_0|^2\right) \to 1 \quad \text{as} \quad |H_0|^2 \to (1 + a_2)^2 . \tag{5.2.36}$$

Again in the $a_2 = 1$ example, the cdf diagram shows that for 10 per cent of the frequencies the power transfer function is more than 13 dB below its mean value. The cdf curves are arranged so that the mean power, $1 + a_2^2$, always corresponds to 0 dB. A flat channel ($a_2 = 0$) would be a line at $|H_0|^2 = 0$ dB.

Finally, with a frequency pdf, the moments such as those for the delay can be formally stated, for example the mean group delay can be written

$$\langle \tau_g \rangle = E\left\{\tau_g(\omega)\right\} = \int \tau_g(\omega)\, p_\omega(\omega)\, \mathrm{d}\omega. \tag{5.2.37}$$

5.2.8.1 Summary

The phase difference between the source contributions is the generic random variable for the statistical approach to the short-term variation of the envelope of the received signal. In the static scenario here, the phase difference is implemented by varying the frequency. However, the mobile position can be varied instead, as developed in Chapter 2, and which will be applied below in Section 5.4.

5.2.9 Effect of finite bandwidth on the fading channel

Many systems are not as narrowband as suggested by the above cw (continuous wave), i.e. infinitesimal bandwidth, approach. 'Wideband' systems are usually taken to mean that the bandwidth covers several frequency selective fades, and there are systems, such as the GSM (single user bandwidth $\sim$200 kHz) which fall between the truly narrowband and wideband cases. Here the effect on the fading of the bandwidth is quantified for the two-path channel.

The power within the band B rad s^{-1} is denoted

$$\begin{aligned} \gamma_B &= \frac{1}{B}\int_{\omega}^{\omega+B} \left|H^{(2)}(\omega)\right|^2 \mathrm{d}\omega \\ &= 1 + a_2^2 + 2a_2\chi\cos\left(\omega\tau_2 - \alpha_2\right), \end{aligned} \tag{5.2.38}$$

where

$$\chi = \frac{\sin\left(B\tau_2/2\right)}{B\tau_2/2}. \tag{5.2.39}$$

The minimum and maximum values of the power are now

$$\gamma_{B\ \max} = 1 + a_2^2 + 2a_2\chi, \qquad \gamma_{B\ \min} = 1 + a_2^2 - 2a_2\chi, \tag{5.2.40}$$

which shows that the dynamic range, $\gamma_{B\ \max}/\gamma_{B\ \min}$, is reduced from the value of the cw ($B\tau_2 = 0$) case. This is the effect of the finite bandwidth on the magnitude (Andersen *et al.*, 1988). The minimum power is the more important quantity, and is given by

$$\begin{aligned} \gamma_{B\ \min} &\approx 2(1-\chi), \quad a_2 \approx 1, \\ &\approx \frac{1}{3}\left(\frac{B\tau_2}{2}\right)^2, \quad B\tau_2 \ll 1, \end{aligned} \tag{5.2.41}$$

i.e. proportional to the bandwidth squared for the narrowband case.

5.2.9.1 Statistical quantities

The pdf is found in the same way as above, and is

$$p_\gamma(\gamma) = \frac{1}{\pi\sqrt{(2a_2\chi)^2 - \left(\gamma_B - \left(1 + a_2^2\right)\right)^2}} \tag{5.2.42}$$

and the cdf follows as

$$P_\gamma\left(\left|H^2(\omega)\right|^2 \le \gamma_B\right) = 1 - \frac{1}{\pi}\cos^{-1}\left(\frac{\gamma - \left(1 + a_2^2\right)}{2a_2\chi}\right). \tag{5.2.43}$$

In Figure 5.2.8, the cdf is given for the case $a_2 = 0.95$ and a range of bandwidths. Again, the curves are arranged so that the mean is always at 0 dB.

The mean power (averaged over values of the relative phase of the two paths) is the same as for the cw case,

$$\begin{aligned} \langle\gamma_B\rangle &= 1 + a_2^2 + 2a_2\chi\langle\cos\alpha_2\rangle \\ &= 1 + a_2^2. \end{aligned} \tag{5.2.44}$$

The second moment of the power is

$$\begin{aligned} \langle\gamma_B^2\rangle &= \left(1 + a_2^2\right)^2 + (2a_2\chi)^2\langle\cos^2\alpha_2\rangle \\ &= 1 + a_2^2 + 2(a_2\chi)^2, \end{aligned} \tag{5.2.45}$$

and the variance of the power follows as

$$\sigma_\gamma^2 = 2(a_2\chi)^2. \tag{5.2.46}$$

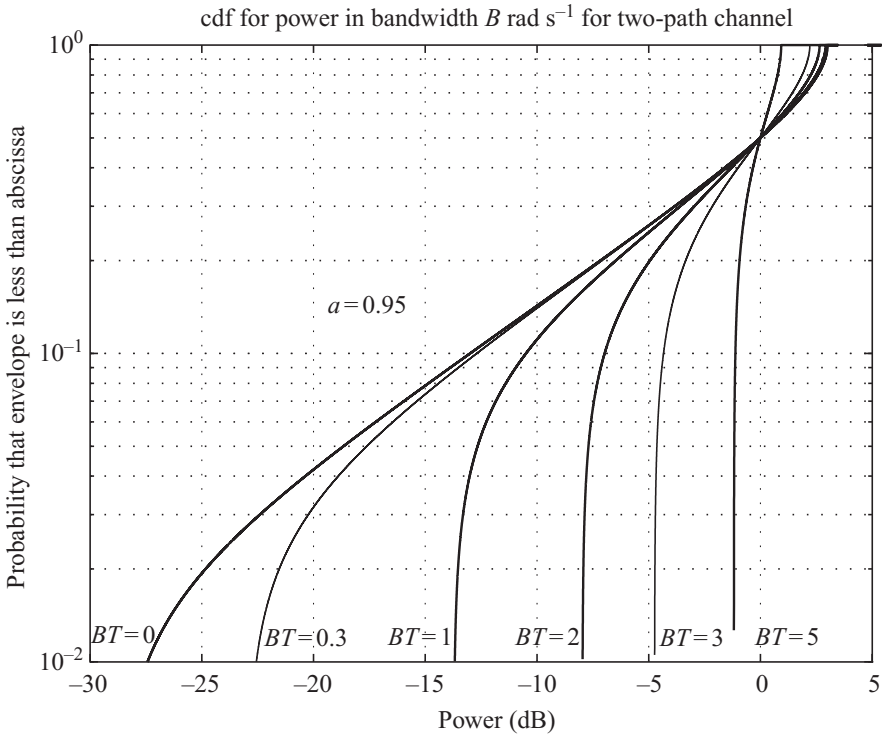


Figure 5.2.8 The cdf for the two-path model with a finite bandwidth B. The curves are located so that the mean (which is the same as the median) for each value of $B\tau_2$ is at 0 dB

5.2.9.2 An effective value of a_2

Note that the cdf can again be written directly in terms of the mean and standard variation. By equating the variances for the cw and finite bandwidth cases, i.e. $\sigma_\gamma^2\,(B\tau_2 = 0,\ a_\mathrm{e}) = \sigma_\gamma^2\,(B\tau_2,\ a_2)$, an effective value of a_2 is found as $a_\mathrm{e} = |a_2\chi|$, which can be used in an 'equivariance' narrowband two-source model. However, for matching up the probability distributions, equating the dynamic ranges for the two cases gives a different a_e, namely,

$$a_\mathrm{e} = \frac{1 + a_2^2}{2a_2\chi} - \sqrt{\left(\frac{1 + a_2^2}{2a_2\chi}\right)^2 - 1}, \quad a_2,\ \chi \neq 0. \tag{5.2.47}$$

This value of a_e can be used in the simpler cw model specified by $(a_\mathrm{e},\ \tau_2)$ instead of the finite bandwidth channel specified by $(a_2,\ \tau_2,\ B)$, giving similar fading distributions. As an example, $a_2 = 0.95$ and $B\tau_2 = 2$ corresponds to $a_\mathrm{e} = 0.55$ for the cw case.

5.2.9.3 Summary

The effect of the finite bandwidth is to reduce the dynamic range of the envelope variations relative to the cw case. The importance of this is that the fades are not as deep. This is also evident in the RSSI signal example of Figure 1.4.8. In the two-path model, the finite bandwidth effect is the same as decreasing the a_2 value. The effect

on the phase is similar, with the rate of phase variation reduced, and this topic is covered in the group delay discussion below.

5.2.10 Interpretation of negative group delay

From the expressions for the minima of the group delay for the minimum phase case, equation (5.2.21), it is of interest that this can be a negative quantity, a phenomenon seen in the many-path case as well (Vaughan, 1986). It is a surprising result in the sense that it may be interpreted as an advance in time-of-arrival (except for the 'time-of-flight' term, which is removed) for the envelope of a narrowband signal passing through the channel. Deriving the conditions for maximum and negative group delay from the two-source model, with an example, is therefore helpful for understanding the channel behaviour. The topic is taken even further in Section 5.3. This section draws from Poletti and Vaughan (1992).

5.2.10.1 Example

For the two-path model, the proportion of frequencies over which the group delay is negative is $\cos^{-1}(a_2/\pi\tau_2)$. Mathematically, the maximum proportion is 0.5, when $a_2 = 0$, but the group delay is zero in this limit. An instructive exercise is to find the minimum (most negative) mean group delay that can be imposed on a signal of fixed bandwidth. This is synthesising the two-path model for negative group delay. Denote the signal bandwidth B rad s^{-1}, which is centred in a region of negative group delay, as shown in Figure 5.2.9.

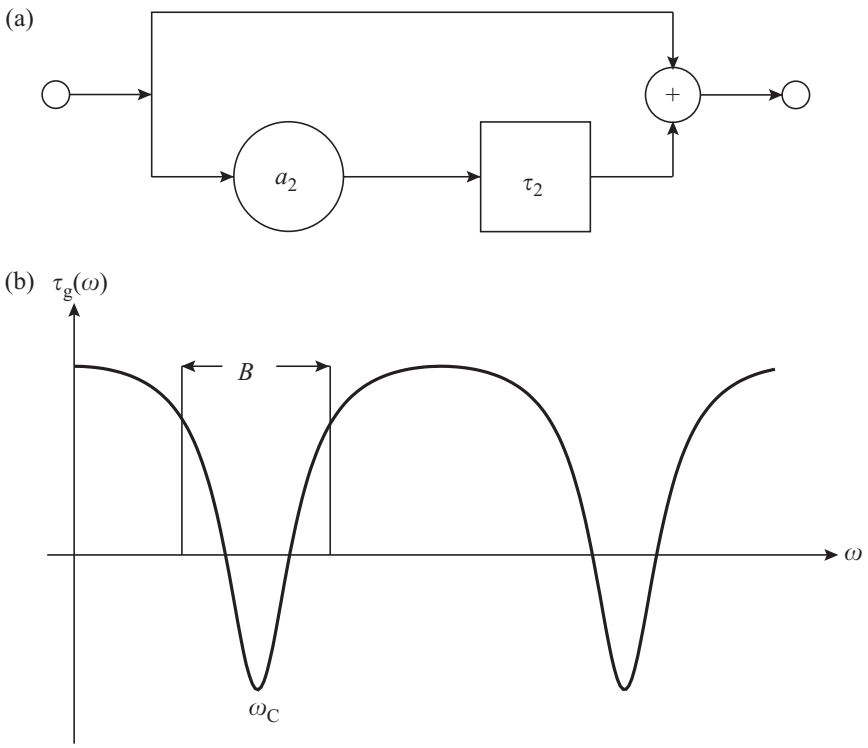


Figure 5.2.9 The two-path model with a bandwidth B centred on a group delay minimum

The input signal for transmission through the channel is a modulated sinc pulse, $x(t) = \mathrm{sinc}(Bt)\cos(\omega_C t)$, where ω_C is the centre of a negative group delay region. This ideal signal has a flat spectrum and zero phase change over the bandwidth. The group delay of the channel is then exactly the delay imposed on the narrowband pulse by the channel. The mean group delay for this negative group delay region is (cf. equation 5.2.23)

$$\left\langle \tau_g^{(2)} \right\rangle_B = \frac{\int_{\pi/\tau - B/2}^{\pi/\tau + B/2} \tau_g^{(2)}(\omega) |H^{(2)}(\omega)|^2 \, \mathrm{d}\omega}{\int_{\pi/\tau - B/2}^{\pi/\tau + B/2} |H^{(2)}(\omega)|^2 \, \mathrm{d}\omega}, \tag{5.2.48}$$

where the integral can be taken over any of the negative group delay regions, since the transfer function is periodic. Normalising by the delay difference, the *relative mean group delay* is

$$\frac{\left\langle \tau_g^{(2)} \right\rangle_B}{\tau_2} = \frac{a_2^2 - a_2\chi}{1 + a_2^2 - 2a_2\chi} \tag{5.2.49}$$

and this is shown in Figure 5.2.10(a) as a function of a_2 and with $B\tau_2$ as a parameter.

The dotted line in Figure 5.2.10(a) ('Rihaczek bandwidth') is a distortion threshold measure discussed later in Section 5.3. For a minimum mean group delay, the value for a_2, for the above modulated sinc pulse, is

$$a_0 = \frac{B\tau_2/2}{\sin(B\tau_2/2)} - \sqrt{\left(\frac{B\tau_2/2}{\sin(B\tau_2/2)}\right)^2 - 1}, \tag{5.2.50}$$

which gives the minimum relative mean group delay as (cf. Andersen *et al.*, 1988; Poletti and Vaughan, 1992)

$$\frac{\langle \tau_g^{(2)} \rangle_{,\,\min}}{\tau_2} = \frac{1}{2}\left(1 - \left(1 - \frac{\sin(B\tau_2/2)}{B\tau_2/2}\right)^2\right)^{-1/2} \approx \frac{\sqrt{3}}{B\tau_2}, \quad B\tau_2 \ll 1. \tag{5.2.51}$$

The minima for the various $B\tau_2$ are also evident in Figure 5.2.10(a). The function is symmetric about $a_2 = 1$, so the maxima of the group delays are related to the minima in this sense, i.e. the minimum group delays are negative and result from minimum phase fades, and the maxima are positive, coming from non-minimum-phase fades. Figure 5.2.11 depicts the relative group delay as a function of a_2.

The equation for minimum mean group delay can be written as

$$\left\langle B\tau_g^{(2)} \right\rangle_{\min} \approx \sqrt{3}, \quad B\tau_2 \ll 1, \tag{5.2.52}$$

so the timing excursions of a signal passed through the two-path channel are limited to $\sim\sqrt{3}/B$ in the narrowband case. This value also represents the limiting value for the many-path case with a delay profile duration of τ_2.

To demonstrate the effect of negative group delay, Figure 5.2.12 shows a delayed sinc pulse and the resulting output pulse after passing through the fade of a two-path channel with $a_2 = 0.998$, and $B\tau_2 = 0.001$ (e.g. $\tau_2 = 100\,\mathrm{ns}$, $B = 10^3\,\mathrm{rad\,s^{-1}}$), for

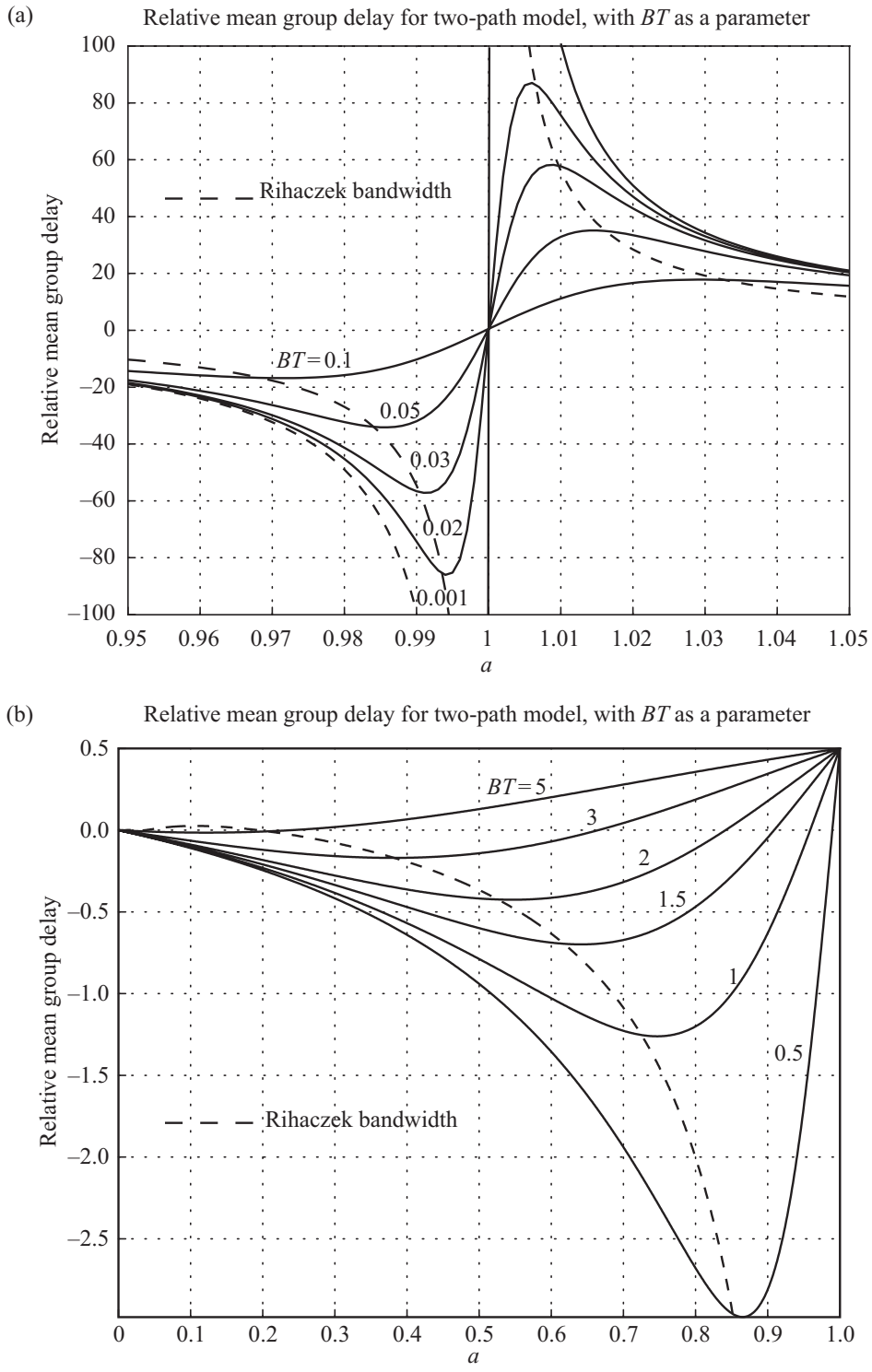


Figure 5.2.10 *The mean group delay relative to the delay difference $\tau_2 = T$ over a frequency band B centred on the minima ($a_2 < 1$) and maxima ($a_2 > 1$) of group delay, for the two-path model; $a_2 = a$ is ratio of source amplitudes. (a) Small $B\tau_2$ case, (b) large $B\tau_2$ case (negative values only)*

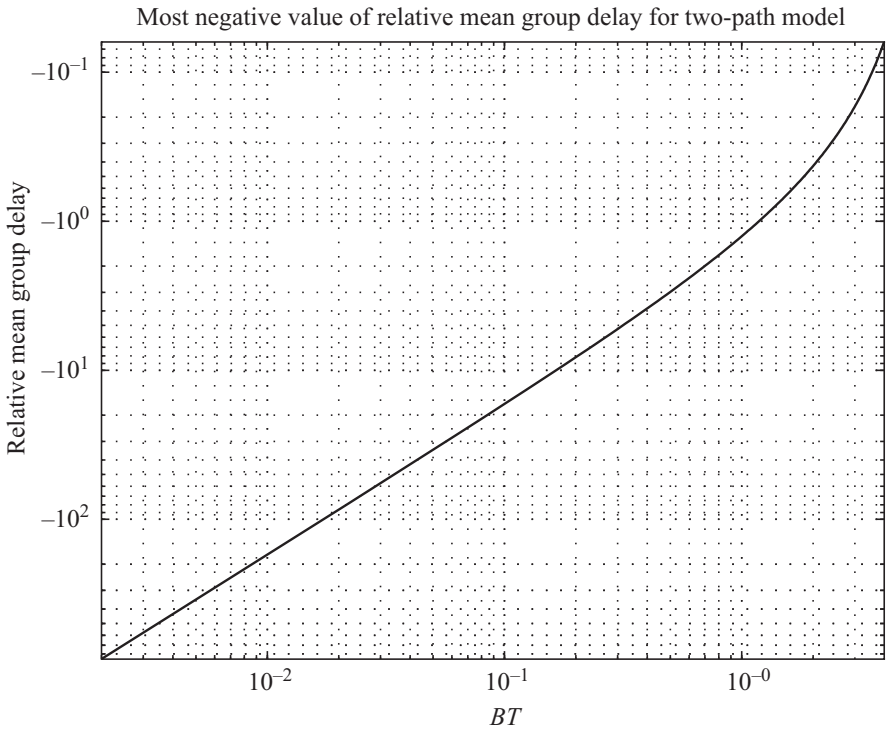


Figure 5.2.11 *The maximum negative (or maximum positive) group delay that an envelope of a modulated sinc pulse experiences on passing through a two-path static channel which is optimised for minimum (or maximum) group delay. The abscissa $B\tau_2$ is the bandwidth–delay difference product*

which $a_{\text{opt}} = 0.9997$. The choice of carrier frequency, $\omega_{\text{C}} = (2n - 1)\pi/\tau_2$, is not important as long as it occupies the centre of B. The envelope of the output signal clearly has a mean delay time less than that of the input, and the envelope has an advance in time of about $400\tau_2$ (e.g. $40\,\mu$s) from eyeballing the data. This can be compared with the value from equation (5.2.37) as $\tau_{\text{g}}/\tau_2 \approx -490$. Note that there is distortion present, which makes it difficult to assign a value of the mean delay time. The distortion is contributed to by the amplitude dispersion of the channel. More dispersion occurs by using $a_2 = a_0$ (this point is where the transfer function magnitude is at a minimum), at which stage the pulse becomes distorted to the point where the concept of group delay corresponding to the envelope delay falls down. Similarly, by shortening the pulse duration to the main lobe size by windowing or using an equivalent Gaussian-shaped pulse (see following section), the amplitude dispersion of the pulse itself contributes distortion. Mathematically, the mean excess time is still correct, of course, being the energy-weighted centre of the signal, but this measure does not directly reflect distortion of the envelopes. With wider bandwidths, the output pulse appears as two separated pulses because the carrier interference has caused the complex envelope of the pulse to go through zero (meaning a π phase shift in the carrier) before the end of the pulse energy flow, and again the interpretation of group delay as envelope delay has broken down owing to the gross distortion. Examples are given in Section 5.3.

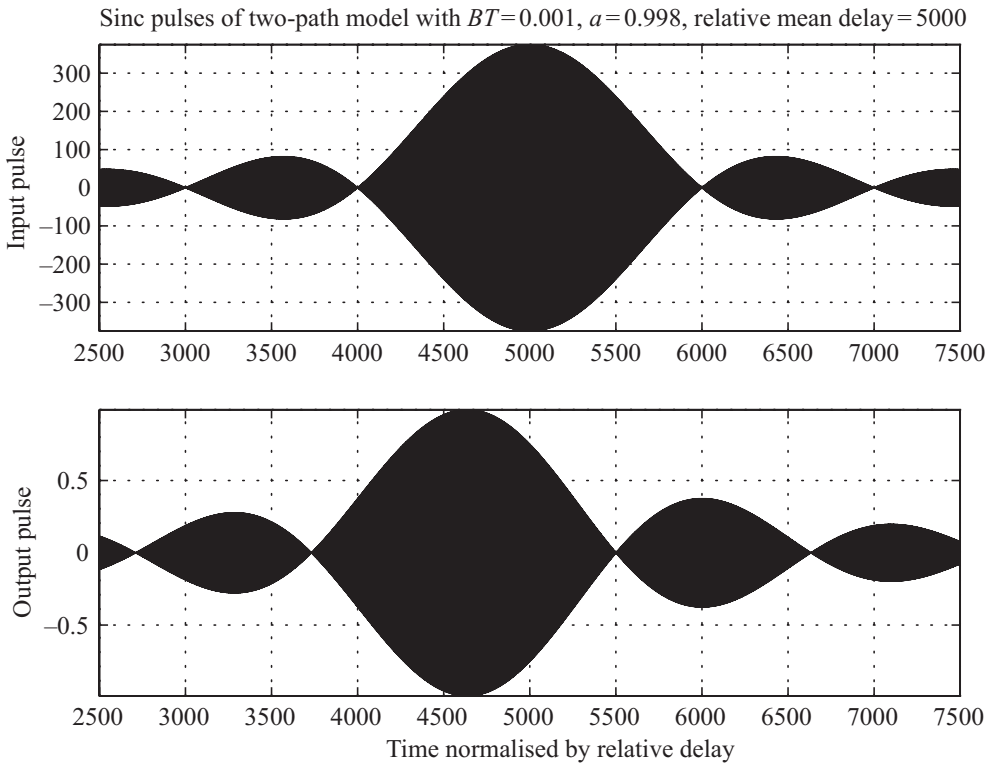


Figure 5.2.12 *A modulated sinc pulse before and after passing through a two-path channel with negative mean group delay. The bandwidth is narrow (in the sense $B\tau_2 = 0.001$) and the pulse envelope is advanced in time as a result of passing through the two-path channel*

5.2.10.2 Mechanism for negative group delay

The physical time-domain mechanism for the negative (and also positive) group delay is elucidated by writing the output signal, $y(t)$, as (Poletti and Vaughan, 1992)

$$\begin{aligned} y(t) &= x(t)\cos(\omega_C t) + ax(t-\tau_2)\cos(\omega_C(t-\tau_2)) \\ &\approx (x(t) - ax(t-\tau_2))\cos(\omega_C t), \end{aligned} \tag{5.2.53}$$

since $\omega_C \tau \approx \pi$ at the fade. The mean time is thus reduced by destructive interference of the carrier, whereby the value of the output pulse envelope is reduced less at early times than it is at late times, which creates an effective advancing in time of the envelope. Note that in the example above, the channel gain is about −50 dB. A channel occupying a region near a fade will invoke large excursions of the group delay. The classical envelope advance interpretation of the group delay is clearly confined to be much less than the signal main lobe duration, i.e. $B\tau_2 \ll 1$. Conversely, increasing the bandwidth can be viewed as limiting the excursions of the mean group delay.

For a very narrow channel and a moving terminal, the jitter of the signal time caused by the changing group delay in the channel can create problems with symbol and carrier timing synchronisation. The wider the bandwidth of the (narrowband)

signal, the less the timing jitter. Again, some caution should be exercised with interpreting the mean group delay as an envelope shift – distortion of the envelope itself is also an important factor in the sense that the timing jitter may be interpreted as a timing ambiguity. This ambiguity is a mechanism that may lead to synchronisation errors and an associated higher (than otherwise expected) bit error ratio. The behaviour follows from the simple two-path channel model when the distortion is bad. In practice, a faded many-path channel can create these problems as well.

For wideband signals, where $B\tau_2 \gg 1$, the two-path model gives two resolvable impulses in the impulse response. Here, the jitter could be considered as τ_2, but the assumption is made that with a resolvable delay difference, the synchroniser should lock on the first one, which in the context of the static two-path scenario definition, is always the larger. The jitter in this case goes to zero because the scenario is static. However, in a dynamic scenario the two impulses have independently changing powers, and in this more general two-path case, timing jitter of τ_2 can occur if the impulse that is dominant (higher power) toggles between the two delay values. A rake equaliser may be gainfully employed in this situation.

5.2.10.3 Summary

The existence of negative group delay can be readily found in simulations of many-path models and by measurement of actual channels. This behaviour has a simple explanation, and its mechanism has been illustrated here using the two-path model. The interpretation of the phase derivative of the transfer function as group delay requires constraints for both the signal and the channel because amplitude dispersion is not taken into account in the group delay. The mean delay (group delay averaged over a band) should not be interpreted as an envelope delay (or advance) unless the amount of distortion, i.e. the dispersion, can be neglected. The conditions for this are investigated in the following section. Finally, the same discussion applies to positive group delays, this case corresponding to the non-minimum phase channel ($a_2 > 1$).

5.2.11 Coherence bandwidth

An important parameter in a frequency selective fading channel is the frequency separation for which the fading becomes effectively independent in the statistical sense (Appendix B). This frequency separation is defined from the autocorrelation coefficient function of the channel transfer function. It is presented here as independent of frequency, i.e. it is assumed wide sense stationary. The autocorrelation coefficient function, or normalised *frequency coherence function* (fcf), is estimated over a finite bandwidth by

$$\hat{C}(\Omega) = \frac{\int H(\omega)H^*(\omega+\Omega)\,\mathrm{d}\omega}{\int |H(\omega)|^2\,\mathrm{d}\omega} \tag{5.2.54}$$

(the hat denotes estimate) and is the Fourier transform of the normalised power delay profile (the estimate sign is omitted from the delay profile here)

$$\hat{C}(\Omega) \quad \Leftrightarrow \quad \frac{P(\tau)}{\int P(\tau)\,\mathrm{d}\tau} = \frac{\langle |h(\tau)|^2 \rangle}{\int \langle |h(\tau)|^2 \rangle\,\mathrm{d}\tau}. \tag{5.2.55}$$

The estimate corresponds to the true value only if the functions are ergodic and when the integral limits go to infinity. For the static, deterministic two-path scenario, the estimate is exact, the averaging is unnecessary, and using the form

$$P(\tau) = \frac{1}{1+a_2^2}\left[\delta\left(\tau - \frac{\tau_2}{2}\right) + a_2^2\,\delta\left(\tau + \frac{\tau_2}{2}\right)\right]$$

$$\Updownarrow$$

$$C^{(2)}(\Omega) = \frac{1}{1+a_2^2}\left[\mathrm{e}^{\mathrm{j}(\Omega\tau_2/2)} + a^2\,\mathrm{e}^{-\mathrm{j}(\Omega\tau_2/2)}\right], \tag{5.2.56}$$

i.e. the magnitude of the fcf (this is a power quantity)

$$|C^{(2)}(\Omega)| = \frac{\sqrt{\left(1-a_2^2\right)^2 + (2a_2\cos\Omega\tau_2)^2}}{1+a_2^2}$$

$$= \left|\cos\frac{\Omega\tau_2}{2}\right| \quad \text{for } a_2 = 1. \tag{5.2.57}$$

The coherence bandwidth $\Omega_\mathrm{C} = 2\pi B_\mathrm{C}\,\mathrm{rad\,s^{-1}}$ is defined as the frequency span from the maximum (unity) of the frequency coherence function to where the magnitude of the function first drops to a value C_C, i.e. from

$$|C(\Omega = \Omega_\mathrm{C})| = C_\mathrm{C}. \tag{5.2.58}$$

C_C is taken by various authors from between $1/\mathrm{e} = 0.37$ to 0.9 (cf. Bello and Nelin, 1963; Cox, 1972; Gans, 1972; Cox and Leck, 1975). A change of C_C scales nonlinearly the coherence bandwidth, so any results derived from some value of C_C are also scaled in some way. For the two-path model, the coherence bandwidth in Hz is thus

$$B_\mathrm{C}^{(2)} = \frac{1}{\pi\tau_2}\cos^{-1}\sqrt{\frac{C_\mathrm{C}^2\left(1+a_2^2\right)^2 - \left(1-a_2^2\right)^2}{4a_2^2}}$$

$$= \frac{1}{\pi\tau_2}\cos^{-1}C_\mathrm{C}, \quad \text{for } a_2 = 1. \tag{5.2.59}$$

Figure 5.2.13 depicts the two-path functions.

The coherence function is periodic in Ω for the two-path channel. The period is $2\pi/\tau_2$ and so the coherence bandwidth decreases with increasing delay difference between the two incident waves. Also, the coherence bandwidth decreases with increasing relative amplitudes, a_2. When a_2 is small, the coherence bandwidth becomes undefined, as the coherence function does not drop down to C_C. It is evident that Ω_C is minimum for $a_2 = 1$, i.e. equal powers in each path.

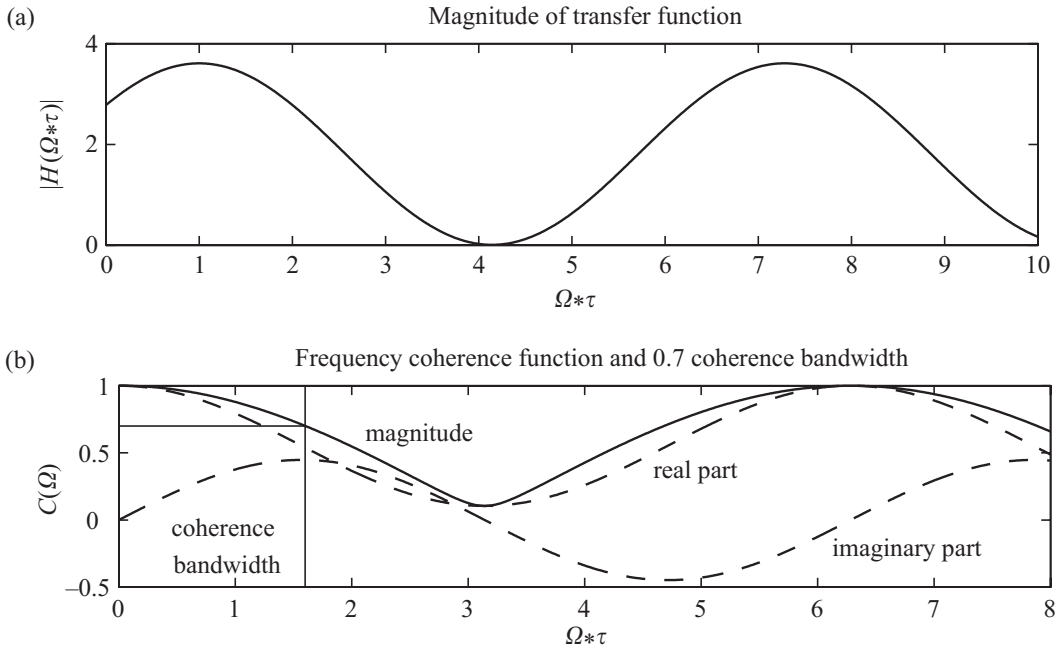


Figure 5.2.13 *The coherence bandwidth, $\Omega_C\ rad\ s^{-1}$, for the two-source scenario with $a_2 = 0.9$. (a) The magnitude of the transfer function of the channel and (b) the coherence function with the $C_C = 0.7$-coherence bandwidth is indicated*

The minimum of $|C^{(2)}(\omega)|$ occurs when $\Omega = \pi/\tau_2$, and its value is

$$\left|C_C^{(2)}\left(\Omega = \frac{\pi}{\tau_2}\right)\right| = \frac{1 - a_2^2}{1 + a_2^2}. \tag{5.2.60}$$

This is plotted in Figure 5.2.14 in order to indicate the limits for the existence of the coherence bandwidth as a function of the ratio of powers in the two incident waves. The coherence bandwidth is not particularly sensitive to the value of the relative amplitude a_2, especially when a_2 is near 1.

5.2.12 Coherence bandwidth–delay spread product

While the delay spread is a measure of the channel-time dispersion, the coherence bandwidth is a measure of the fading rate with changing frequency. The ideal communications channel has a zero delay spread and infinite coherence bandwidth. For the two-path model, the delay spread increases while the coherence bandwidth decreases for increasing relative delay τ_2 and increasing relative amplitude a_2. The coherence bandwidth and the delay spread are thus inversely related, but the exact relationship is not simple.

The product of these two parameters was taken for experimental channels using $C_C = 0.75$, by Gans (1972), who found an empirical law that $B \cdot s$ was constant and approximately equal to 1/8. Fleury (1996) views the product as an uncertainty

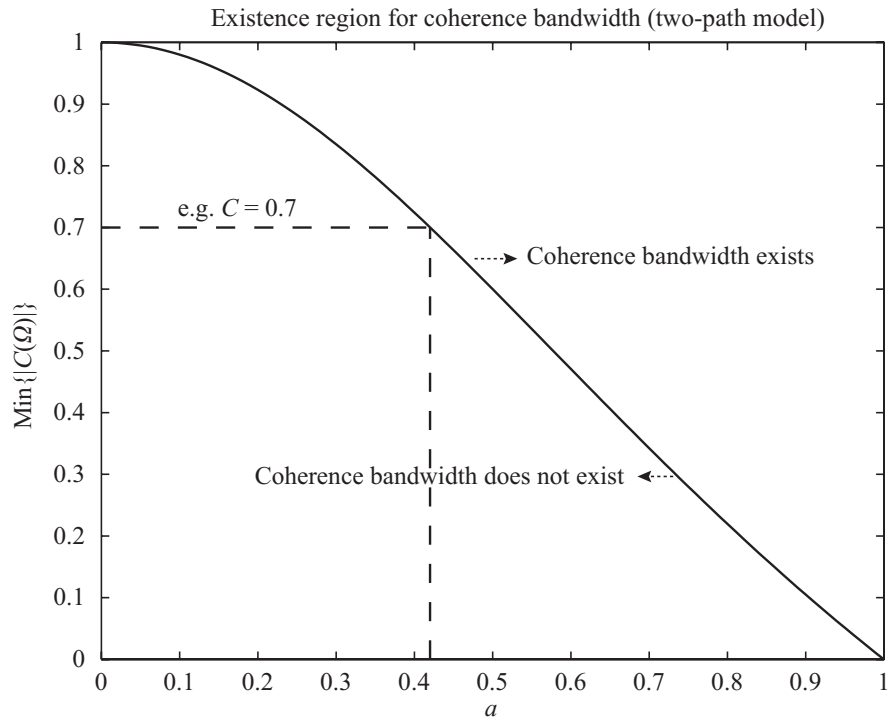


Figure 5.2.14 *The minimum value of the two-path coherence function versus the relative power of the sources. The coherence bandwidth does not exist (i.e. is infinite) when $C_C < min\{|C(\Omega)|\}$ which is to the left of the curve*

principle and gives a lower bound for the many-path channel as $1/(2\pi)\cos^{-1} C_C$, so that

$$B_C(C_C) \cdot s \geq \frac{1}{2\pi}\cos^{-1}(C_C) \tag{5.2.61}$$

and also notes that the bound is obtained for the two-path case with equal powers. Recall that this case, $a_2 = 1$, corresponds to maximum delay spread.

For the two-path channel, the product is

$$B_C^{(2)} \cdot s^{(2)} = \frac{a_2}{\pi\left(1 + a_2^2\right)}\cos^{-1}\sqrt{\frac{C_C^2(1 + a_2^2)^2 - (1 - a_2^2)^2}{4a_2^2}}, \tag{5.2.62}$$

which of course is independent of the relative delay τ_2. This product is depicted in Figure 5.2.15.

The dependence of the product on the relative amplitudes a_2 is weak, with a greater influence being the choice of C_C. The $B_C^{(2)} \cdot s^{(2)}$ product is a minimum when a_2 is 1, i.e. when the frequency fades are the deepest. In this case, the expression reduces and agrees with Fleury's result in equation (5.2.61)) as noted above, and for the value $C_C = 0.75$ is also in close agreement with Gans' law, i.e. $B_C^{(2)} \cdot s^{(2)} = 1/(2\pi)\cos^{-1}(0.75) \approx 0.115 \approx 1/8$.

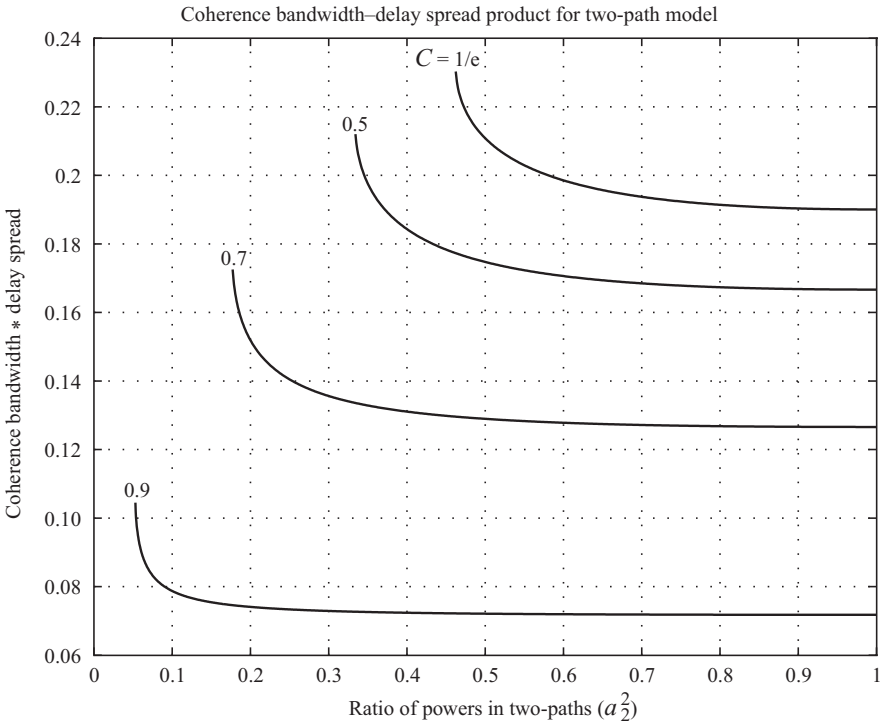


Figure 5.2.15 *The product of the coherence bandwidth–delay spread product as a function of the relative powers in the incident waves, a^2, for the two-source model. The coherence bandwidth, and thence the $B^{(2)}s^{(2)}$ product, does not exist for small a_2*

In the two-path model then, the virtually constant value of the product allows the delay spread to be calculated from a measured correlation bandwidth, or vice versa. However, in a general many-path case, the expression for the product must be recalled as a lower limit. It should always be borne in mind that the choice of C_C for the coherence bandwidth affects the value of the product. Note, however, that there is no upper limit for the $B \cdot s$ product in a mathematical sense. In the two-path model, for example, the coherence bandwidth B_C may be undefined (depending on the value of a_2 in the channel and the choice of C_C in the definition) which may be interpreted as $B \cdot s$ being infinite. An interesting example occurs when there are two strong contributions (cf. $a_2 \rightarrow 1$ in the two-path model) plus a third small contribution a_3 further away in delay time. Figure 5.2.16 shows the example. In this case, the strong pair dominates the frequency correlation function, which goes close to zero regularly. The third path contribution adds a small amount of ripple to the frequency correlation function. However, the third path can increase the delay spread without bound, so that the $B \cdot s$ product can be also unbounded even though the coherence bandwidth remains essentially constant (see also equations (5.3.7) and (5.3.8) below). In practice, contributions at very long delay times become buried in the noise owing to the increasing path loss with delay time, and so the $B \cdot s$ product does have an upper bound which is governed by propagation conditions and signal-to-noise ratio.

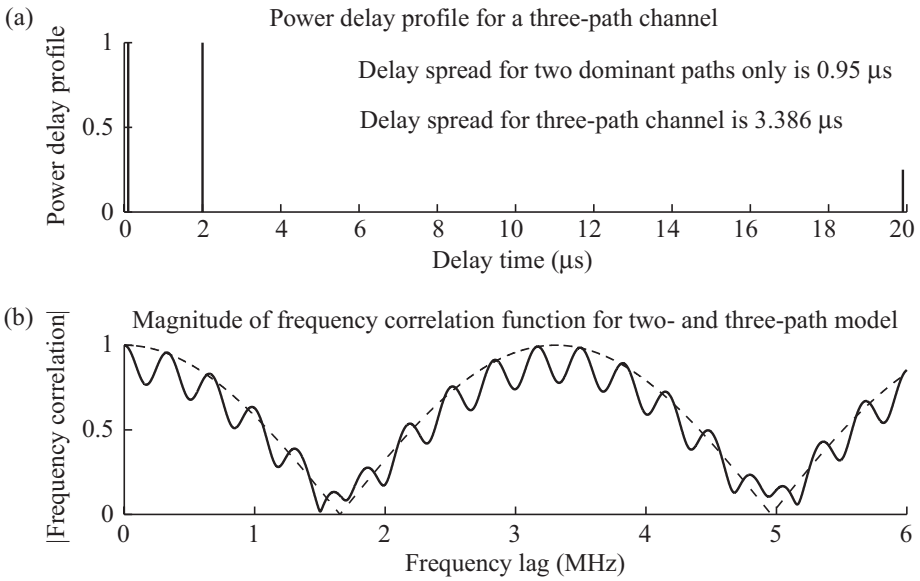


Figure 5.2.16 *(a) An example where two large contributions close together at early time, and a third, small contribution at late time make up the impulse response. The delay spread is dependent on the time delay of the third contribution. (b) The frequency coherence function is dominated by the two strong paths, and these govern the definition of the coherence bandwidth. In this example, the $B \cdot s$ product is bounded by the delay spread, which may be large*

An analogous product, that of the spread of the Doppler frequencies and the correlation distance, is discussed below for the moving terminal.

5.2.12.1 Summary

Behaviour of the group delay and delay spread is important in the performance of communications systems operating in frequency selective channels. The existence of negative group delay can be readily found in simulations of many-path models and by measurement of actual channels. This unlikely behaviour has an explanation, and its mechanism has been illustrated here using the two-path model. The interpretation of the phase derivative of the transfer function as a pure group delay requires constraints for both the signal and the channel because amplitude dispersion is not taken into account in the usual definition for group delay. The mean delay (group delay averaged over a band) should not be interpreted as an envelope delay (or advance) unless the amount of distortion, i.e. the dispersion, can be neglected. The same interpretation applies to the more familiar positive group delays. In many-path channels, the coherence bandwidth can remain constant, but the delay spread can continue to increase such that the coherence bandwidth-delay spread product does not have a well defined general upper bound. In the two-path model, the product follows Gans' empirical law and this is the lower bound.

5.3 Dispersion metrics from time–frequency theory for the static frequency-selective channel

Section 5.2 demonstrated that the mean group delay, which is equivalent to the mean excess delay time, cannot always be interpreted as envelope delay around a deep fade, owing to the distortion caused by the channel. In this section, the dispersion, and in particular the delay spread, is investigated in the frequency domain. The purpose is to gain deeper insight into the distortion mechanisms and to develop dispersion metrics which suit the frequency selective fading channel. The channel under study is static and frequency dependent. However, the discussion can, by analogy, apply also to a position- (or time-) dependent mobile channel. This section is independent of the successive sections so the reader can skip directly to later sections without a loss of continuity. This section draws heavily on Cohen (1989), who reviews time–frequency theory including the concepts of local and global variances for time and frequency, and the control of the time–frequency distribution through the properties of an integral kernel, and many of the original contributors are listed in this reference rather than cited here; parts of this section are taken from Poletti and Vaughan (1992), who investigated the concept of negative group delay from time–frequency theory.

5.3.1 Terminology

The terminology is as follows. The *distortion* refers to changes in a signal passed through a channel and the *dispersion* is the variance of the channel impulse response, i.e. dispersion is associated with the channel, although a signal after passing through a dispersive channel can be referred to as dispersed. The *global variance* in time is the delay spread squared and the local (in frequency) variance is referred to as the *local delay spread* squared, and these terms are used interchangeably in this section.

Both the mean group delay, which is also referred to here as simply the mean delay time, and the dispersion are (conditional) moments of a time–frequency distribution describing the channel. The theory of time–frequency distributions is therefore called on and the available results are compared with expansions of the channel transfer function about a local frequency within the channel. Time–frequency theory is for signals, rather than channel behaviour, and the adaptation here attempts to avoid, without ignoring, some of the interpretation problems which currently accompany time–frequency theory. The dispersion terms are summarised in Table 5.1.

5.3.2 Time–frequency distribution discussion

5.3.2.1 Introduction

The use of generalised time–frequency distribution theory (Claasen and Mecklenbrauker, 1980; Cohen, 1989) gives a basis for discussing the dispersion of a signal. The goal of the time–frequency theory is to find a distribution, i.e. a joint time–frequency function, that represents the energy of a signal simultaneously in time and frequency. The signal here is the impulse response of a static, frequency selective channel.

Table 5.1 Summary of dispersion terms for the time–frequency theory (left column) and its application to static, frequency-fading channels (right column). The prime denotes differentiation with respect to the argument and the angle brackets denote averaging with respect to the subscript

Time–frequency signal dispersion parameter (t is signal time)	Fading channel dispersion parameter signal is the impulse response (τ is delay time)
local (in frequency) mean time $t(\omega) = -\phi'(\omega)$	mean delay time at a given frequency (channel group delay) $\tau_g(\omega) = -\phi'(\omega)$
instantaneous (in time) frequency $\omega_{inst}(t) = \phi'(t)$	instantaneous frequency at a delay time $\omega_{inst}(\tau) = \phi'(\tau)$
instantaneous (in time) frequency variance $\sigma_\omega^2(t)$ (signal instantaneous bandwidth squared)	channel instantaneous bandwidth squared at a given delay time
local (in frequency) time variance $\sigma_t^2(\omega)$ (local dispersion of signal)	local delay spread squared $\sigma_\tau^2(\omega)$ (local dispersion of channel)
global mean time $\langle t \rangle$	mean delay time (= mean group delay of channel)
global time variance σ_t^2	delay spread squared, $s^2 = \sigma_\tau^2$

The spectrogram is the traditional time–frequency tool, using a short-time Fourier transform over a sliding window to characterise changing spectra. This immediately introduces ambiguities if the signal changes occur within the window. Also, the window itself introduces a fundamental tradeoff between the resolution in the time and frequency domains. This tradeoff is expressible as a form of the uncertainty principle given as

$$\sigma_t \sigma_\omega \geq \tfrac{1}{2}, \tag{5.3.1}$$

where σ_t^2 and σ_ω^2 are the variances of the signal and its spectrum respectively[2] and which are of course moments of the time–frequency distribution. The equality holds for Gaussian form only. The 'uncertainty' here is better expressed as the *certainty* that a small duration signal has a wide bandwidth or a narrow bandwidth supports only long duration signals. However, for statistical signals, the uncertainty relation is recognised as equivalent to the limit on the coherence bandwidth (B Hz, where the normalised frequency correlation function has dropped to a value C_C)–delay spread (s)

[2] Papoulis (1962) notes that the principle holds for signals decaying faster than $t^{-1/2}$ asymptotically, and that the equality condition cannot hold for causal signals.

product from Section 5.2 after Fleury (1996),

$$B \cdot s \geq \frac{1}{2\pi} \cos^{-1}(C_C). \tag{5.3.2}$$

The time–frequency theory is currently considered incomplete and its development to date has been often characterised by attempts to formulate and interpret time-varying spectra. Most applications beyond the periodogram have been limited to visual interpretation of a 2D or contour plot of the time–frequency distribution. Generalisation is hindered in particular by the local time variance (here, the delay spread squared at a given frequency) being poorly defined in the sense that it can have any number of expressions or that it violates intuitive behaviour (Cohen, 1989). However, one form of the global variance (here, the delay spread squared) features complementary dispersion contributions from the magnitude and phase variations of the channel. This suggests that a generalised group delay, containing both phase and magnitude terms, can be used to quantify the dispersion of the general channel. This is supported by local expansion of the transfer function (Poletti and Vaughan, 1992) in the following section. However, for fading channels, the phase and magnitude tend to be related about the deeper fades (the log of the magnitude and the phase can be Hilbert transforms about the fade region). This means that the inclusion of just phase terms is sufficient, at least about the deep fades and this is the region causing the greatest signal distortion. A result is, after Rihaczek (1969), that the absolute phase curvature is a suitable measure for quantifying the dispersion of channels with frequency selective fades.

The use of several time variables is eased by collating their definitions, along with their Fourier frequency symbols, in Table 5.2. The dual use of Fourier frequency symbols results from the definition of signal and correlation transforms, e.g. the Fourier transform is written $x(t) \Leftrightarrow X(\omega)$ and an autocorrelation Fourier transform is written $R_x(\alpha) \Leftrightarrow |X(\omega)|^2$.

The time–frequency approach is based on the time–frequency distribution, $P(t, \omega)$, which is for describing the instantaneous energy distribution of a signal. Here the signal is the impulse response, $h(\tau)$, of a linear, time-invariant channel, and the time–frequency distribution becomes denoted $P(\tau, \omega)$.

Table 5.2 The time and frequency symbols used in this section

Time quantity	Time symbol	Fourier frequency symbol
signal time for signal $s(t)$	t	ω
delay time in impulse response (or signal) $h(\tau)$	τ	θ
displacement time for impulse response autocorrelation function	α	ω
dummy delay time variable used in autocorrelation integral	T	θ

5.3.2.2 Correlation function

The autocorrelation function, in time–frequency tradition, is typically expressed with symmetrical displacements and includes a key feature – a weighting function, or window, written $r(T, \alpha)$, but which contains dependence on the delay time τ in the autocorrelation:

$$R_h^{(r)}(\tau, \alpha) = \frac{1}{2\pi} \int r(T - \tau, \alpha) h\left(T - \frac{\alpha}{2}\right) h^*\left(T + \frac{\alpha}{2}\right) \mathrm{d}T. \tag{5.3.3}$$

The 2π factor relates to the Fourier transform convention, which is traditional in the time–frequency context (see below). The weighting function is chosen to emphasise the autocorrelation function contributions near some delay time τ (i.e. about $T = \tau$). This can be viewed as trying to focus on the impulse response energy at time τ, defined by the windowed autocorrelation. In time–frequency theory, the choice of the window r alters $P(\tau, \omega)$ so that it can be arranged to suit intuitive ideas of what a time-varying spectrum should look like. The time–frequency distribution is a Fourier transform of the autocorrelation function, here arranged as

$$\frac{1}{2\pi} \int R_h^{(r)}(\tau, \alpha)\, \mathrm{e}^{-\mathrm{j}\omega\alpha}\, \mathrm{d}\alpha = P(\tau, \omega) \quad \Leftrightarrow \quad R_h^{(r)}(\tau, \alpha) = \int P(\tau, \omega)\, \mathrm{e}^{\mathrm{j}\omega\alpha}\, \mathrm{d}\omega \tag{5.3.4}$$

($\Leftrightarrow$ denotes a Fourier transform)[3]. In stating a time–frequency relationship between R and P there is an inherent tradeoff ('uncertainty') made available between the domains, and affected, or ideally controlled, by r. A short-duration weighting function, or window, centred at delay time τ, means a wide spectrum at τ; the converse is also true. Research issues in the time–frequency community have included seeking different weighting functions r to suppress unwanted ('unintuitive') cross products in $P(\tau, \omega)$ and maintaining the distribution as a positive and real function.

5.3.2.3 Generalised form of time–frequency distribution

The Fourier transform of the autocorrelation weighting function is $k(\theta, \alpha)$, denoted by

$$k(\theta, \alpha) \quad \Leftrightarrow \quad r(T, \alpha) = \int k(\theta, \alpha) \mathrm{e}^{\mathrm{j}\theta T}\, \mathrm{d}\theta. \tag{5.3.5}$$

Combining equations (5.3.3)–(5.3.5) results in a general form of time- and frequency-shift-invariant distribution given by (Claasen and Mecklenbrauker, 1980; Cohen, 1989)

$$P(\tau, \omega) = \frac{1}{4\pi^2} \int \int \int k(\theta, \alpha)\, h\left(T - \frac{\alpha}{2}\right) h^*\left(T + \frac{\alpha}{2}\right) \mathrm{e}^{\mathrm{j}\theta T - \mathrm{j}\alpha\omega - \mathrm{j}\theta\tau}\, \mathrm{d}T\, \mathrm{d}\alpha\, \mathrm{d}\theta, \tag{5.3.6}$$

in which context $k(\theta, \alpha)$ is referred to as the kernel function, and following Cohen (1989), the integral limits are from minus to plus infinity. This general form offers the

[3] Note that electrical engineering tradition has the 2π term on the right-hand side equation, rather than on the left-hand side equation used often in time–frequency analysis and used above.

advantage that general results can be deduced and properties of specific distributions[4] elucidated. The equation contains the ambiguity function (cf. Rihaczek, 1969) of $h(\tau)$,

$$\chi(\theta, \alpha) = \int h\left(T - \frac{\alpha}{2}\right) h^*\left(T + \frac{\alpha}{2}\right) \mathrm{e}^{\mathrm{j}\theta T}\,\mathrm{d}T. \tag{5.3.7}$$

This function was introduced by Ville (1948) and applied to radar by Woodward (1953). It is the two-dimensional autocorrelation $h(\tau)$ with its delayed and frequency-shifted replica, $h(\tau + \alpha)\mathrm{e}^{\mathrm{j}\theta\tau}$, and it is used to indicate the tradeoff ('ambiguity') between the time delay time resolution and frequency shift resolution. A short delay time means a broad spectral support, for example. Substituting $\chi(\theta, \alpha)$, equation (5.3.6) is re-written as

$$P(\tau, \omega) = \frac{1}{4\pi^2} \int\int k(\theta, \alpha)\chi(\theta, \alpha)\mathrm{e}^{-\mathrm{j}(\theta\tau + \omega\alpha)}\,\mathrm{d}\theta\,\mathrm{d}\alpha, \tag{5.3.8}$$

i.e. the Fourier transform pair is

$$P(\tau, \omega) \quad \Leftrightarrow \quad k(\theta, \alpha)\chi(\theta, \alpha) = \int\int P(\tau, \omega)\mathrm{e}^{\mathrm{j}(\theta\tau + \omega\alpha)}\,\mathrm{d}\tau\,\mathrm{d}\omega. \tag{5.3.9}$$

5.3.2.4 Kernel choice for defining distribution properties

The properties of the time–frequency distribution, $P(\tau, \omega)$, are controlled by the choice of kernel function (i.e. the autocorrelation window function). Elaboration of this is given in Cohen (1989) and a summary is given here. The kernel is kept independent of delay invariant, i.e.

$$h(\tau + \tau_0) \rightarrow P(\tau + \tau_0, \omega) \tag{5.3.10}$$

$$h(\tau)\ \mathrm{e}^{-\mathrm{j}\omega_0\tau} \rightarrow P(\tau, \omega + \omega_0). \tag{5.3.11}$$

To ensure that the autocorrelation function is even, $k(\theta, \alpha) = k^*(-\theta, -\alpha)$. This condition also enforces the (intuitive) requirement that $P(\tau, \omega)$ be real. There are moment-derived properties which are considered desirable because they relate directly to the one-dimensional moments. The total, or global, energy being constant and normalised, i.e.

$$\int\int P(\tau, \omega)\,\mathrm{d}\omega\,\mathrm{d}\tau = 1, \tag{5.3.12}$$

which allows $P(\tau, \omega)$ to be used directly as a pdf, requires $k(0, 0) = 1$. The zeroth moments (marginals), should yield the power delay profile $P(\tau) = \langle|h(\tau)|^2\rangle = |h(\tau)|^2$ and the spectral magnitude $P(\omega) = \langle|H(\omega)|^2\rangle = |h(\omega)|^2$, respectively. Also, the kernel should be constrained so that what corresponds intuitively to the

[4] Historically, the Wigner ($k = 1$), Page ($k = \exp(\mathrm{j}\theta\alpha/2)$) and Rihaczek ($k = \exp(\mathrm{j}\theta\,|\alpha|\,/2)$) distributions, for example, have been useful for a given application, but less useful for general signals (cf. Cohen, 1989).

instantaneous frequency is the mean frequency at a given delay time. This is given by the first conditional moment of frequency, denoted

$$\omega_{\text{inst}}(\tau) = \langle \omega \mid \tau \rangle = \frac{\int \omega P(\tau, \omega)\, \mathrm{d}\omega}{\int P(\tau, \omega)\, \mathrm{d}\omega} = \frac{\int \omega P(\tau, \omega)\, \mathrm{d}\omega}{P(\tau)}. \tag{5.3.13}$$

Similarly, for the group delay, which is the mean delay time at a given frequency, and is the first conditional moment of delay time,

$$\tau_{\text{g}}(\omega) = \langle \tau \mid \omega \rangle = \frac{\int \tau P(\tau, \omega)\, \mathrm{d}\tau}{\int P(\tau, \omega)\, \mathrm{d}\tau} = \frac{\int \tau P(\tau, \omega)\, \mathrm{d}\tau}{P(\omega)}. \tag{5.3.14}$$

Cohen (1989) gives the result that the marginal properties require $k(\theta, 0) = 1$ and $k(0, \alpha) = 1$, and the first moment properties require $k'_\tau(\theta, \alpha)\ |_{\alpha=0} = 0$ and $k'_\theta(\theta, \alpha)\ |_{\theta=0} = 0$, respectively. The kernel is normally a 'product kernel', meaning one containing the variables θ and α only as a product, $\theta\alpha$, so

$$\begin{aligned} k(\theta, \alpha) = k(\theta\alpha) : k(0) = 1, &\quad \text{for zeroth moments,} \\ k'(0) = 0, &\quad \text{for first moments.} \end{aligned} \tag{5.3.15}$$

However, the interest in time–frequency distribution theory is in departing from $k'(0) = 0$ in order to try to find distributions which describe the behaviour of real-world signals. Cohen (1989, p. 954) notes in particular that strictly, the instantaneous frequency is the phase derivative with respect to time only if an analytic signal is used (i.e. the analytic impulse response would have to be formed), whereas an analytic signal form is not required in the above formulation, and points out from this that the time–frequency theory is not currrently consistent with respect to the instantaneous frequency. By symmetry, the group delay, which is the (negative) phase derivative with respect to frequency, falls into the same dubious category. With this inconsistency in mind, some of the possible time–frequency results are given below in order to find common ground with results from local signal expansion, and ascertain their suitability for providing dispersion measures for frequency selective fades in channels.

5.3.2.5 Local variances: instantaneous bandwidth and local delay spread

The bandwidth of a channel at delay time τ can be measured by the conditional spectral variance, expressed using the notation above, and this can be called the instantaneous bandwidth (Cohen and Lee, 1990; Poletti, 1993)

$$\sigma^2_\omega(\tau) = \langle \omega^2 \mid \tau \rangle - \langle \omega \mid \tau \rangle^2 \quad \text{(instantaneous bandwidth squared).} \tag{5.3.16}$$

Similarly, the delay spread at a given frequency, ω, is measured by the conditional delay time variance, which we refer to as a local delay spread,

$$\sigma^2_\tau(\omega) = \langle \tau^2 \mid \omega \rangle - \langle \tau \mid \omega \rangle^2 \quad \text{(local delay spread squared).} \tag{5.3.17}$$

For a linear channel this defines the dispersion of the channel as a function of frequency and is the focus of this section (Poletti and Vaughan, 1992). The quantities of equations (5.3.16) and (5.3.17) are also defined formally for signals in Cohen (1989). A number of forms for this may be obtained for different time-frequency kernels as demonstrated below.

The local mean time (mean delay at a given frequency) of the channel is the first conditional moment of time and is found from the derivative of the characteristic function at $\theta = 0$. The result, in terms of the product kernel, is (cf. Cohen and Lee, 1988)

$$\langle \tau \mid \omega \rangle P(\omega) = A^2(\omega) \left\{ -k(0)\phi'(\omega) + 2k'(0)\frac{A'(\omega)}{A(\omega)} \right\}, \tag{5.3.18}$$

where the channel power transfer function is denoted

$$|H(\omega)|^2 = \left| A(\omega)\, \mathrm{e}^{\mathrm{j}\phi(\omega)} \right|^2 = A^2(\omega) = P(\omega) \tag{5.3.19}$$

for a static channel. For $k(0) = 1$ and $k'(0) = 0$, this moment becomes the group delay as noted above.

The mean square time at a given frequency is the second conditional moment, found from the second derivative of the characteristic function, here with $k(0) = 1$, as (cf. Cohen and Lee, 1990),

$$\begin{aligned} &\langle \tau^2 \mid \omega \rangle P(\omega) \\ &= A^2(\omega) \left(2k''(0) \left[\frac{A''(\omega)}{A(\omega)} + \left(\frac{A'(\omega)}{A(\omega)} \right)^2 \right] - 2k'(0) \left[\phi''(\omega) + 2\frac{A'(\omega)}{A(\omega)}\phi'(\omega) \right] \right) \\ &\quad + A^2(\omega) \left[\frac{1}{2} \left(\frac{A'(\omega)}{A(\omega)} \right)^2 - \frac{1}{2}\frac{A''(\omega)}{A(\omega)} + \left(\phi'(\omega) \right)^2 \right] \end{aligned} \tag{5.3.20}$$

and so the local delay spread squared, $\langle \tau^2 \mid \omega \rangle - \langle \tau \mid \omega \rangle^2$, follows as (Poletti and Vaughan, 1992)

$$\begin{aligned} \sigma_\tau^2(\omega) = &\frac{1}{2} \left[\frac{A'(\omega)}{A(\omega)} \right]^2 \left[1 - 8\left(k'(0) \right)^2 + 4k''(0) \right] \\ &- \frac{1}{2} \left[\frac{A''(\omega)}{A(\omega)} \right] \left[1 - 4k''(0) \right] - 2k'(0)\phi''(\omega). \end{aligned} \tag{5.3.21}$$

This is a general expression for the local delay spread of the channel in terms of two parameters, $k'(0)$ and $k''(0)$, and the first and second derivatives of the log magnitude and phase of the transfer function (see below). Different forms for the local delay spread may thus be derived depending on the chosen kernel. The problem is to identify one which is appropriate for the frequency selective channel.

5.3.2.6 Global time variance: delay spread

The mean delay is found by integrating the local mean time over the frequency band of the channel. It is the expected value of the first conditional moment of time, here normalised with the signal having unit energy,

$$\langle \tau \rangle = \int \langle \tau \mid \omega \rangle P(\omega)\, \mathrm{d}\omega = \int\int \tau P(\tau, \omega)\, \mathrm{d}\omega\, \mathrm{d}\tau = \int \tau P(\tau)\, \mathrm{d}\tau. \tag{5.3.22}$$

Similarly the mean square time, or expected value of the second conditional moment, is denoted $\langle \tau^2 \rangle$. The delay spread squared is the global variance, which follows as

$$\sigma_\tau^2 = \langle \tau^2 \rangle - \langle \tau \rangle^2 = \int \langle \tau^2 \mid \omega \rangle P(\omega)\, \mathrm{d}\omega - \left[\int \langle \tau \mid \omega \rangle P(\omega)\, \mathrm{d}\omega \right]^2. \tag{5.3.23}$$

Cohen and Lee (1990) have shown that this may be put in the alternative form

$$\sigma_\tau^2 = \int \sigma_\tau^2(\omega) P(\omega)\, \mathrm{d}\omega + \int [\langle \tau \mid \omega \rangle - \langle \tau \rangle]^2\, P(\omega)\, \mathrm{d}\omega, \tag{5.3.24}$$

i.e. the delay spread squared is given by the mean of the local delay spread squared plus the variance of the local mean delay time about the global mean delay time.

The delay spread is independent of the values of $k'(0)$ and $k''(0)$. (It has already been assumed that $k(0) = 1$.) The mean time is readily found as independent of $k'(0)$ and is equal to the mean group delay. The delay spread squared follows as (Poletti and Vaughan, 1992)

$$\sigma_\tau^2 = \int \left(\left(\frac{A'(\omega)}{A(\omega)} \right)^2 + \left(\phi'(\omega) \right)^2 \right) A^2(\omega)\, \mathrm{d}\omega - \left(\int \phi'(\omega)\, A^2(\omega)\, \mathrm{d}\omega \right)^2. \tag{5.3.25}$$

From a time–frequency viewpoint, the delay spread is a robust measure of dispersion because it is independent of $k'(0)$ and $k''(0)$.

5.3.2.7 Interpretation of local delay spread

A difficulty of time–frequency theory lies with the local delay variance, $\sigma_\tau^2(\omega)$, i.e. the local delay spread squared, in that it can be negative depending on the choice of k. This fundamentally faults the current state of time–frequency theory in the sense that the theory pursues what is intuitively correct. However, some insight for finding an expression for the local dispersion can still be gained while avoiding the inappropriate aspects of time–frequency theory. Some examples of different kernels illustrate the situation.

5.3.2.8 Example 1: $k = 1$

This is the Wigner (1932) distribution, which has the local mean time equal to the group delay, and the local dispersion is

$$\sigma_\tau^2(\omega) = \frac{1}{2} \left[\frac{A'(\omega)}{A(\omega)} \right]^2 - \frac{1}{2} \left[\frac{A''(\omega)}{A(\omega)} \right], \tag{5.3.26}$$

which can be negative (Claasen and Mecklenbraauker, 1980; Cohen, 1989).

5.3.2.9 Example 2: $k'(0) = 0$ and $k''(0) = 1/4$

This kernel removes the second term in the Wigner local variance (Cohen and Lee, 1988), and so a positive expression for the local delay spread squared is obtained as

$$\sigma_\tau^2(\omega) = \left[\frac{A'(\omega)}{A(\omega)}\right]^2. \tag{5.3.27}$$

This local delay spread can be related to single term local expansion results given below. The local mean delay time is again the group delay. Cohen and Lee's global variance equation (here for the delay spread squared) then takes the form

$$\sigma_\tau^2 = \int \left(A'(\omega)\right)^2 \, \mathrm{d}\omega + \int \left(-\phi'(\omega) - \langle\tau\rangle\right)^2 A^2(\omega) \, \mathrm{d}\omega. \tag{5.3.28}$$

5.3.2.10 Example 3: $k'(0) = -1/2$ and $k''(0) = 1/4$

This choice equates the local delay spread squared with the phase curvature (Poletti, 1993),

$$\sigma_\tau^2(\omega) = \phi''(\omega), \tag{5.3.29}$$

and is closely related to the Rihaczek (1968) time which is proportional to the absolute value of the phase curvature (see below). In this choice a linear phase produces no dispersion, but phase curvature can be negative, which again gives negative dispersion. However, if the poles and zeros in the transfer function are reasonably isolated, (so that there is locally a Hilbert transform relationship between log magnitude and phase derivatives (Poletti and Vaughan, 1990)), then the local dispersion from *Example 2* tends to produce maximum values when the phase curvature is large. In such a case the two definitions for *Example 2* (using the log magnitude derivative) and *Example 3* (using the phase double derivative) produce similar results.

The local mean time for this example is

$$\langle\tau \mid \omega\rangle = -\left(\phi'(\omega) + \frac{A'(\omega)}{A(\omega)}\right). \tag{5.3.30}$$

The association of this purely real expression for the mean time with the Rihaczek bandwidth is used below as a criterion for suggesting them as dispersion measures appropriate for the frequency selective fading channel.

The delay spread squared for this example is (Poletti and Vaughan, 1992)

$$\sigma_\tau^2 = \int \phi''(\omega) A^2(\omega) \, \mathrm{d}\omega + \int \left[-\left(\phi'(\omega) + \frac{A'(\omega)}{A(\omega)}\right) - \langle\tau\rangle^2\right]^2 A^2(\omega) \, \mathrm{d}\omega. \tag{5.3.31}$$

A simple example to demonstrate the different results is a flat channel (all-pass filter). This has a constant magnitude frequency response, and so the local dispersion according to *Example 2* is zero. However the dispersion according to *Example 3* is given by the phase curvature, $\phi''(\omega)$, which is therefore more appropriate for the flat channel case. The use of $\left|\phi''\right|$ as a dispersion measure is addressed further below.

5.3.2.11 Summary

There are multiple different definitions of dispersion available from time–frequency theory. The local dispersion is the delay spread squared at a given frequency, and the global dispersion is the mean delay spread squared. The choice of local dispersion relation is arbitrary, but some choices can be more illuminating than others for particular types of system. The local dispersion given in *Example 2*, $(A(\omega)'/A(\omega))^2$, is the only form that is always positive. However, it is unsuitable for dispersive channels that may be also flat (constant magnitude), for example. The local delay spread can be formulated to contain explicit separate phase and amplitude contributions. These two contributions may be referred to as the magnitude and phase dispersion components, as pointed out by Ville (1948). In a frequency-selective fading channel, this form of delay spread is therefore a more revealing description of the dispersion than those based on the group delay only, and also includes the mean time, which in Ville's equation, is zero. Sketches of the dispersion components for the two-path channel are given below.

5.3.3 Local approximation of transfer function

Further insight into dispersion is available from local expansion (Poletti and Vaughan, 1992; Poletti, 1993) of the channel transfer function and from comparison of the results with those from the time–frequency theory.

5.3.3.1 Single term expansion

The magnitude variation may be included by expanding the log magnitude as well as the phase of the transfer function (Poletti and Vaughan, 1992), i.e.

$$H(\omega) \approx H(\omega_c)\mathrm{e}^{[A'(\omega_c)/A(\omega_c)+\mathrm{j}\phi'(\omega_c)][\omega-\omega_c]}. \tag{5.3.32}$$

The output of the channel for a narrowband input signal $x(t)\mathrm{e}^{\mathrm{j}\omega_c t} \Leftrightarrow X(\omega-\omega_c)$ is then given by (reverting to the conventional electrical engineering Fourier transform definition (Papoulis, 1962)),

$$y(t) = H(\omega_c)\frac{1}{2\pi}\int X(\omega-\omega_c)\mathrm{e}^{[A'(\omega_c)/A(\omega_c)+\mathrm{j}\phi'(\omega_c)][\omega-\omega_c]}\mathrm{e}^{\mathrm{j}\omega t}\mathrm{d}\omega. \tag{5.3.33}$$

The use of t here is just to differentiate between the signal time t and the channel delay time τ. The demodulated signal $y(t)\exp(-\mathrm{j}\omega_c t)$ follows as (Poletti and Vaughan, 1992)

$$\begin{aligned} y(t)\,\mathrm{e}^{-\mathrm{j}\omega_c t} &= H(\omega_c)\frac{1}{2\pi}\int X(\omega)\mathrm{e}^{\mathrm{j}[t+\phi'(\omega_c)-\mathrm{j}A'(\omega_c)/A(\omega_c)]\omega}\mathrm{d}\omega \\ &= H(\omega_c)\,x\left(t-\left[-\phi'(\omega_c)+\mathrm{j}\frac{A'(\omega_c)}{A(\omega_c)}\right]\right), \end{aligned} \tag{5.3.34}$$

i.e. the distortion caused by the channel can be viewed as the signal $x(t)$ being scaled by $H(\omega_c)$ and delayed in complex time by $-\phi'(\omega)+\mathrm{j}A'(\omega)/A(\omega)$, indicated in

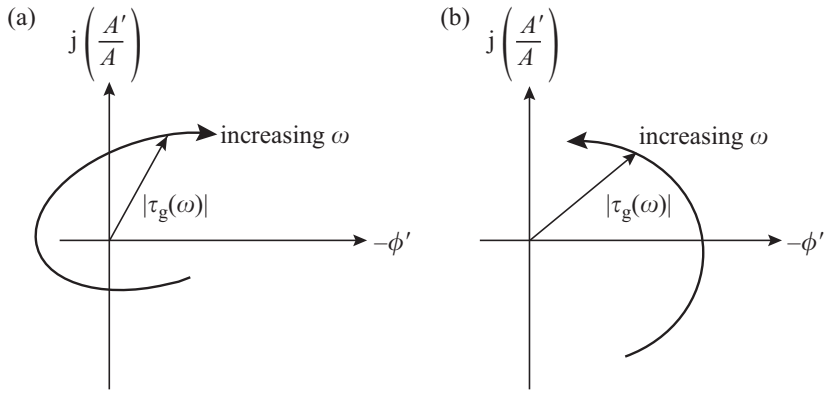


Figure 5.3.1 The locus of the complex delay for (a) a minimum phase fade and (b) a non-minimum fade

Figure 5.3.1. When $A'/A = 0$, the distortion is zero and there is pure (real) delay. For a minimum phase fade, two frequencies about the fade centre frequency will see this condition. For a non-minimum phase fade, this may not be the case as indicated in the figure where the locus is away from the $\mathrm{j}(A'/A)$ axis.

The above form for the delay suggests that a *complex group delay* may be defined this way, i.e. as the log derivative of the transfer function (cf. Rihaczek, 1968),

$$\tau_g(\omega) = -\phi'(\omega) + \mathrm{j}\frac{A'(\omega)}{A(\omega)}, \quad \left|\tau_g(\omega)\right| = \sqrt{\left(\phi'(\omega)\right)^2 + \left(\frac{A'(\omega)}{A(\omega)}\right)^2}, \tag{5.3.35}$$

which can be likened to the makeup local mean time of *Example 3*, namely, the purely real $\tau_m(\omega) = -\phi'(\omega) - A'(\omega)/A(\omega)$.

The complex group delay is the equivalent of the *complex instantaneous frequency* of a signal $x(t) = A(t)\mathrm{e}^{\mathrm{j}\phi(t)}$, being $\omega_{\mathrm{inst}}(t) = A'(t)/A(t) + \mathrm{j}\phi'(t)$, as introduced by Hahn (1964).

5.3.3.2 Rihaczek bandwidth

An important form of local variance is one proportional to the absolute value of the phase curvature. The 'dynamic bandwidth', B_d, was defined by Rihaczek (1968) as the frequency range centred at ω over which the phase $\phi(\omega)$ is linear to within $\pi/4$. The quadratic part of the Taylor expansion for the phase, $(\omega - \omega_c)^2\phi''(\omega_c)/2$, is set to $\pi/4$ at the edges of the band, i.e. at $(\omega - \omega_c) = B_d/2$, which gives

$$B_d(\omega) = \sqrt{\frac{2\pi}{|\phi''(\omega)|}}. \tag{5.3.36}$$

Rihaczek's dynamic bandwidth therefore defines the region over which the group delay derivative is approximately constant. From this measure, a signal limited to the above bandwidth will experience a delay equal to the group delay, and experience little

dispersion. Signals centred at ω with a larger bandwidth, $B > B_\mathrm{d}$, will experience more dispersion.

Alternatively stated, signals centred at ω will be more dispersed if they have a duration which is less than a local delay spread of

$$\sigma_\mathrm{d}(\omega) = \frac{1}{B_\mathrm{d}}(\omega) = \sqrt{\frac{|\phi''(\omega)|}{2\pi}}. \tag{5.3.37}$$

Hence the absolute phase curvature gives a measure of the local dispersion of the system. This suits fading channels, since, although the magnitude (and thence the amplitude dispersion) is not explicitly included, the local magnitude and phase variations are related in the vicinity of the fades as noted above.

From *Example 3* above, whose local variance is related to the Rihaczek bandwidth, the local delay time is recalled as $\left(-\phi'(\omega) - A'(\omega)/A(\omega)\right)$. This local delay time and Rihaczek bandwidth thus form associated dispersion parameters which seem the most suitable for the frequency fading channel.

Finally, it is noted that other workers have taken a similar approach to Rihaczek. Wetzel (1967) also used the absolute phase curvature to discuss dispersion bandwidths. Anderson and Askne (1974) note that a Gaussian pulse $s(t) = \exp(-t^2/2\tau_0^2)$ propagating through a quadratic phase channel holds its Gaussian form, but its duration is stretched to $\tau^2 = \tau_0^2 + (\phi(\omega_0))^2/\tau_0^2$. Lin *et al.* (1989) discuss a Gaussian pulse $s(t) = \exp(-t^2/2\tau_0^2)$, which is stretched by a factor of a to be $s(t) = \exp(-t^2/2(a\tau_0)^2)$ by passing through a quadratic phase channel. They show that if the channel limits the (Gaussian) spectrum of the input pulse to where it has decayed to 1/e of its peak value at $\omega = \omega_0$, then that bandwidth is given by

$$\Delta\omega = \frac{b(a)}{\sqrt{2\pi}}\sqrt{\frac{2\pi}{|\phi''(\omega_0)|}}, \tag{5.3.38}$$

where

$$b(a) = \frac{\sqrt{2}\left(a^2 - 1\right)^{1/4}}{2\pi}. \tag{5.3.39}$$

As an example, for 10 per cent pulse stretching ($a = 1.1$), then $b = 0.152$.

5.3.3.3 Summary

As with the time–frequency results being dependent on an arbitrary choice of the kernel of the time–frequency distribution, a difficulty with these local expansion results is that they are dependent on the order of the expansion (Poletti and Vaughan, 1992; Poletti, 1993). Some of the results obtained from time–frequency analysis and local expansions can be related. Although the question of a rigorous general measure of dispersion remains unanswered, the examples quoted offer a feel for how specific definitions can suit specific situations. For fading channels, where the phase and magnitude are locally related (Hilbert transform between log magnitude and phase), the Rihaczek bandwidth has stood the test of time (–frequency) and seems a sound choice

as a simple dispersion measure. A mean delay time of $\left(-\phi'(\omega) - A'(\omega)/A(\omega)\right)$ accompanies an example of a time–frequency distribution having a variance proportional to the Rihaczek bandwidth and so is suitable for characterising a frequency fading channel.

5.3.4 Mean delay for a signal in a channel

Example 3 of the previous discussion has noted that, in general, the local (in frequency) mean time of a normalised signal

$$x(t) \quad \Leftrightarrow \quad X(\omega)\mathrm{e}^{\mathrm{j}\Phi(\omega)}, \tag{5.3.40}$$

can be defined in terms of its phase dispersion (group delay) and its amplitude dispersion, i.e.

$$\tau_{\mathrm{m}}^{(\mathrm{g})}(\omega) = \langle t \mid \omega \rangle_t = -\left(\Phi'(\omega) + \frac{X'(\omega)}{X(\omega)}\right), \tag{5.3.41}$$

which, when referred to a channel, can be called a *generalised group delay*. Again, t is used to indicate the signal quantity.

Treating the situation from a traditional group delay perspective, the mean time of a modulated pulse, $s(t) = x(t)\mathrm{e}^{\mathrm{j}\omega_{\mathrm{c}}t}$, after passing through a channel $|H(\omega)|\,\mathrm{e}^{\mathrm{j}\phi(\omega)}$, is (Poletti and Vaughan, 1992)

$$\tau_{\mathrm{g}} = \langle \tau_{\mathrm{g}}(\omega) \rangle_\omega = -\frac{\int \left(\Phi'(\omega - \omega_{\mathrm{c}}) + \phi'(\omega)\right) |H(\omega)|^2 \, X^2(\omega - \omega_{\mathrm{c}}) \, \mathrm{d}\omega}{\int |H(\omega)|^2 \, X^2(\omega - \omega_{\mathrm{c}}) \, \mathrm{d}\omega}, \tag{5.3.42}$$

which shows how the mean time of a signal which has passed through a channel depends on both the signal as well as the channel. Figure 5.3.2 depicts the signal and channel quantities.

To interpret the group delay of the channel, $-\phi'(\omega)$, as the difference in the mean times (from the *Example 3* definition) of the input and output signals, therefore requires some constraints of the signal itself. Specifically, if the signal pulse $x(t)$ is even and with linear phase, then the $\Phi'(\omega)$ term (belonging to the signal) drops out; if also the spectral magnitude of the signal passing through the channel is flat so that the amplitude term drops out, then the difference in the local mean times ('generalised

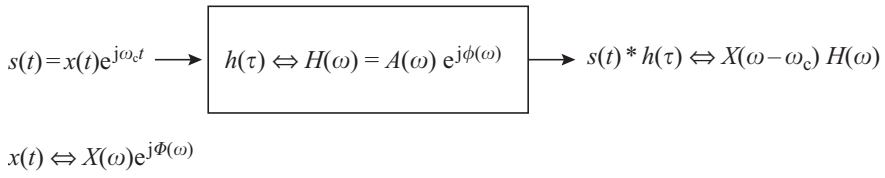


Figure 5.3.2 The signal and transfer function parameters for the dispersion relations. t denotes signal time and τ indicates the delay time of the channel impulse response

group delay') of the input and output signals is the same as the channel group delay. This situation was used in the sinc pulse example demonstrating the negative group delay in the previous section.

5.3.5 Dispersion in the two-path model

The basic dispersion quantities, brought in from the above discussion, for the two-path model, $h(\tau) = \delta(\tau) + a_2\delta(\tau - \tau_2) \Leftrightarrow A(\omega)\,\mathrm{e}^{\mathrm{j}\phi(\omega)}$, are:

$$(A(\omega))^2 = 1 + a_2^2 + 2\,a_2 \cos\omega\,\tau_2. \tag{5.3.43}$$

$$A'(\omega) = \frac{-a_2\tau_2 \sin(\omega\,\tau_2)}{A(\omega)}, \tag{5.3.44}$$

$$A''(\omega) = -\frac{a_2\tau_2^2\left(a_2\left(1+\cos^2\omega\,\tau_2\right) + \left(1+a_2^2\right)\cos\omega\,\tau_2\right)}{A^3(\omega)}, \tag{5.3.45}$$

$$\phi(\omega) = \tan^{-1}\left(\frac{-a_2 \sin\omega\,\tau_2}{1 + a_2\,\cos\omega\,\tau_2}\right), \tag{5.3.46}$$

$$\phi'(\omega) = \frac{-a_2\,\tau_2\,(a_2 + \cos\omega\,\tau_2)}{A^2(\omega)}, \tag{5.3.47}$$

$$\begin{aligned}\phi''(\omega) &= \frac{a_2\,\tau_2^2 \sin(\omega\,\tau_2)}{A^2(\omega)} - 2\frac{(a_2\,\tau_2)^2\;(a_2 + \cos\omega\,\tau_2)\sin\omega\,\tau_2}{A^4(\omega)} \\ &= \frac{a_2\,\tau_2^2\;\left(a_2^2 - 1\right)\;\sin\omega\tau_2}{\left(1+a_2^2\right)^2 + 4\,a_2\left(1 + a_2^2 + a_2\;\cos\omega\,\tau_2\right)\cos\omega\tau_2}\end{aligned} \tag{5.3.48}$$

The delay spread squared, from the general time–frequency result of equation 5.3.25, is

$$\begin{aligned}\sigma_\tau^2 &= \frac{\int\left(\left(A'(\omega)/A(\omega)\right)^2 + \left(\phi'(\omega)\right)^2\right)\;A^2(\omega)\,\mathrm{d}\omega}{\int P(\omega)\,\mathrm{d}\omega} - \left(\frac{\int\phi'(\omega)\;A^2(\omega)\,\mathrm{d}\omega}{\int P(\omega)\,\mathrm{d}\omega}\right)^2 \\ &= \left(\frac{a_2\;\tau_2}{1+a^2}\right)^2,\end{aligned} \tag{5.3.49}$$

which is the delay spread squared from Section 5.2. (The normalisation is required here since the basic expression used above for the two-path channel model is not normalised.)

The channel quantities are shown in Figure 5.3.3 for a single fade period of the two-path model. The magnitude quantities are the solid lines and the phase quantities are dotted. Plot (a) is the magnitude in dB and the minimum phase; (b) is the basic amplitude and phase terms in the dispersion, $A'(\omega)/A(\omega)$, $\phi'(\omega)$; (c) is the amplitude and phase dispersion contributions, $(A'(\omega)/A(\omega))^2$, $(\phi'(\omega))^2$; and

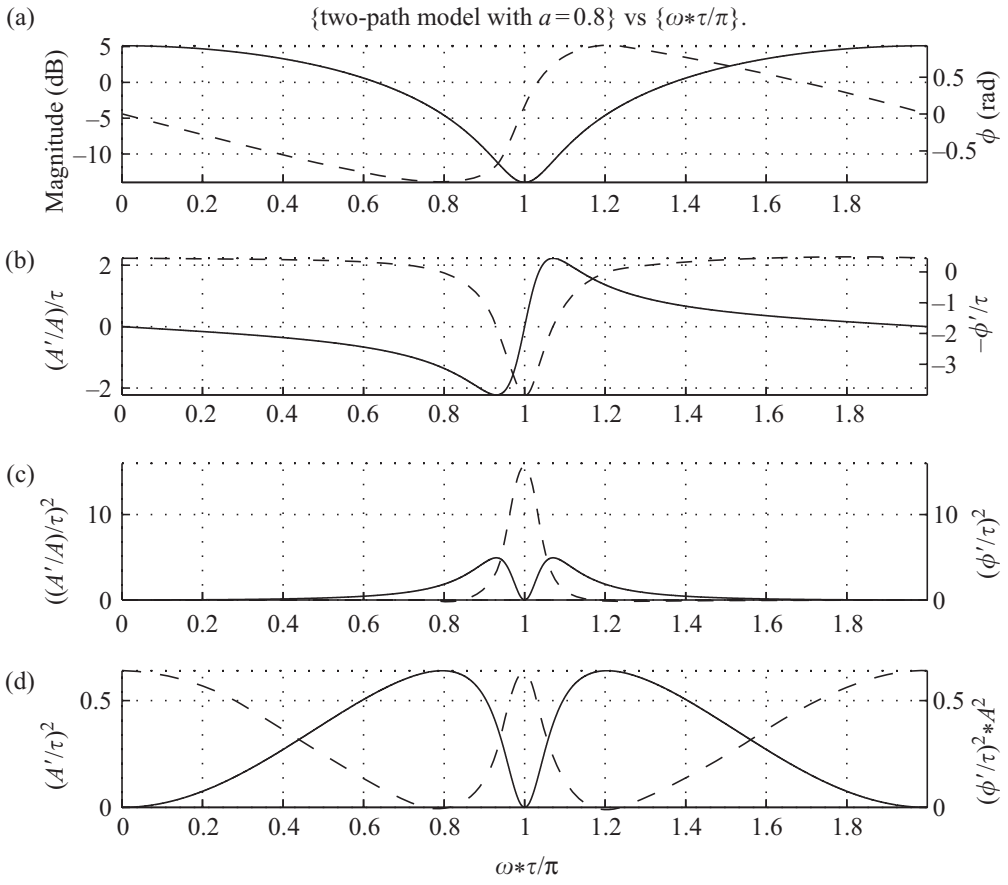


Figure 5.3.3 The channel parameters in the minimum phase two-path model. The solid curves are for the amplitude quantities (left-hand scale) and the phase terms are the dotted plots (right-hand scale). In (d), the integral of the sum of the two terms is the second time moment $\langle\tau^2\rangle$

(d) is the energy weighted amplitude and phase dispersion terms, $(A'(\omega)/A(\omega))^2 \cdot A^2(\omega)$, $(\phi'(\omega))^2 \cdot A^2(\omega)$.

In Figure 5.3.3 (c), the dispersion terms are seen to concentrate about the fade location, with the amplitude dispersion complementing the phase dispersion in the sense that they are peaking in different locations within the fade. The phase dispersion is the larger and peaks at the fade location. The complementary nature of the contributions is more obvious when the dispersions are weighted by the energy, in (d). Here, their sum, which is the integrand for the mean square delay, is constant,

$$\left(\frac{A'(\omega)}{A(\omega)}\right)^2 \cdot A^2(\omega) + \left(\phi'(\omega)\right)^2 \cdot A^2(\omega) = (a_2\tau_2)^2 \,. \tag{5.3.50}$$

The non-minimum phase case is given in Figure 5.3.4, where the value of a_2 is 1.25, i.e. the reciprocal of the 0.8 value used in the minimum phase case above. The magnitudes are therefore the same. The difference is that the phase dispersion

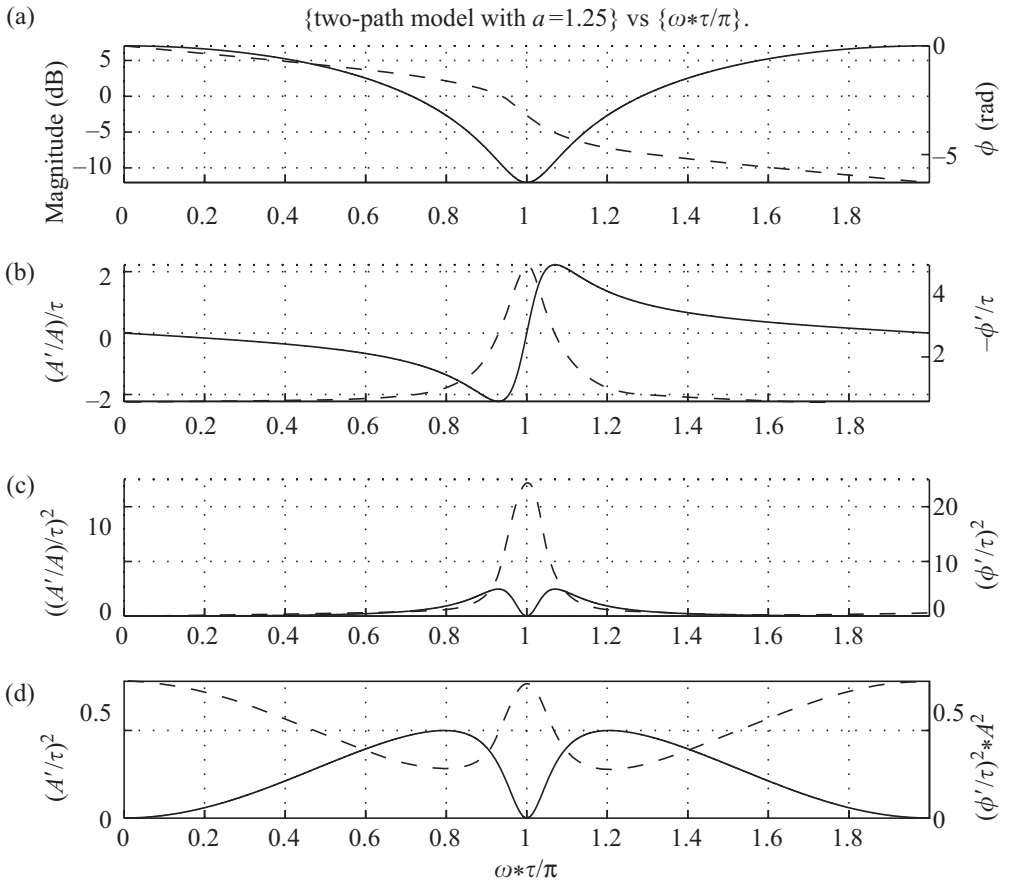


Figure 5.3.4 The channel parameters for the non-minimum phase channel, equivalent to that in Figure 5.3.3

is higher (by definition, the phase excursion is larger) and so its dominance over the amplitude dispersion (which is unchanged) is more pronounced.

The second derivative for the phase, $\phi''(\omega)$, is plotted in Figure 5.3.5(a), for the mimimum phase case, and (b) shows the Rihaczek bandwidth. The non-minimum phase case for $\phi''(\omega)$ has the same magnitude and form but is reversed in polarity, and the Rihaczek bandwidth, related to the absolute phase curvature, is the same for both cases. So the channel distortion, using the Rihaczek bandwidth as a measure, is the same for the minimum phase and the non-minimum phase cases, despite the phase trajectories being different.

The Rihaczek bandwidth has minima symmetrically located about the centre of the fade in the two-path transfer function, evident in the figure. These minima are located at

$$(\omega\tau_2)_{\min} = \Omega = \cos^{-1}\left(\frac{1+a_2^2}{4a_2} - \frac{1}{4a_2}\sqrt{1+34a^2+a^4}\right), \qquad (5.3.51)$$

i.e. at $\omega\tau_2 = (\pi - \Omega)$ either side of the fade position. For the $a_2 = 0.8$ example, which corresponds to a fade of about 16 dB below the mean level, the minima of

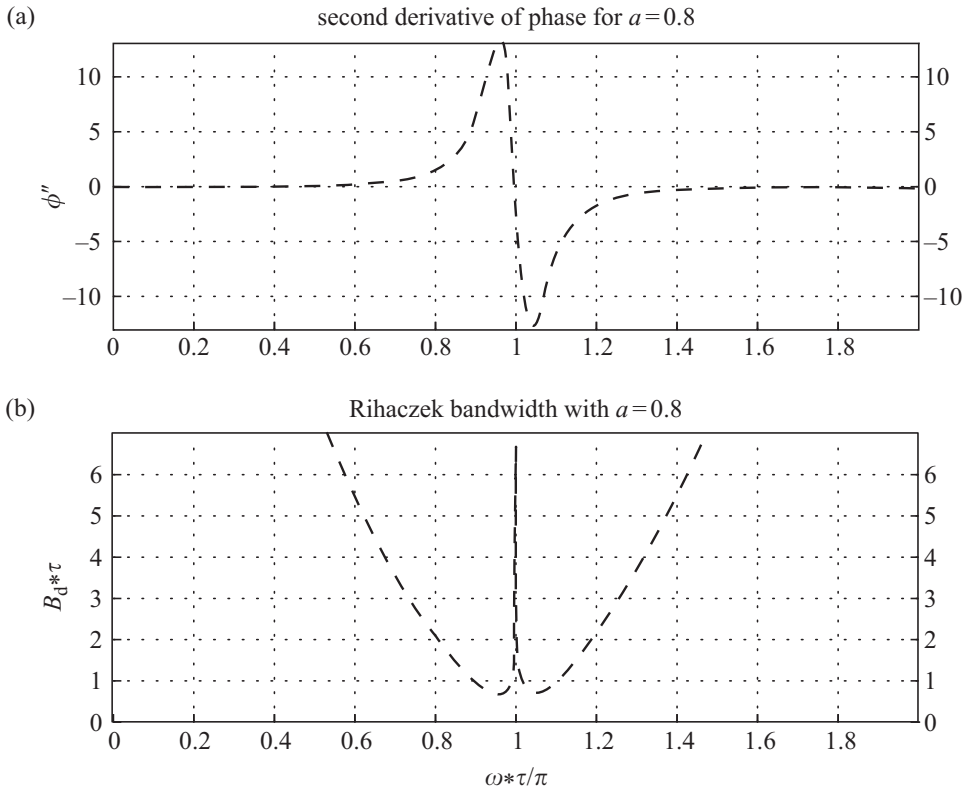


Figure 5.3.5 (a) The second phase derivative of the minimum phase case and (b) the Rihaczek bandwidth (same for both cases)

the Rihaczek bandwidth are located at about $\Omega = 0.04\pi$ from the fade frequency. The minima do not shift much with a_2, being at about 0.13π for $a_2 = 0.5$, and Ω approaching 0 for a_2 approaching 1.

The Rihaczek bandwidth can now be expressed as a function of the two-path parameter a_2,

$$B_{\mathrm{d}}(a_2) \cdot \tau_2 = \sqrt{\frac{2\pi\ \tau_2^2}{|\phi''\,(\Omega(a_2))|}} \tag{5.3.52}$$

and this is plotted in Figure 5.3.6.

The curve can be interpreted as the maximum allowable bandwidth for a signal to be 'undispersed' after passing through the two-path channel, as a function of a_2. This curve runs parallel to the extrema of the group delay in Figure 5.2.10. Using this measure of dispersion, signals should be confined to regions on the outside (away from the origin) of the dotted curves. 'Undispersed' here means distorted by less than some amount. Within the bandwidth B_{d}, the allowed deviation from linearity is $\pi/4$, but an acceptable dispersion depends on the signal itself and other communications considerations. So the value of B_{d} can be scaled accordingly, for example to be confined to be linear to within $\pi/8$ instead of $\pi/4$. This would reduce B_{d} and move the dotted curves of Figure 5.2.10 further from the origin.

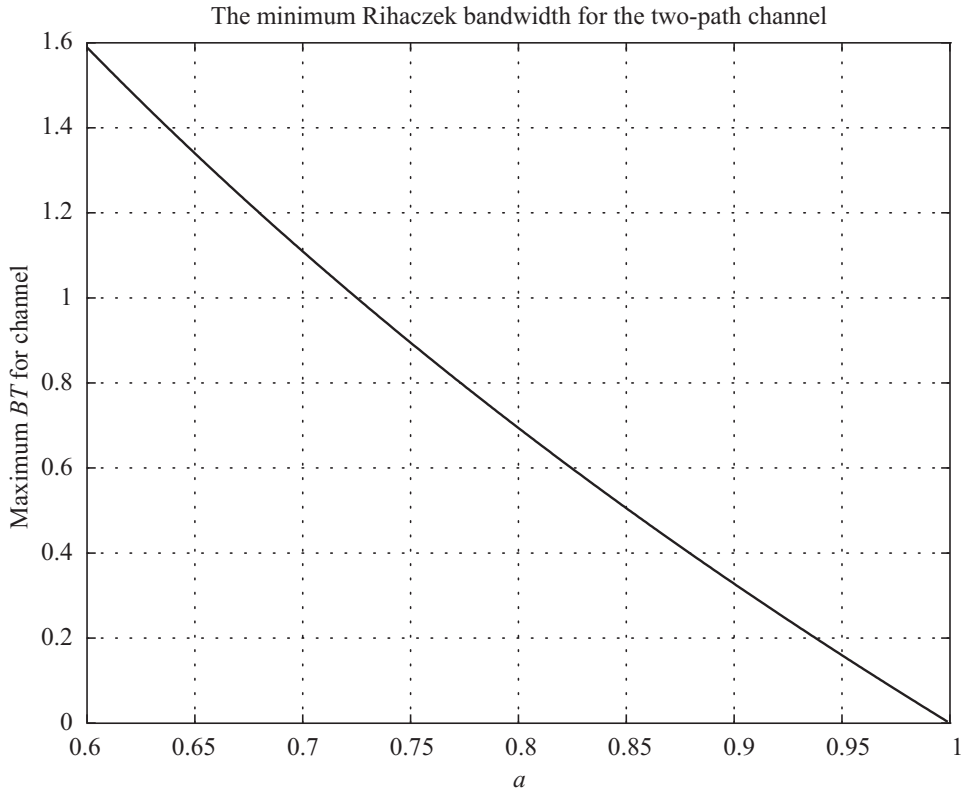


Figure 5.3.6 The minimum Rihaczek bandwidth which can be found in a given two-path model. This bandwidth can be interpreted, for the narrowband case, as the maximum signal bandwidth allowable for 'undispersed' signals emanating from the two-path channel with a given fade depth (a_2). The bandwidth is expressed in the diagram as a product with the notation $BT = B_d\tau_2$

5.3.5.1 Examples

A couple of examples illustrate the distortion for the two-path model. The input pulse to the two-path channel is a Gaussian envelope with modulation. The envelope has the same duration at the half-power points as that of the sinc pulse of bandwidth B used in the previous section. The effective bandwidth of the Gaussian pulses and the bandwidth of the sinc pulses are then of the same order. The signal carrier is at the fade frequency again, the region of the channel which invokes the most dispersion. The channel is $a_2 = 0.99$, corresponding to an extremely deep fade – some 40 db below the mean power level: see Figure 5.2.2. Figure 5.3.7 shows the input pulse and output pulses using three bandwidths: (a) $B\tau_2 = 0.005$, (b) $B\tau_2 = 0.03$, and (c) $B\tau_2 = 0.15$. These are chosen to correspond to (a) 'undispersed' (the signal band B occupying less than the Rihaczek bandwidth, i.e. $B < B_d$), (b) 'threshhold dispersed' ($B = B_d$), and (c) 'dispersed' ($B > B_d$). The operating points may be estimated in Figure 5.2.10 and Figure 5.3.6. The output pulse shows in: (a) an obvious negative group delay ($\sim 70\tau_2$), mild distortion with the envelope inverting for a small duration,

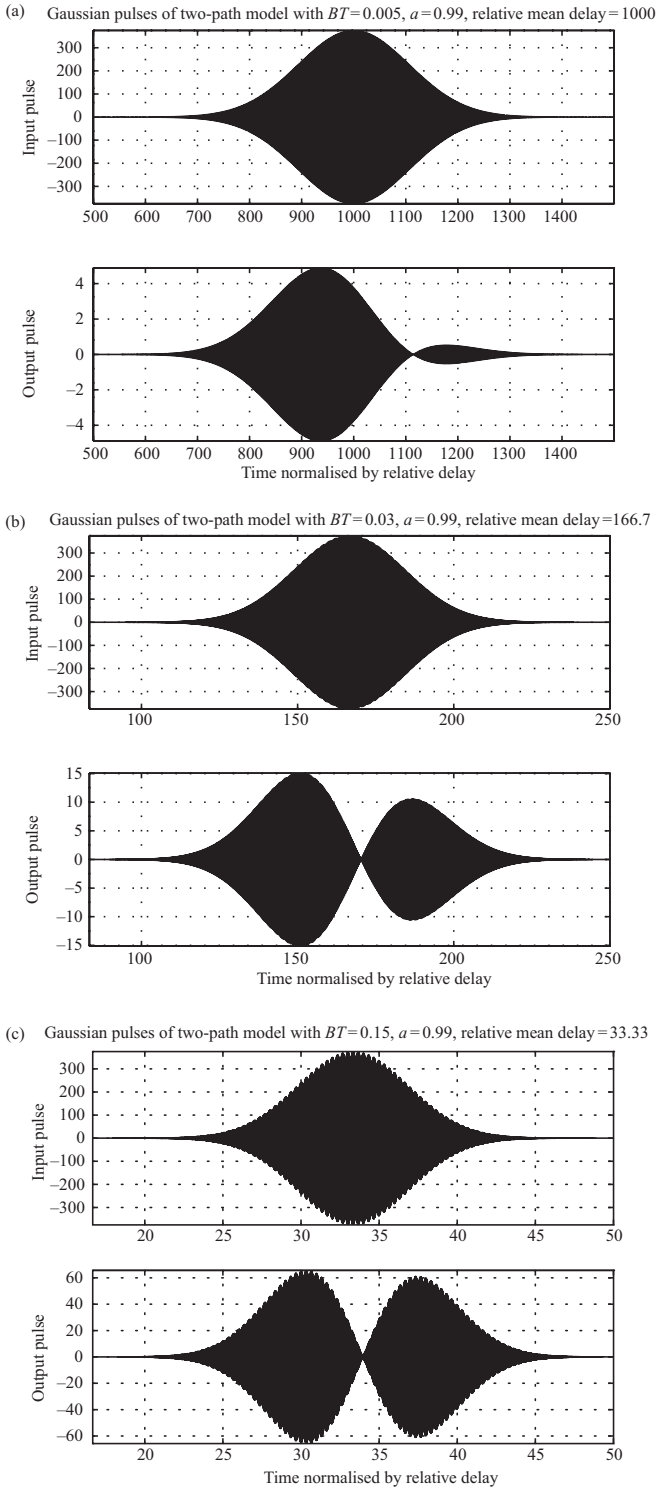


Figure 5.3.7 The Gaussian envelope input pulse and output pulse for a deep frequency-faded ($a_2 = 0.99$ corresponding to a 40 dB fade) two-path channel. (a) Sufficiently narrowband envelope for low dispersion, (b) envelope bandwidth similar to the Rihaczek bandwidth for threshhold dispersion, and (c) envelope bandwidth exceeding the Rihaczek bandwidth for high distortion

and gain ~ -35 dB; (b) a negative group delay but with serious distortion creating an element of ambiguity in the pulse position, gain ~ -25 dB; (c) a negative group delay with gross distortion creating ambiguity as to whether there is positive or negative group delay, and a gain ~ -12 dB. The gain is not the main issue for distortion, and indeed tends to be compensated by an AGC circuit normally used in receivers. Also, signals are usually angle modulated and so after signal limiting only the phase distortion is present.

The mechanism for potential reception timing problems caused by digitally modulated signals occupying the channel around the fades is that not only does the optimum symbol synchronisation position move from its undispersed position, but also the distortion due to the channel could cause a synchroniser to get confused. The distortion can appear as an ambiguity in position of the pulse position, a result of the envelope going negative at late time (or going negative at early time for positive group delays). In practice, timing problems may be aggravated by the signal having low signal-to-noise ratios (low channel gain) at the fades.

Finally, it is noted that the above examples were for the narrowband case (in the sense $B\tau_2 \ll 1$), where the delay difference τ_2 cannot be resolved to within the signal bandwidth B. For wider bandwidths ($B\tau_2 > 1$), such as those of Figure 5.2.10, the delay difference becomes resolvable, and the two paths of the channel convey signals which can effectively be separated at the receiver by simple signal processing, such as a rake-type equaliser. Each component of the received signal (one for each of the two paths) is just a scaled version of the original, and there is no distortion in either of them, in the context of the two-path channel model here. The Rihaczek bandwidth, dotted also in Figure 5.2.10, is thus not to be interpreted as a distortion limit in this sense, although mathematically it is the same dispersion measure as for the narrowband case.

5.3.6 Summary

In this section, time–frequency theory has been examined in the context of frequency selective channels. The formulation offers a Bello-like distribution but with several definitions of variance which are presented as the delay spread squared. Examples are given to demonstrate how different results suit different situations, such as the flat channel. Expansion of the transfer function also gives different possible results, depending on the order of the expansion. Compatible results from the two approaches provide a guide to a choice of dispersion metric which is suitable for channels with frequency selective fades. A local mean delay time of $\left(-\phi'(\omega) - A'(\omega)/A(\omega)\right)$ accompanies a local delay spread proportional to $\sqrt{|\phi''(\omega)|}$ which is from the reciprocal of the Rihaczek bandwidth, also known as the 'dynamic bandwidth'. These are suitable at a fade in the transfer function, which is where the greatest distortion is generated. This local delay spread measure is the same for minimum- and non-minimum-phase fades. Finally, examples of dispersed Gaussian pulses illustrate the amount of distortion resulting from 'undispersed' and 'dispersed' channels, i.e. channels that have a local delay spread which is respectively less or more than the duration of the communications symbol signals.

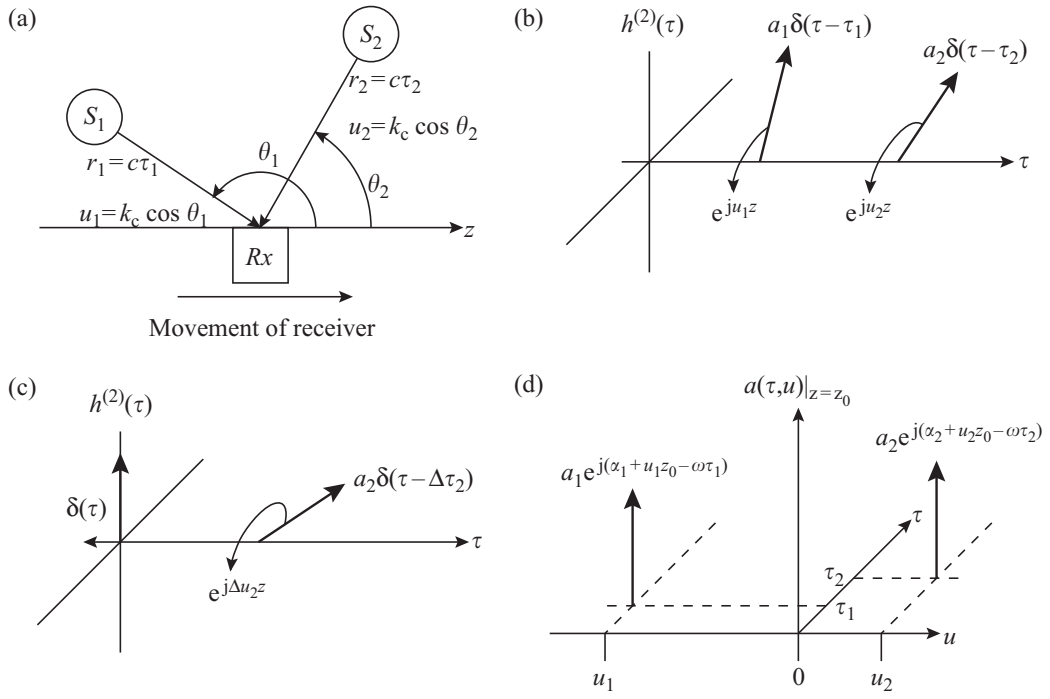


Figure 5.4.1 *(a) The scenario for a moving receiver in the presence of two scatterers. The receiver is in the far field of the individual scatterers and in the near field of an array comprising the scatterers as elements. The far field assumption means that the angles θ_1 and θ_2 are independent of distance z over the region of interest. (b) The impulse response in absolute delay time, (c) impulse response with the delay time and phase normalised to the first received path, (d) the effective scattering distribution at receiver position $z = z_0$*

5.4 Moving terminal in a static two-source scenario

A moving terminal introduces the terminal position as an argument into the transfer function of the channel. The scenario is otherwise the same as in the static terminal case, as shown in Figure 5.4.1(a). (In the figure, the delay of each effective scatterer is here taken as proportional to its distance from the receiver. However, the scatterers can have a common distance and their own delays, of course.) The transfer functions (b) and (c) and the effective scattering distribution (d) are also given.

5.4.1 Two-dimensional transfer function

The development of the two-dimensional transfer function is recalled here from Chapter 2. The distances to the scatterers, r_1 and r_2, under the assumption that the mobile is in the far field of the individual sources, change as

$$\begin{aligned} r_1(z) &\approx r_1(0) - z\ \cos\theta_1, \\ r_2(z) &\approx r_2(0) - z\ \cos\theta_2, \end{aligned} \tag{5.4.1}$$

Alternatively, by viewing the sources as a two-element array, the mobile can be considered as in the near field of such an array. So the transfer function can be viewed as the signal from the near field pattern of the two-source array comprising the two scatterers. Expressed as a two-dimensional scalar, the near field in the region of interest is written

$$E^{(2)}(\theta) = A_1 \mathrm{e}^{\mathrm{j}\alpha'_1}\, \mathrm{e}^{-\mathrm{j}(kr_1(0) - kz\cos\theta_1)} + A_2 \mathrm{e}^{\mathrm{j}\alpha'_2}\, \mathrm{e}^{-\mathrm{j}(kr_2(0) - kz\cos\theta_2)} \tag{5.4.2}$$

in which $A_1\mathrm{e}^{\mathrm{j}\alpha'_1}$ and $A_2\mathrm{e}^{\mathrm{j}\alpha'_2}$ represent the complex amplitudes of the incident (scalar) signals at the antenna. The prime on the alpha is used in order to distinguish between the phases of the physical scatterers (α') and the effective scatterers (α). The transfer function also depends on the far field pattern of the receiver, $h(\theta)$ (here as a scalar), and becomes, from equation (2.1.2),

$$\begin{aligned} H^{(2)}(\omega, z) = {} & h(\theta_1)\, A_1 \mathrm{e}^{\mathrm{j}\alpha'_1}\, \mathrm{e}^{-\mathrm{j}(kr_1 - kz\cos\theta_1)} \\ & + h(\theta_2)\, A_2 \mathrm{e}^{\mathrm{j}\alpha'_2}\, \mathrm{e}^{-\mathrm{j}(kr_2 - kz\cos\theta_2)}. \end{aligned} \tag{5.4.3}$$

Recall that the validity of this expression is based on the assumptions: that the receiver is in the far field of the scatterers (here the two scatterers can be viewed as the elements of the two-source array); that the locus of the receiver is sufficiently short and the scatterers' patterns are sufficiently non-directional to cause no amplitude change in the waves incident at the receiver positions; and finally that the range of frequencies is restricted such that the scatterers and receiving antenna have no frequency dependent properties. Finite bandwidth effects are addressed below.

The transfer function is now scaled, and with a change in terminology, written as (cf. equation (5.2.12))

$$H^{(2)}(\omega, z) = 1 + a_2 \mathrm{e}^{\mathrm{j}(\alpha_2 - k\Delta r_2 + \Delta u_2 z)} \tag{5.4.4}$$

where now 1 and $a_2\mathrm{e}^{\mathrm{j}\alpha_2}$ represent the scaled complex amplitudes of the effective scatterers through the substitutions,

$$\frac{A_2\,|h(\theta_2)|}{A_1\,|h(\theta_1)|} = a_2, \tag{5.4.5a}$$

$$\alpha'_2 - \alpha'_1 + \mathrm{phase}\{h(\theta_2)\} - \mathrm{phase}\{h(\theta_1)\} = \alpha_2. \tag{5.4.5b}$$

The difference in the ranges from the receiver to the two scatterers is denoted

$$\Delta r_2 = r_2 - r_1 \tag{5.4.6}$$

and the spatial Doppler frequency difference is

$$\Delta u_2 = k_\mathrm{C}\,(\cos\theta_2 - \cos\theta_1)\,. \tag{5.4.7}$$

The frequency and space dependence is then made explicit by writing the function as

$$H^{(2)}(\omega, z) = 1 + a_2 \mathrm{e}^{\mathrm{j}(\alpha_2 - \omega\Delta\tau_2 + \Delta u_2 z)} \tag{5.4.8}$$

where

$$\Delta\tau_2 = \frac{\Delta r_2}{c} \tag{5.4.9}$$

is the difference in the time delays of the signals from the two scatterers only if the scattered signals are launched simultaneously (i.e. the two scatterers are excited simultaneously in signal time).

The impulse response is shown in Figure 5.4.1(b). This can be scaled and shifted as

$$h^{(2)}(\tau, z) = \delta(\tau) + a_2 e^{j\alpha_2}\delta(\tau - \Delta\tau_2)e^{j\Delta u_2 z}. \tag{5.4.10}$$

This is the same as in the static two-path, except that now the second path, relative to the first, is rotating with distance z at a rate given by the Doppler frequency difference Δu_2. This is shown in Figure 5.4.1(c).

The power transfer function is

$$\left|H^{(2)}(\omega, z)\right|^2 = 1 + a_2^2 + 2a_2\cos(\alpha_2 + \omega\Delta\tau_2 - \Delta u_2 z), \tag{5.4.11}$$

which is the two-dimensional form of equation (5.2.13).

The variation of the frequency dependent fading is the same as that of the static terminal (one-dimensional channel), namely the fades occur regularly, at

$$\omega_n = \frac{1}{\Delta\tau_2}\big((2n-1)\pi - \alpha_2 + \Delta u_2 z\big) \quad n = 0, \pm 1, \ldots; \tag{5.4.12}$$

and so the frequency spacing of the fades is governed by $\Delta\tau_2$. In the general case, recall that the frequency dependency is found from the autocorrelation in frequency, which is in turn given by the Fourier transform of the power delay profile (see Appendix B),

$$P(\tau) \quad \Leftrightarrow \quad R_H(\omega). \tag{5.4.13}$$

So for the two-path case here, with $P(\tau) = \left|h^{(2)}(\tau)\right|^2$, the periodicity in angular frequency is $2\pi/\tau_2$. Similarly, the spatial fades are at

$$z_n = \frac{1}{\Delta u_2}\big((2n-1) - \alpha_2 + \Delta\tau_2\omega\big), \quad n = 0, \pm 1, \ldots; \tag{5.4.14}$$

so that the difference in the cosine angles (i.e. the difference in the spatial Dopper frequencies of the two sources, Δu_2) governs the spatial periodicity, with the fade positions depending on frequency and the phase difference between the scatterers. Again, the general case (Appendix B) is given by

$$P(u) \quad \Leftrightarrow \quad R_H(z). \tag{5.4.15}$$

Here we have a two-path scenario,

$$P^{(2)}(u) = \delta(u - u_1) + a_2^2\delta(u - u_2) \tag{5.4.16}$$

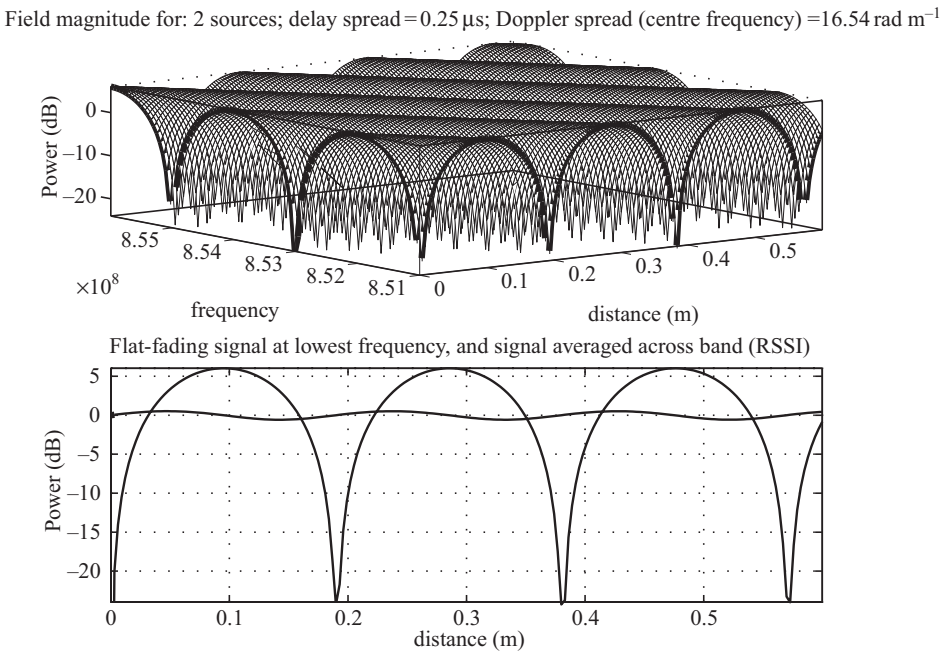


Figure 5.4.2 *(a) An example of a two-path channel power transfer function depicted in space and frequency. The fading is regular with a period in frequency governed by the delay spread and the period in space (or time, since the terminal is moving) given by the spatial Doppler spread. (b) A cut of the transfer function, showing the spatial fading at a single frequency and the much decreased fading of the total power in the band*

whose Fourier transform gives the distance periodicity. A plot showing the behaviour is given in Figure 5.4.2. The scenario comprises: $a_1 = a_2$; $\alpha_2 = 0$; $\theta_1 = 30°$, $\theta_2 = 170°$; $\Rightarrow \lambda \Delta u_2 = 11.63\,\text{rad}\,\lambda^{-1} \Rightarrow \Delta u_2 = 32.9\,\text{rad}\,\text{m}^{-1}$ at 851 MHz; ($\Rightarrow$ spatial fading separation of $2\pi/\Delta u_2 = 0.58 = 0.19$ m at 851 MHz); $\Delta r_2 = 150\,\text{m} \Rightarrow$ $\Delta \tau_2 = 0.5\,\mu\text{s}$ ($\Rightarrow$ frequency fading separation of $1/\Delta\tau_2 = 2$ MHz). The mean delay is 5.25 μs and the delay spread is 0.25 μs.

5.4.2 *Effect of finite bandwidth on the two-dimensional power transfer function*

The same results as for the static case apply to the moving receiver case. The power transfer function of the two-path channel becomes

$$\gamma_B(z) = \left|H_B^{(2)}(z)\right|^2 = \frac{1}{B}\int_{-B/2}^{B/2} \left|H^{(2)}(\omega, z)\right|^2 \,\mathrm{d}\omega$$
$$= 1 + a_2^2 + 2a_2 \frac{\sin(B\tau_2/2)}{B\tau_2/2} \cos(u_2 z + \alpha_2)\,. \qquad (5.4.17)$$

In a deep fade at z_0, $a_2 \approx 1$, i.e. $u_2 z_0 + \alpha_2 \approx \pi$; the energy is a minimum,

$$\gamma_B(z_0) \approx \frac{1}{3}\left(\frac{B\tau_2}{2}\right)^2, \quad B\tau_2 \ll 1; \tag{5.4.18}$$

where the latter expansion is truncated for powers of $(B\tau_2)^4$, and is the same result as in Section 5.2 for the static terminal. Again, the important interpretation is that the power transfer function does not go to zero for a non-zero bandwidth channel.

5.4.3 Doppler frequency moments

With the moving receiver, the Doppler effect comes into play. It is recalled that in the model, the Doppler frequency is not dependent on the position of the mobile. Analogously to the delay spread, the Doppler spread is the standard deviation of the spatial Doppler frequency distribution. For the scatterers being at a fixed bearing relative to the mobile (physically, the sources are far away relative to the trajectory of the receiver), the Doppler power profile requires no averaging, and is

$$P(u) = \left\langle |a(u)|^2 \right\rangle = |a(u)|^2. \tag{5.4.19}$$

The Doppler moments are

$$\langle u \rangle = \frac{\int u P(u)\,\mathrm{d}u}{\int P(u)\,\mathrm{d}u} = \frac{\int u|a(u)|^2\,\mathrm{d}u}{\int |a(u)|^2\,\mathrm{d}u} \tag{5.4.20}$$

and with

$$\langle u^2 \rangle = \frac{\int u^2 P(u)\,\mathrm{d}u}{\int P(u)\,\mathrm{d}u} = \frac{\int u^2|a(u)|^2\,\mathrm{d}u}{\int |a(u)|^2\,\mathrm{d}u} \tag{5.4.21}$$

the *Doppler spread* has, by analogy, identical form to the delay spread for the two-path model in equation (5.2.10), namely

$$s_u^{(2)} = \frac{a_2 \Delta u_2}{1 + a_2^2}. \tag{5.4.22}$$

In the example above, the spatial Doppler spread is $u_2/2 = 16.5\,\mathrm{rad\,m^{-1}}$ at 1 GHz. Again, strictly speaking these should be called the instantaneous moments since no averaging was used, although, as with the delay spread, the Doppler spread is the same as the instantaneous Doppler spread in a static scenario. Mathematically, the Doppler power profile is

$$P(u) = \langle P(u, \tau) \rangle_\tau = \langle |a(u, \tau)|^2 \rangle_\tau \tag{5.4.23}$$

stating that the averaging is over the delay time variable, analogously to the power delay profile defined as the average over distance (namely, the phase mixing process). In practical measurements, however, the averaging for a Doppler power profile can be over time or position instead of delay time, and averaging details should be documented with measurement results.

5.5 Statistics of the 'few-path' model

When there are multiple paths in a channel, there is a problem with describing the situation when there are too many paths to warrant a deterministic description but too few paths to warrant a Gaussian result from the central limit theorem. As we move from two paths to Gaussian, the pdfs of the signals change markedly. This transition region is referred to as the 'few-path model'. The three-path model is briefly discussed before the few-path statistics, which cannot be accurately described by Gaussian-derived results.

5.5.1 Three-path model

The three-path model is tractable, although somewhat unwieldy. For example, in the static case, the transfer function can be written directly from the summation as

$$|H^{(3)}(\omega)|^2 = 1 + a_2^2 + a_3^2 + 2a_2\cos(\omega\tau_2) + 2a_3\cos(\omega\tau_3) + 2a_2a_3\cos(\omega(\tau_3 - \tau_2)) \tag{5.5.1}$$

where the scaling is for $a_1 = 1$. The coherence bandwidth, $\Omega_c^{(3)}$ rad s^{-1}, must be expressed transcendentally as

$$\left|C_c^{(3)}\right|^2 = \frac{1 + a_2^2 + a_3^2 + 2a_2^2\cos\left(\Omega_c^{(3)}\tau_2\right) + 2a_3^2\cos\left(\Omega_c^{(3)}\tau_3\right) + 2a_2^2a_3^2\cos\left(\Omega_c^{(3)}(\tau_3 - \tau_2)\right)}{1 + a_2^4 + a_3^4 + 2\left(a_2^2 + a_3^2 + a_2^2a_3^2\right)}. \tag{5.5.2}$$

The three-path instantaneous delay spread can be found directly,

$$s^{(3)} = \frac{\sqrt{(a_2a_3(\tau_3 - \tau_2))^2 + (a_2\tau_2)^2 + (a_3\tau_3)^2}}{1 + a_2^2 + a_3^2}, \tag{5.5.3}$$

and the three-path spatial Doppler spread has identical form. Often, two or more of the three-path parameters are made dependent in order to allow further progress with closed form channel characterisation.

5.5.1.1 Application

The three-path model is considered suitable for point-to-point links with atmospheric multipath where more than three paths (direct, ground bounce, and more than one refracted ray – see Figure 5.5.1 is considered unusual (Parl, 1983). As a consequence, the probability functions have been popular research topics in connection with outage in microwave point-to-point links (e.g. Rummler, 1979; Sakagami and Hosoya, 1982; Parl, 1983). For mobile communications, the three-path model seems to offer little new insight over the two-path model to the workings of the real world (many-path) channel, and is not accurate enough, particularly at the tails of the distributions, to simulate properly the many-path channel.

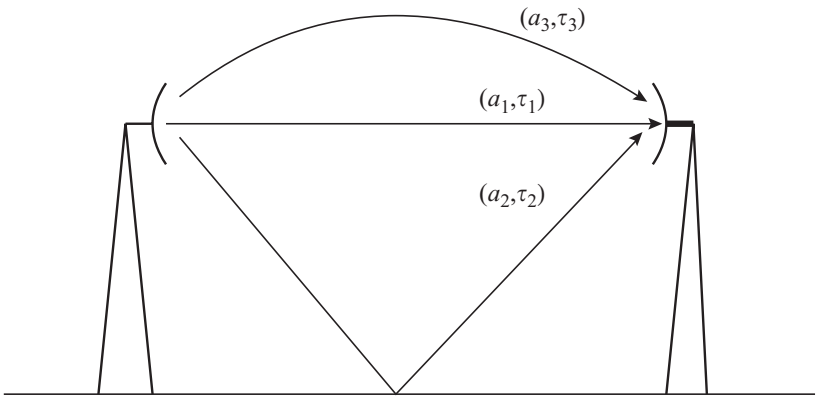


Figure 5.5.1 The three-path model is typically for point-to-point links where there is a direct, single ground bounce and refraction path

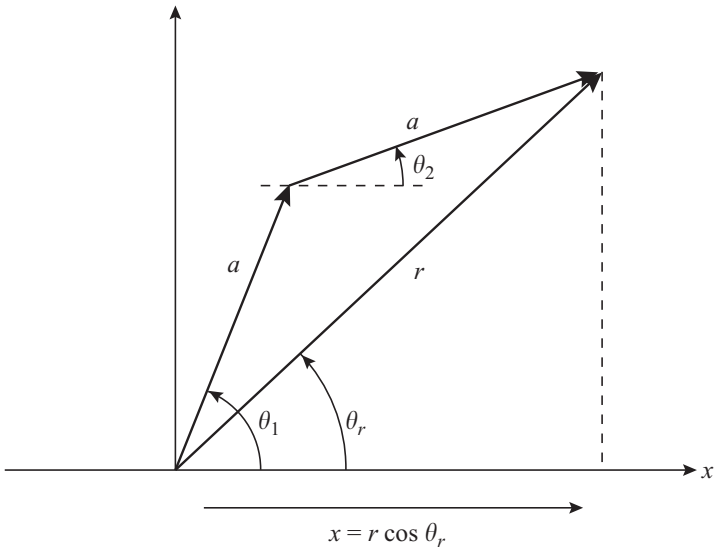


Figure 5.5.2 The summation of $n = 2$ vectors of length a and random angle θ_n to give a resultant vector of envelope r and in-phase component x

5.5.2 The few-path model

As more paths are added to the two-path model, the statistics of the deterministic channel move from the closed form to those of the many-path, i.e. uniform phase and Rayleigh envelope. During this transition (the 'few-path' channel), the probability distributions are complicated and need to be evaluated on an individual basis.

5.5.2.1 Equal amplitudes

The mathematical situation is given in Figure 5.5.2, where only $n = 2$ paths are depicted. The angles of the vectors are uniformly distributed from 0 to 2π. The

resultant vector length is the envelope r and its projection on the abscissa (the in-phase component) is $x = r\cos\theta_r$, where θ_r is the angle of the resultant. The projection onto the ordinate (the quadrature component) has the same statistics as x. The largest value of both r and $\pm x$, is na.

5.5.2.2 Power of in-phase and quadrature components

The simplest case is to consider the sum of n independent, normally distributed random variables with zero mean and variance σ^2. The pdf of the components is denoted $\mathcal{N}(0, \sigma^2)$. The sample mean is denoted υ; its expected value is zero and its variance is $\sigma_v^2 = 2\sigma^4/n$. The quantity $\eta = \sum_{i=1}^{n} (x_i/\sigma)^2$ is always positive, and its distribution is chi-squared with n degrees of freedom. This corresponds to the power in the resultant in-phase or resultant quadrature components. The chi-squared pdf with n degrees of freedom (see Appendix C) is

$$p_\eta(\eta) = \frac{1}{2^{n/2}\,\Gamma(n/2)}\eta^{n/2-1}\,\mathrm{e}^{-\eta/2}, \quad \eta > 0, \tag{5.5.4}$$

and is shown in Figure 5.5.3, demonstrating the rapidly changing situation with the number of path components for the few-path model. So the statistics, including the cumulative probability, confidence intervals (Papoulis, 1962), etc., for the

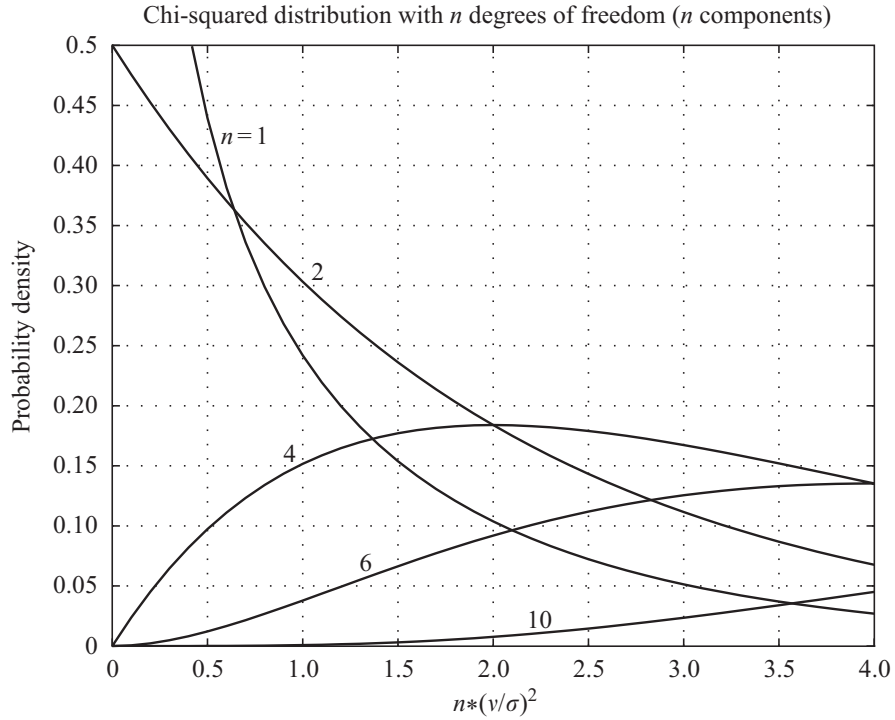


Figure 5.5.3 The chi-squared pdf with n degrees of freedom corresponds to the power of the resultant in-phase and quadrature signals when the n components are normal, zero mean, independent random variables with variance σ^2

power of the resultant in-phase and quadrature components can be readily calculated. This power is averaged over many different realisations of the random amplitude and phase for the given number of components. The power of the envelope is simply the sum of the powers of the components, i.e. has a chi-squared distribution with $2n$ degrees of freedom.

The special case for equal and constant amplitudes and random phases is reviewed below because of its importance in signal simulations where common amplitudes are often used for the random sources. The case would not normally arise with real-world propagation signals, but may arise when combining weighted (to make the signals equal amplitude), random signals in signal-processing stages. Further statistical reference results for simulations are in Section 7.2. The results below are drawn from Slack (1946) and the parameter a represents the common amplitude of the components, and can be set to one to simplify the formulations.

For n contributions with equal amplitudes and random phases, the pdf of the amplitude of the resultant in-phase and quadrature components is

$$p_{x^{(n)}}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (J_0(a\beta))^n \cos \beta x \, \mathrm{d}\beta, \tag{5.5.5}$$

i.e. the cosine Fourier transform of $(J_0(a\beta))^n$. The one-sided form, for $|x|$, is

$$p_{x^{(n)}}(|x|) = \frac{1}{\pi} \int_{0}^{\infty} (J_0(a\beta))^n \cos(\beta x) \, \mathrm{d}\beta. \tag{5.5.6}$$

This is shown in Figure 5.5.4. As n becomes large, the pdf approaches Gaussian with variance $na^2/2$, i.e.

$$p_{x^{(n)}}(x) \approx \frac{1}{a\sqrt{n\pi}} \mathrm{e}^{-x^2/(na^2)}, \quad n \text{ large}. \tag{5.5.7}$$

For $n > \sim 10$, the Gaussian distribution is a very good approximation except at the extremes of the tails, where the finite-n distributions are bounded. Figure 5.5.4 includes the case $n = 8$, which cannot be distinguished from the Gaussian case on the scale used. In the figure, the resultant signal is normalised by the standard deviation $a\sqrt{n/2}$ and the distribution function is scaled correspondingly. The $n = 1$ case is the familiar 'bathtub' curve

$$p_{x^{(1)}}(|x|) = \frac{1}{\pi\sqrt{a - x^2}}. \tag{5.5.8}$$

The figure shows that the form of the pdfs changes drastically for the small values of n.

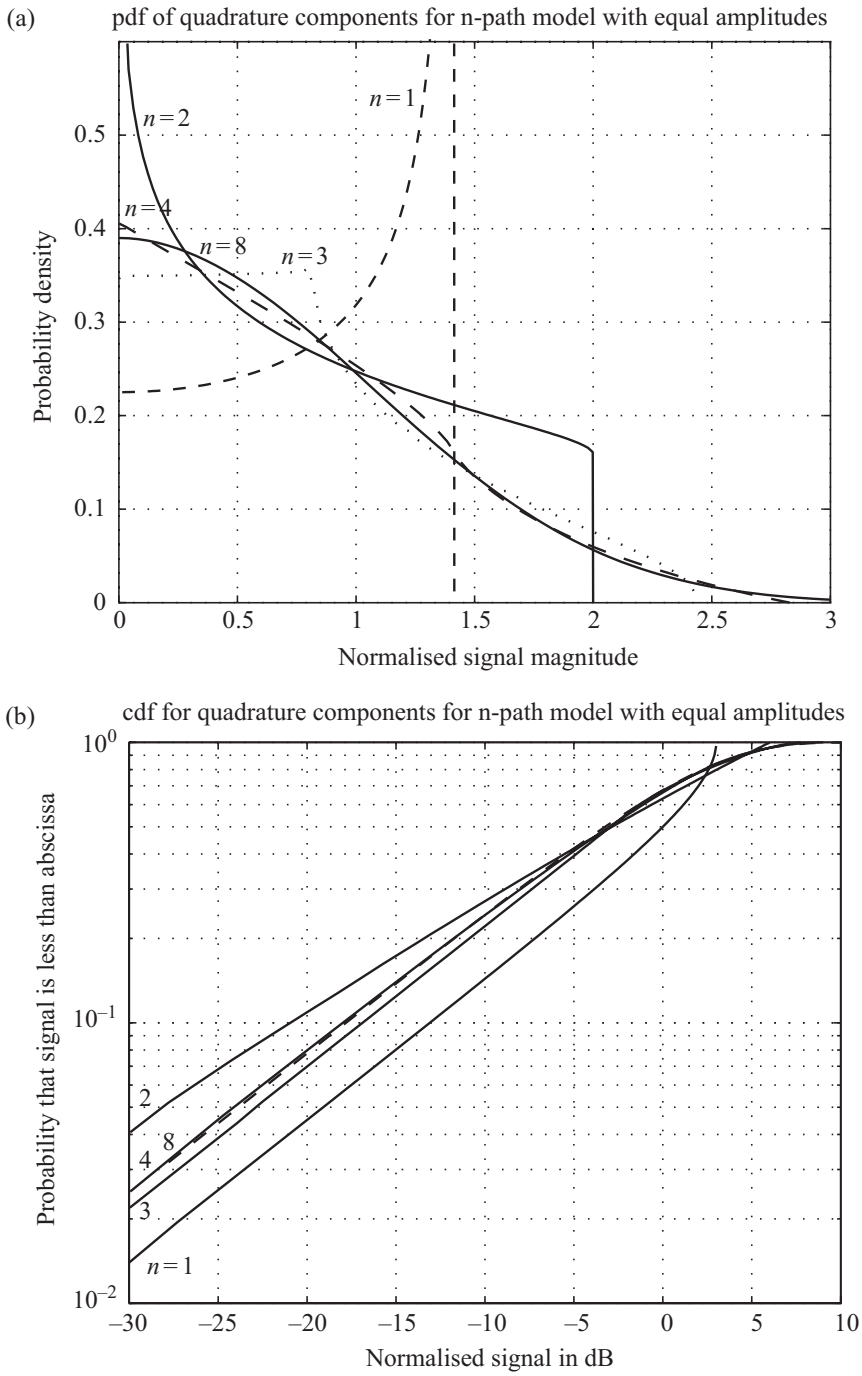


Figure 5.5.4 *(a) The pdf for the one-sided form of the resultant in-phase and quadrature signal of the common-amplitude few-path mode. (b) The cdf for the in-phase and quadrature components for the n-path model*

The cdf for the in-phase and quadrature components is, here for the one-sided form (where x_0 is positive)[5],

$$P_{x^{(n)}}(|x| \leq x_0) = 2\int_0^{x_0} p_{x^{(n)}}(x)\,\mathrm{d}x$$
$$= \frac{2}{\pi}\int_0^{\infty} \frac{1}{\beta}\,(J_0(a\beta))^n \sin\beta x_0\,\mathrm{d}\beta. \tag{5.5.9}$$

This is also shown in Figure 5.5.4. For the $n = 1$ case, the solution can be found as

$$P_{x^{(1)}}(|x| \leq x_0) = \frac{2}{\pi}\sin^{-1} x_0. \tag{5.5.10}$$

Although the pdfs are very different for the small-n cases, the cdfs do not reflect these changes so much. Only the $n = 1, 2$ lines deviate from the others by more than a couple of dB.

5.5.2.3 Envelope

The pdf for the envelope is

$$p_{r^{(n)}}(r) = r\int_0^{\infty} \beta J_0(r\beta)\ (J_0(a\beta))^n\ \mathrm{d}\beta \tag{5.5.11}$$

and this is similarly well approximated (cf. the Gaussian distribution above) when $n > \sim 10$ by the Rayleigh distribution,

$$p_{r^{(n)}}(r) \approx \frac{2r}{na^2}\,\mathrm{e}^{-r^2/(na^2)}, \quad n \text{ large}, \tag{5.5.12}$$

where the variance $\sigma^2 = a^2 n/2$ is that of the phase components and the envelope variance is $\sigma_r^2 = (4 - \pi)a^2(n/2)$.

The envelope pdf for the $n = 1$ case reduces to a $(\sin x)/x$ form which has a delta function characteristic, i.e.

$$p_{r^{(1)}}(r) = \delta(r - a), \tag{5.5.13}$$

and this result is intuitively obvious from considering $n = 1$ (i.e. $r = a$) in Figure 5.5.4.

The envelope squared, $R = r^2$, is the power of the signal, and the pdf of the power follows as

$$p_{R^{(n)}}(R) = \frac{1}{2}\int_0^{\infty} \beta\ J_0(\sqrt{R}\beta)\ (J_0(a\beta))^n\ \mathrm{d}\beta. \tag{5.5.14}$$

[5] Note from $P(|x| \leq na) = 1$, the result $\int_0^{\infty}(1/\beta)\,(J_0(a\beta))^n \sin(\beta na)\,\mathrm{d}\beta = \pi/2, \quad n \geq 1.$

The cdf for the envelope is[6]

$$P_{r^{(n)}}(r \le r_0) = \int_0^{r_0} \int_0^{\infty} r\, J_1(r\beta)\, (J_0(a\beta))^n \, \mathrm{d}\beta \, \mathrm{d}r$$
$$= r_0 \int_0^{\infty} J_1(r_0\beta)\, (J_0(a\beta))^n \, \mathrm{d}\beta. \tag{5.5.15}$$

The envelope quantities are depicted in Figure 5.5.5, with the scaling again by the standard deviation of the Gaussian component. Again, there are extreme changes in the form of the pdf as the number of components increases from $n = 1$. (The delta function $n = 1$ case is not included.) At $n = 4$, the form is becoming more Rayleigh-like, and by $n = 8$, the pdf is essentially Rayleigh, except at the upper tail. The cdf shows the variation (again, the $n = 1$ case is not included, but would be a vertical line on the plot), with the oscillation about the Rayleigh line, which is essentially the same as the $n = 8$ line. Very little difference is seen between the $n = 4$ and $n = 8$ cdfs.

5.5.2.4 Closed forms

For the cases $n = 2$ and $n = 3$, i.e. the two-path and three-path channels, there are some closed form expressions which are worth noting since the numeric integration is more demanding for lower values of n. For the in-phase and quadrature components, the pdf is

$$p_{x^{(2)}}(x) = \frac{1}{a\pi^2} K\left(\sqrt{1 - \left(\frac{x}{2a}\right)^2}\right), \quad x < 2a, \tag{5.5.16}$$

where $K(k)$ is the complete elliptic integral of the first kind with modulus k. This blows up at $x = 0$ and has zero value for $|x| > 2a$, as seen in the plot. The cdf, here for the absolute value of x exceeding a positive number x_0, is

$$P_{x^{(2)}}(|x| \le x_0) = 1 - 2\int_{x_0}^{2a} p_{x^{(2)}}(x)\, \mathrm{d}x. \tag{5.5.17}$$

The integral can be expressed as a converging series (Slack, 1946), but it is also straightforward to apply numerical integration in this case.

The two-path model of Section 5.4.2 also gives expressions for the pdf and cdf for the power (envelope squared), and transformation gives the envelope functions as

$$p_{r^{(2)}}(r) = \frac{2}{\pi\sqrt{(2a)^2 - r^2}}, \tag{5.5.18}$$

and

$$P_{r^{(2)}}(r \le r_0) = \frac{2}{\pi} \sin^{-1}\left(\frac{r_0}{2a}\right). \tag{5.5.19}$$

[6] Note from $P_{r^{(n)}}(r \le na) = 1$, the result $\int_0^{\infty} J_1(na\,\beta)\, (J_0(a\beta))^n \, \mathrm{d}\beta = 1/(na), \quad n \ge 1.$

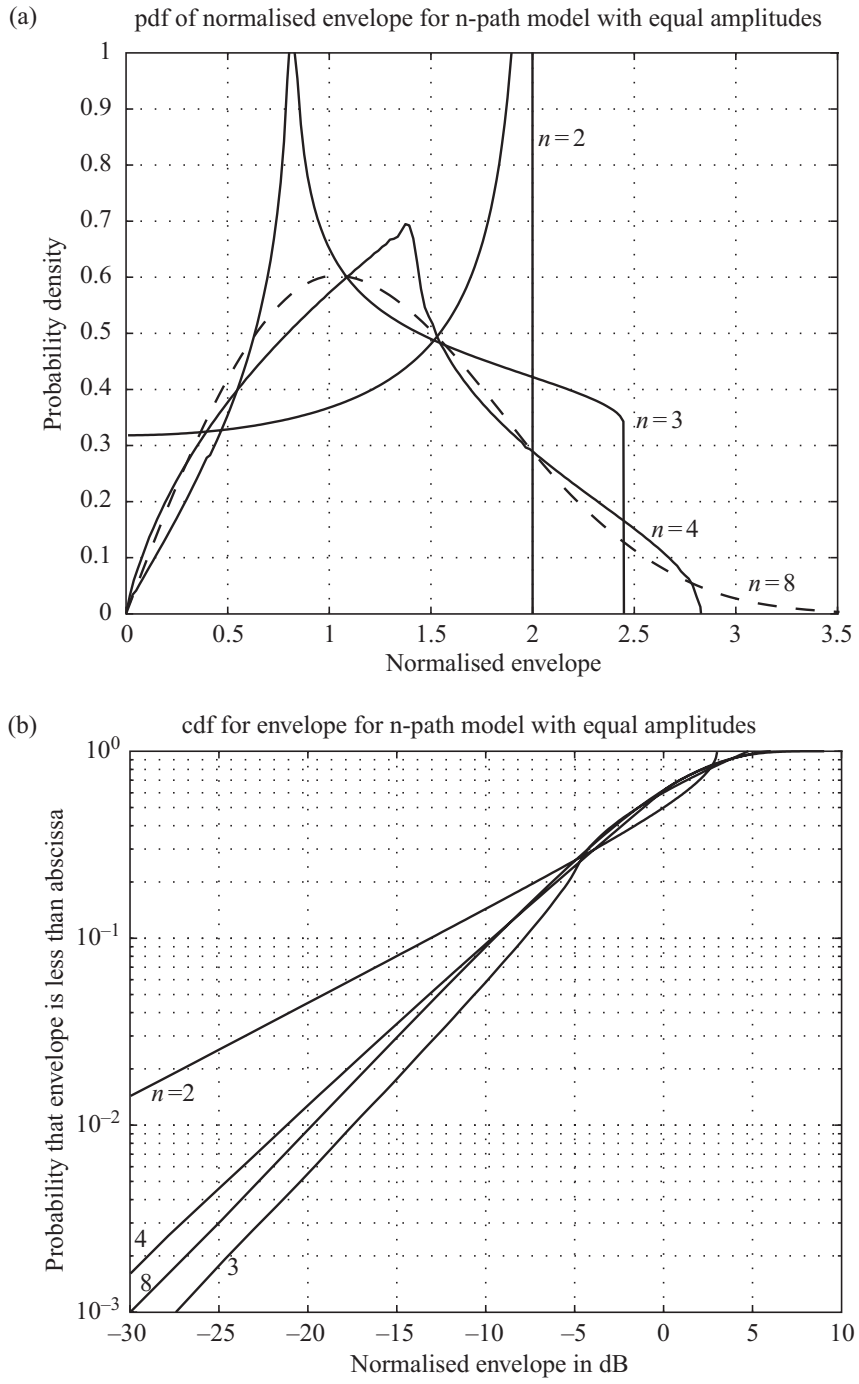


Figure 5.5.5 *(a) The pdf for the resultant envelope of the n-path channel with equal amplitudes and uniformly distributed phase scaled by $a\sqrt{n/2}$. (b) The cdf of the envelope*

Finally, for the three-path model (Slack, 1946; after Nicholson, 1920),

$$p_{r^{(3)}} = \frac{r}{\pi^2 a\sqrt{ar}}\, K\left(\frac{a + r\sqrt{(3a-r)(a+r)}}{4a\sqrt{ar}}\right), \quad a < r < 3a, \qquad (5.5.20)$$

which blows up at $r = a$ and is taken as zero for $r > 3a$, as indicated in Figure 5.5.5. The envelope is normalised by the phase component standard deviation $a\sqrt{n/2}$ in Figure 5.5.5. An alternative normalisation is the Rayleigh envelope standard deviation which is

$$\sigma_r = \sqrt{\frac{4-\pi}{4}(2\sigma)^2} = 0.9265\, a\sqrt{n/2},$$

or some 0.66 dB different from the component standard deviation, i.e. for this normalisation, the cdf curves would shift by 0.66 dB.

5.5.2.5 Different amplitudes case

For the case where the component amplitudes a_i are different, Bennett (1948) draws attention to Kluyver's (1906) formula which corresponds to the cdf for the resultant envelope,

$$P_{r_u^{(n)}}(r \leq r_0) = r_0 \int_0^\infty J_1(r_0\beta) \prod_{i=1}^{n} J_0(a_i\beta)\, \mathrm{d}\beta, \tag{5.5.21}$$

which reduces to the uniform amplitude case for common a_i.

5.6 References

ANDERSEN, J. BACH, EGGERS, P.C.F., and ANDERSEN, B.L.: 'Propagation aspects of datacommunication over the radio channel', *Proceedings of Eurocon*, June 1988, pp. 301–305.

ANDERSON, D.G. and ASKNE, J.I.H.: 'Wave packets in strongly dispersive media', *Proc. IEEE*, 1974, **62**, pp. 1518–1523.

BELLO, P.A. and NELIN, B.D.: 'The effect of frequency selective fading on the binary error probabilities of incoherent and differentially coherent matched filter receivers', *IEEE Transactions*, November 1973, **Com-21**, pp. 1214–1224.

BENNETT, W.R.: 'Distribution of the sum of randomly phased components', *Quarterly of Applied Mathematics*, 1948, **5** (4), pp. 385–393.

CLAASEN, T.A.C.M. and MECKLENBRAUKER, W.F.G.: 'The Wigner distribution – a tool for time-frequency signal analysis', Part 1: *Philips J. Res.*, 1980, **35** (3), pp. 217–250; Part 2: *Philips J. Res.*, 1980, **35** (4/5), pp. 276–300; Part 3: *Philips J. Res.*, 1980, **35** (6), pp. 372–389.

COHEN, L.: 'Time–frequency distributions-a review', *Proc. IEEE*, July 1989, **77** (7), pp. 941–981.

COHEN, L. and LEE, C.: 'Instantaneous, frequency, its standard deviation and multi-component signals', *SPIE*, 1988, **975** Advanced Algorithms for and Architectures for Signal Processing III, pp. 186–208.

COHEN, L. and LEE, C.: 'Instantaneous bandwidth for signals and spectrogram', *Proceedings of the International Symposium on Signal Processing and Applications*, Gold Coast, Australia, 1990, pp. 2451–2454.

COX, D.C.: 'Delay-Doppler characteristics of multipath propagation at 910 MHz in a suburban mobile radio environment', *IEEE Transactions*, September 1972, **AP-20** (5), pp. 625–635.

COX, D.C. and LECK, R.P.: 'Correlation bandwidth and delay spread multipath propagation statistics for 910 Mhz urban mobile radio channels', *IEEE Transactions*, 1975, **Com-23** (11), pp. 1271–1280.

FLEURY, B.H.: 'An uncertainty relation for WSS processes and its application to WSSUS systems', *IEEE Transactions*, December 1996, **Com-44** (12), pp. 1632–1635.

GANS, M.J.: 'A power spectral theory of propagation in the mobile-radio environment', *IEEE Transactions*, February 1972, **VT-21** (1).

HAHN, S.L.: 'Complex variable frequency electric circuit theory', *Proc IEEE*, June 1964, **52**, pp. 735–736.

KLUYVER, J.C.: 'A local probability problem', *Proceedings of the Royal Academy of Sciences*, Amsterdam, 1906, **6**, p. 341.

LEE, W.C.Y.: *Mobile Communications Engineering*, McGraw-Hill, 1982.

LIN, K.H., YEH, K.C., SOICHER, H., REINISCH, B.W., and GAMACHE, R.R.: 'Vertical ionograms and dispersive bandwidth for an oblique path', *Radio Science*, July-August 1989, **24** (4), pp. 519–526.

NICHOLSON, J.W.: 'Generalisation of a theorem due to sonine', *Quarterly Journal of Pure and Applied Mathematics*, 1920, **48**, p. 321.

PAPOULIS, A.: *The Fourier Integral and its Applications*, McGraw-Hill, 1962, (reissued 1987).

PARL, S.A.: 'Characterisation of multiple parameters for line-of-sight microwave propagation', *IEEE Transactions*, 1983, **AP-31**, pp. 938–948.

POLETTI, M.A.: 'The development of instantaneous bandwidth via local signal expansion,' *Signal Processing*, 1993, **31**, pp. 273–281.

POLETTI, M.A. and VAUGHAN, R.G.: 'Reduction of multipath fading effects in single variable modulations', *Proceedings of the International Symposium on Signal Processing and Applications*, Gold Coast, Australia, 1990, pp. 672–676.

POLETTI, M.A. and VAUGHAN, R.G.: 'Signal dispersion and negative group delay', Internal Report, Industrial Development Division, DSIR, March 1992.

RICE, S.O.: 'Mathematical analysis of random noise', *Bell System Technical Journal*, 1944, **23**, pp. 282–332.

RICE, S.O.: 'Statistical properties of a sine wave plus random noise', *Bell System Technical Journal*, 1948, **27**, pp. 109–157.

RIHACZEK, A.W.: 'Signal energy distribution in time and frequency,' *IEEE Trans.*, May 1968, **IT-14** (3), pp. 369–374.

RIHACZEK, A.W.: *Principles of High Resolution Radar*, McGraw-Hill, New York, 1969.

RUMMLER, W.D.: 'A new selective fade model: application to propagation data', *Bell System Technical Journal*, May–June 1979, **58** (5), pp. 1037–1071.

SAKAGAMI, S. and HOSOYA, Y.: 'Some experimental results on in-band amplitude dispersion and a method of estimating linear amplitude dispersion', *IEEE Transactions*, 1982, **COM-30**, pp. 1875–1877.

SLACK, M.: 'The probability distributions of sinusoidal oscillations combined in random phase', *Proceedings of the IEE*, 1946, **93**, Pt.3, pp. 76–86.

VAUGHAN, R.G.: 'Signals in mobile communications – a review', *IEEE Transactions*, November 1986, **VT-35** (4), pp. 133–145.

VILLE, J.: 'Théorie et applications de la notion de signal analytique,' *Cables et Transmission*, 1948, **2 A** (1), pp. 61–74.

WETZEL, L.: 'On the theory of signal distortion due to ionospheric dispersion', *Research Paper P317*, Inst. for defence Anal., Arlington, Va., 22202, July 1967. (Reviewed in Lin *et al.* (1989)).

WIGNER, E.P.: 'On the quantum correction for thermodynamic equilibrium', *Phys Rev*., 1932, **40**, pp. 749–759.

WOODWARD, P.M.: *Probability and Information Theory, with Applications in Radar*, Pergamon Press, New York, 1953.

Chapter 6

Short-term behaviour of many-path models and scenarios

The extension from the two-path model to the N-paths is conceptually straightforward from the point of view that the channel can be described by simple summations,

$$h(\tau, z) = \sum_{i=1}^{N} a_i \mathrm{e}^{\mathrm{j}\alpha_i} \delta(\tau - \tau_i) \mathrm{e}^{\mathrm{j}u_i z} \Leftrightarrow H(\omega, z) = \sum_{i=1}^{N} a_i \mathrm{e}^{\mathrm{j}\alpha_i} \mathrm{e}^{-\mathrm{j}\omega\tau_i} \mathrm{e}^{\mathrm{j}u_i z}. \qquad (6.0.1)$$

This chapter looks at the short-term channel behaviour arising from the many-path model, where no single path is dominant, and discusses the scenarios of sources used in channel simulations.

6.1 Many-path model

The many-path model can be presented as in Figure 6.1.1(a,b). A snapshot of discrete scatterers seen by the mobile is shown in contrast with a sampled model of an ensemble averaged distribution. The model depicted here is a uniform scenario which has an exponential-type delay distribution that is the same from all directions. In a computer simulation, these are conveniently implemented at a single radius and assigned random delay times. However, it is often convenient to impose a delay distribution in each angular direction, and an angular power distribution can be assigned directly on to the amplitudes of effective sources in given directions. Simulations using the scatterers of (a) and (b) generate different signals in the sense that (a) is a snapshot model and does not feature a uniform phase or uniform angular power distribution, i.e. (a) is more realistic for local signal behaviour (assuming the scatterers are appropriately placed); and (b) is a model of an averaged distribution and is more realistic in producing an average expected behaviour.

The field strength at the receiver is now Rayleigh distributed in both the distance coordinate and the frequency coordinate, and this is evident in Figure 1.4.8, from

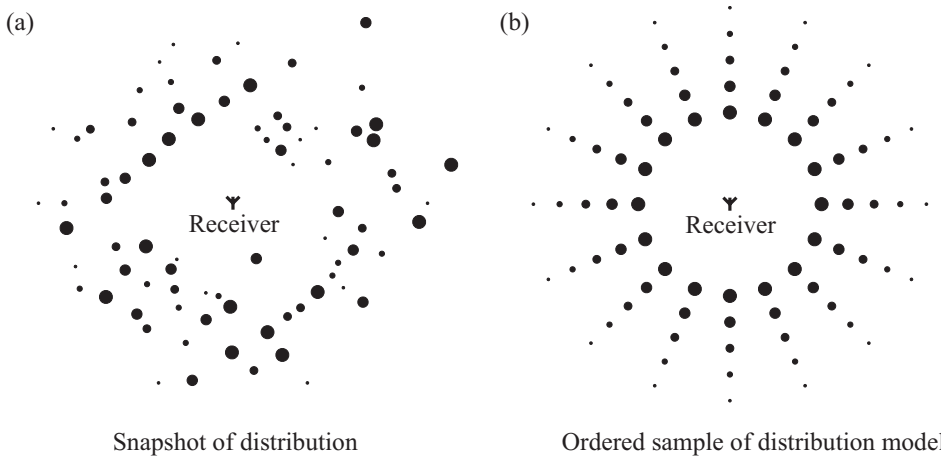


Figure 6.1.1 Indicative sources in space which surround the receiver and have, on average, an exponential delay profile model for simulations. (a) A random snapshot sample of the scatterers and (b) ordered samples of a distribution model. In a simulation using a moving receiver with fixed scatterers in (b), the receiver trajectory should be small relative to the delay differences so that the angular power density is properly preserved

Section 1.4.3. For narrowband channels, the envelope is Rayleigh and the phase is uniform, as already established. The general frequency and distance (quasi-periodic) dependencies of the envelope can be found from the above individual transforms or the combined dependences from the double transform discussed in Chapter 2,

$$P(u, \tau) \Leftrightarrow R_H(\Delta\omega, \Delta z). \tag{6.1.1}$$

Here, $P(u, \tau)$ is recalled as the distribution of power in the direction $u = k_C \cos\theta$ and delay time τ, and $R_H(\Delta\omega,\ \Delta z)$ is the correlation function of the transfer function at frequency difference $\Delta\omega$ and at distance separation Δz. R_H is also denoted by the spectral quantity S_H (Appendix B) since it is the spectrum of the scattering distribution.

6.1.1 Moving terminal in a static many-path scenario: linear systems model

The transfer function of a multipath channel can be modelled as a linear, time-variant system. This section relates also to Chapter 2 but discusses the propagation channel from a linear systems perspective.

The output signal is given by the usual convolution of the system impulse response $h(t)$ and the input signal $x(t)$, i.e.

$$y(t) = \int_{-\infty}^{\infty} h(t - \tau)\, x(\tau)\, \mathrm{d}\tau. \tag{6.1.2}$$

In the general case of mobile communications, the channel is time variant, and $h(t-\tau)$ is replaced by $h(t, \tau)$ which represents the response at time t to an impulse at time τ in a linear, time-variant system. With causality, $h(t, \tau) = 0$ for $\tau > 0$. The output signal can now be expressed as a distribution (i.e. superposition) integral,

$$y(t) = \int_{-\infty}^{t} h(t, \tau)\, x(\tau)\, \mathrm{d}\tau. \tag{6.1.3}$$

The use of the two time parameters can be confusing, and the notion of *propagation time* (τ) of signals through the channels, and *clock time* (t), relating here specifically to the rate of change of the propagation channel. The propagation time quantities are 'short', typically many microseconds, corresponding to many hundreds of metres of free space (1 μs corresponds to 300 m). The clock time is 'slow', and relates directly to the speed of the mobile terminal if there are static scatterers, a fixed beam antenna and a linear mobile trajectory. Scaling by the terminal's velocity V allows the clock time to correspond to the receiver position, i.e. $z = Vt$, as long as the effective scatterers are constant. The output for the linear system can then be written, by substituting z for t, as

$$y(z) = \int_{-\infty}^{z} h(z, \tau)\, x(\tau)\, \mathrm{d}\tau. \tag{6.1.4}$$

A depiction of the changing impulse response is given in Figure 6.1.2. If the input signal to the system is a series of delta functions separated by $\Delta T = \Delta z / V$, which appear on the line $\tau = z = Vt$, i.e.

$$x(Vt) = x(\tau) = \sum_i a_i \delta\left(\tau - i\,\Delta T\right) \tag{6.1.5a}$$

or

$$x\left(\frac{z}{V}\right) = \sum_i a_i \delta\left(\frac{z}{V} - i\,\frac{\Delta z}{V}\right), \tag{6.1.5b}$$

then the output can be written

$$y(z) = y(\tau) = \sum_{-\infty}^{\tau} a_i h\,\left(i\,\frac{\Delta z}{V}, \tau\right), \tag{6.1.6}$$

and the various impulse responses are overlapping in general. The output signal thus stems from many different impulse responses and a statistical description is required for the convolution action of the multipath channel.

6.1.2 Bello functions in mobile communications

One of Bello's (1963) versions of the convolution relation is

$$y(t) = \int h(t, \tau)\, x(t - \tau)\, \mathrm{d}\tau \quad \left(= \int h(z/V, \tau)\, x(z/V - \tau)\, \mathrm{d}\tau\right), \tag{6.1.7}$$

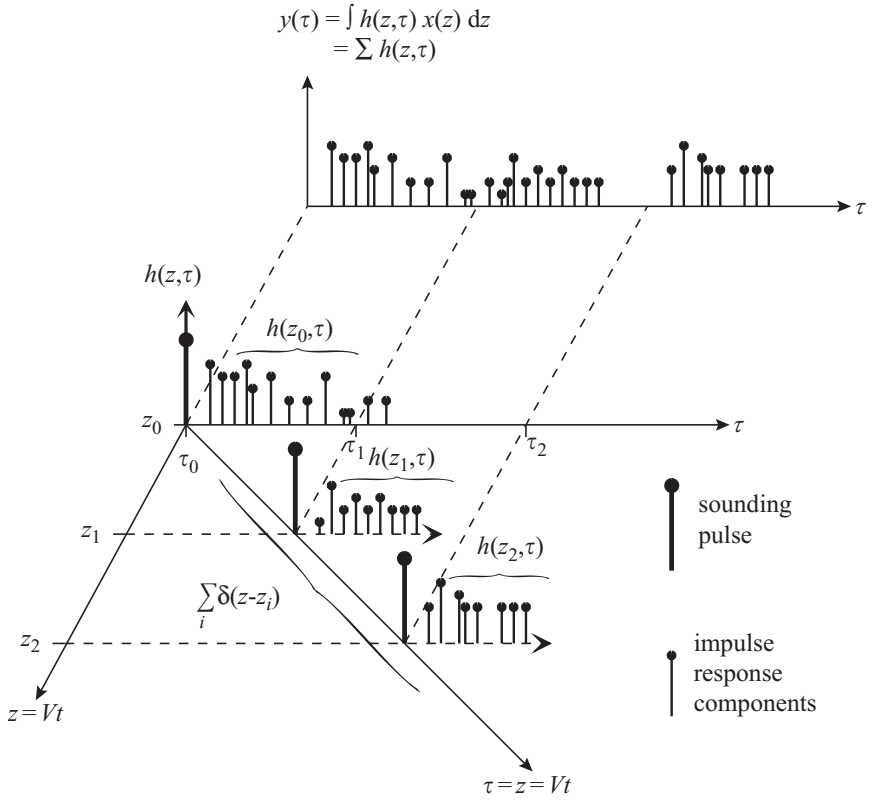


Figure 6.1.2 The changing impulse response of a multipath channel as a function of position z. V is the fixed velocity of the receiver and τ is the impulse response time variable. For an input comprising delta functions (sounding pulses), the output is constructed from the sums of differing impulses, here depicted as an ideal (dispersionless) channel

where $h(t, \tau)$ is called the *input delay-spread function* since the input signal contains the delay. Bello's *output delay-spread function* $h(t + \tau, \tau)$ is defined in a corresponding way. There are several Bello functions comprising $h(t, \tau)$ and its delayed and transformed versions. These were originally derived with propagation through a scintillating atmosphere in mind, where independent (i.e. uncorrelated) scatterers moved around in the physical channel, and the transmit and receive terminals were static. The functions are often quoted for fading channel characterisation. However, in mobile communications, the correlations of the undelayed functions are sufficient for most purposes. These are included in Chapter 2.

6.1.3 Finite bandwidth effects on the many-path impulse response

In most communications applications, the propagation channel can be treated as having effectively an infinite bandwidth. This is because the radio-channel bandwidth (which relates to the communications application), is bandlimited by elements of

the communications system itself. These elements are the antenna and front-end electronics, and this is discussed in Chapter 1.

The following treatment is sometimes referred to as the Turin (1956) model for multipath propagation. Only the delay time is of interest, so the mobile position is fixed and the channel is time invariant. The input signal to the propagation medium is denoted $x(t)\mathrm{e}^{\mathrm{j}\omega_c t}$, and the output from the medium is the convolution

$$\begin{aligned} y(t) &= x(t)\mathrm{e}^{\mathrm{j}\omega_c t} * \sum_i a_i \delta\,(t-\tau_i) \\ &= \mathrm{e}^{\mathrm{j}\omega_c t} \sum_i a_i \mathrm{e}^{-\mathrm{j}\omega_c \tau_i}\; x(t-\tau_i), \end{aligned} \tag{6.1.8}$$

and the a_i are complex. The 'infinite' bandwidth of the non-dispersive channel is again reflected in the unquantised nature of τ_i.

Imposing an equivalent rectangular bandwidth of B rad s^{-1} on the complex envelope $x(t)$, the delays in the radio channel can be resolved, in the classical Fourier sense, to within $T \approx 2\pi/B$. The discrete signal delays of the propagation channel model are thus combined and quantised in time by the filter action of the transmission system, receiving antenna, front end, filters, etc., of the radio channel, into L delay bins, each T s long. The bandlimited signal of the mobile radio channel model is then

$$\tilde{x}(t-\tau_i) \approx x\left(t-(l-\tfrac{1}{2})\,T\right), \quad \text{for } (l-1)T \le \tau_i \le lT,\ l=1,2,\ldots,L;$$

or

$$\tilde{x}(t-\tau_i) \approx x(t-\tau_l), \tag{6.1.9}$$

and the τ_l are quantised, i.e. τ_l are the quantised, discrete time bins. The output from the radio channel is therefore written

$$y(t) = \mathrm{e}^{\mathrm{j}\omega_c t} \sum_{l=1}^{L} A_l\; x(t-\tau_l) \tag{6.1.10}$$

where

$$A_l = \sum_i a_i \mathrm{e}^{-\mathrm{j}\omega_c \tau_i}. \tag{6.1.11}$$

If there are sufficient independent contributions (many i) in each lth bin, then the A_l are independent, complex zero-mean Gaussian variables. Each bandlimited component at τ_l then exhibits independent, Rayleigh envelopes and uniform phase. This is caused by the bandwidth restriction on the signal, and is independent of the details of the propagation channel, i.e. the distribution of a_i (as long as i is large, say $i >\sim 10$). This is the discrete equivalent of the more general statistical approach where integrals are used with continuous distribution models.

The impulse response of the channel is now proportional to

$$\tilde{h}_B(\tau) = \sum_l^L A_l \mathrm{sinc}\left(B(\tau-\tau_i)\right), \tag{6.1.12}$$

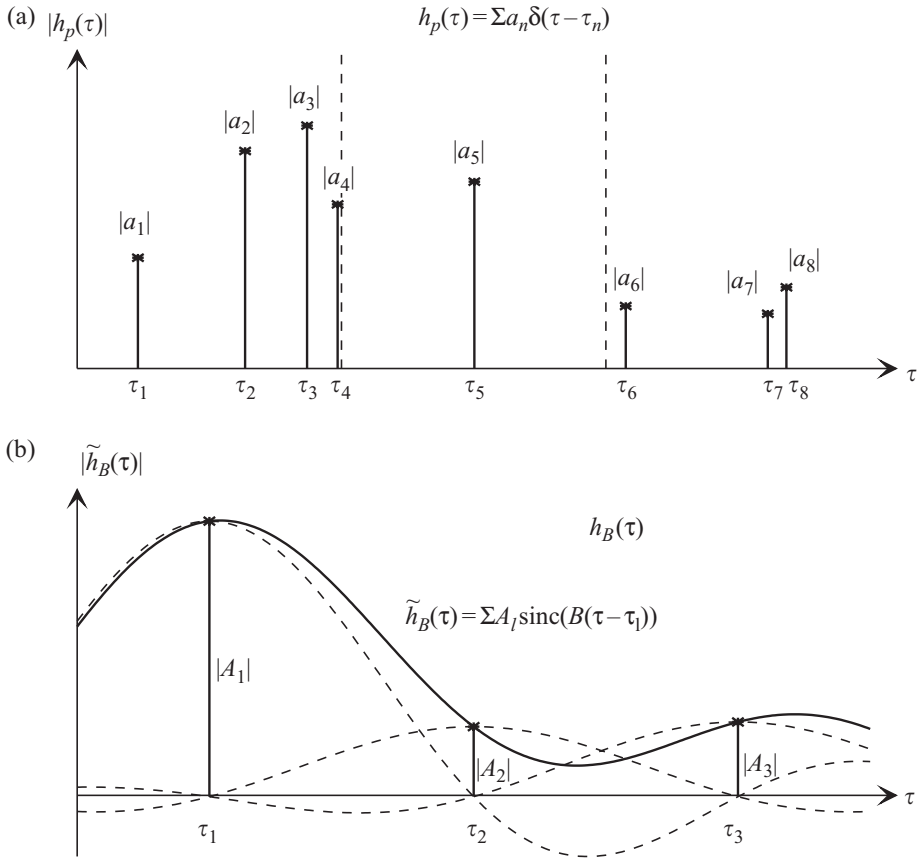


Figure 6.1.3 The impulse response magnitude of (a) the propagation channel, modelled with effectively infinite bandwidth, and (b) the radio channel with finite bandwidth. The A_l can be independent complex Gaussian variables

which is depicted in Figure 6.1.3. With finite rectangular bandwidth, the delta functions that described different paths become sinc functions which are quantised in delay time, or location for a moving terminal and static effective sources.

6.1.3.1 Doppler spectrum

In the Doppler frequency domain, the analogous bandlimiting effect is as follows. The delay time is fixed in the same way that the impulse response dealt with a fixed mobile position. The mathematically convenient delta functions in $a(u)$ or $P(u)$ imply, in the Fourier transform sense, infinite distance covered by the mobile. The directions (delta function locations) to the discrete sources can be estimated from the information over a finite mobile trajectory by super-resolution techniques. In the Fourier transform sense, a finite linear mobile trajectory causes the delta fuctions to become sinc functions and the Doppler profile becomes proportional to the sum

$$\tilde{a}(u) = \sum_i A_l \,\mathrm{sinc}(z_0(u - u_i)), \tag{6.1.13}$$

where the A_l are now the complex Gaussians (Rayleigh-distributed magnitudes) arising from a summation of several discrete random phase sources in each Doppler bin. This represents the directional information of the sources that can be found from moving the receiver a distance z_0, and using the basic Fourier transform, $h(z) \Leftrightarrow a(u)$. The basic Fourier resolution is treated in the following section and super-resolution techniques are given in Chapter 7.

6.1.4 Finite length effect: Fourier angular resolution from a mobile trajectory

A trajectory of length L m offers a Fourier resolution of $u \approx 2\pi/L$ rad m^{-1}. Alternatively stated, the angular Fourier resolution available from the distance L is $\Delta\theta = \theta_{\mathrm{U}} - \theta_{\mathrm{L}}$, where

$$\cos\theta_{\mathrm{L}} - \cos\theta_{\mathrm{U}} = \frac{\lambda}{L} \tag{6.1.14}$$

(cf. Figure 6.3.1).

6.1.4.1 Resolution in the directions of the trajectory

In the direction of motion along the z-axis, we have $\theta_{\mathrm{L}} = 0$ and $\Delta\theta = \Delta\theta^{(E)} = \theta_{\mathrm{U}}$, which results in

$$\begin{aligned} \Delta\theta^{(E)} &= 2\sin^{-1}\sqrt{\frac{\lambda}{2L}} \\ &\approx \sqrt{\frac{2\lambda}{L}}, \quad L \gg \lambda. \end{aligned} \tag{6.1.15}$$

Note that the Doppler spectrum cannot resolve between the different sides of the trajectory (z-axis). This latter case with $\theta_{\mathrm{L}} = 0$ is otherwise analogous to the resolution at endfire of an array of length L.

6.1.4.2 Resolution at broadside to trajectory

Broadside to the mobile trajectory, the resolution is finer. The directions θ_{U} and θ_{L} can be either symmetric about $\pi/2$ (broadside) or we can fix $\theta_{\mathrm{L}} < \theta_{\mathrm{U}} = \pi/2$. The result is

$$\begin{aligned} \Delta\theta^{(B)} &= 2\sin^{-1}\left(\frac{\lambda}{2L}\right) \\ &\approx \frac{\lambda}{L}, \quad L \gg \lambda. \end{aligned} \tag{6.1.16}$$

This is similar to the resolution of a broadside array of length L. The resolution is plotted in Figure 6.1.4. An example is for a trajectory of $L = 2\lambda$, where scatterers can be Fourier resolved to about 30° at broadside and to about 60° in the direction of motion. In other directions, the aperture-limited resolution is between these values.

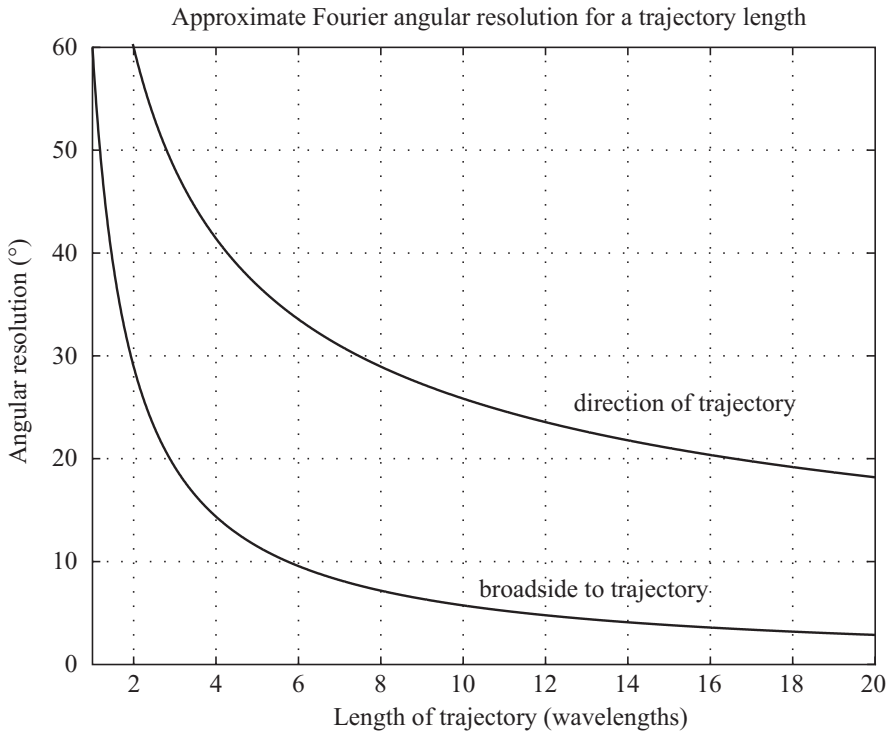


Figure 6.1.4 The Fourier resolution of scatterers in the direction of motion and at broadside to a linear trajectory of a mobile receiver

6.2 Derivatives of the transfer function

6.2.1 Phase derivatives: group delay and random FM

The variation of the phase of the transfer function has a direct bearing on the performance of angle-modulated transmissions. The transfer function $H(\omega, z)$ features phase variation with respect to frequency ω and position z or time through $z = Vt$, where V is the speed and the direction is along the positive z-axis. Denoting the phase of the transfer function with ψ, the group delay is recalled as

$$\tau_{\mathrm{g}} = -\frac{\partial \psi}{\partial z} \tag{6.2.1}$$

and similarly, the random FM (Doppler frequency) is

$$\omega_{\mathrm{D}} = \frac{\partial \psi}{\partial t}. \tag{6.2.2}$$

Normalising by the speed, we get the Doppler shift per unit velocity, or spatial Doppler frequency (rad m^{-1}), written

$$u = \frac{\omega_{\mathrm{D}}}{V} = \frac{\partial \psi}{\partial z}. \tag{6.2.3}$$

So results for this spatial derivative hold, by direct analogy, from the group delay case (Vaughan, 1986).

The derivation for the statistics of the phase derivatives uses a more general approach than in the above sections (Andersen *et al.*, 1990). The notation for the components of the transfer function here is

$$H(\omega, z) = \mathrm{Re}\{H(\omega, z)\} + \mathrm{j}\,\mathrm{Im}\{H(\omega, z)\} = H_\mathrm{R} + \mathrm{j}H_\mathrm{I}. \tag{6.2.4}$$

The phase derivative is written, from

$$\psi = \tan^{-1}\left(\frac{H_\mathrm{I}}{H_\mathrm{R}}\right) \tag{6.2.5}$$

as

$$u = \frac{H_\mathrm{R} H_\mathrm{I}' - H_\mathrm{R}' H_\mathrm{I}}{|H|^2}, \tag{6.2.6}$$

where the derivative notation is, here for the spatial case,

$$H' = H_z'(\omega, z) = \frac{\partial H(\omega, z)}{\partial z} = \mathrm{Re}\{H'\} + \mathrm{j}\,\mathrm{Im}\{H'\} = H_\mathrm{R}' + \mathrm{j}H_\mathrm{I}'. \tag{6.2.7}$$

The Fourier transform expression from the effective scattering distribution, $a(\tau, u)$, is

$$H_z'(\omega, z) = \frac{\mathrm{j}}{2\pi}\int_{-k_\mathrm{C}}^{k_\mathrm{C}}\int_0^\infty u a(\tau, u)\mathrm{e}^{\mathrm{j}(-\omega\tau + uz)}\,\mathrm{d}\tau\,\mathrm{d}u. \tag{6.2.8}$$

The basic many-path model assumes that the effective scattering distribution is distributed as a zero-mean, complex Gaussian. Arranging the real and imaginary parts as a vector of zero mean, real Gaussians

$$\boldsymbol{a}_\mathrm{RI}(\tau, u) = \left[\mathrm{Re}\{a(\tau, u)\},\ \mathrm{Im}\{a(\tau, u)\}\right], \tag{6.2.9}$$

then its covariance matrix is purely real, given here by

$$\begin{aligned}\boldsymbol{M}_a &= E\left\{\boldsymbol{a}_\mathrm{RI}(a, \tau)\,\boldsymbol{a}_\mathrm{RI}(a_0, \tau_0)\right\} \\ &= \begin{bmatrix} \frac{1}{2}P(\tau, u)\delta(\tau - \tau_0)2\pi\delta(u - u_0) & 0 \\ 0 & \frac{1}{2}P(\tau, u)\delta(\tau - \tau_0)\,2\pi\delta(u - u_0)\end{bmatrix}.\end{aligned} \tag{6.2.10}$$

It follows that a vector

$$\boldsymbol{X} = \left[H_\mathrm{R},\ H_\mathrm{I},\ H_\mathrm{R}',\ H_\mathrm{I}'\right] \tag{6.2.11}$$

also comprises zero mean, real Gaussians. (The derivative of a Gaussian will also be Gaussian since the derivative is a linear operation.) The covariance matrix is (cf. Rice,

1944, p. 80; Rice, 1948, equation 4.2; see also in Chapter 7)

$$\boldsymbol{M}_X = \begin{bmatrix} \langle H_\mathrm{R} H_\mathrm{R} \rangle & \langle H_\mathrm{R} H_\mathrm{I} \rangle & \langle H_\mathrm{R} H'_\mathrm{R} \rangle & \langle H_\mathrm{R} H'_\mathrm{I} \rangle \\ \langle H_\mathrm{I} H_\mathrm{R} \rangle & \langle H_\mathrm{I} H_\mathrm{I} \rangle & \langle H_\mathrm{I} H'_\mathrm{R} \rangle & \langle H_\mathrm{I} H'_\mathrm{I} \rangle \\ \langle H'_\mathrm{R} H_\mathrm{R} \rangle & \langle H'_\mathrm{R} H_\mathrm{I} \rangle & \langle H'_\mathrm{R} H'_\mathrm{R} \rangle & \langle H'_\mathrm{R} H'_\mathrm{I} \rangle \\ \langle H'_\mathrm{I} H_\mathrm{R} \rangle & \langle H'_\mathrm{I} H_\mathrm{I} \rangle & \langle H'_\mathrm{I} H'_\mathrm{R} \rangle & \langle H'_\mathrm{I} H'_\mathrm{I} \rangle \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2}P & 0 & 0 & \frac{1}{2}P\bar{u} \\ 0 & \frac{1}{2}P & -\frac{1}{2}P\bar{u} & 0 \\ 0 & -\frac{1}{2}P\bar{u} & \frac{1}{2}P\overline{u^2} & 0 \\ \frac{1}{2}P\bar{u} & 0 & 0 & \frac{1}{2}P\overline{u^2} \end{bmatrix} \tag{6.2.12}$$

where P is the total power and the overbar indicates the expected value. Note that for a symmetric Doppler spectrum, the entries featuring $\bar{u}$ will be zero. (In fact, it is not necessarily intuitive that off-diagonal elements should be non-zero.) The expected values of u and the cross-covariances are discussed further below.

The generality of the approach allows all the statistics to be found. For example, the normalised power of the transfer function, or channel gain,

$$\eta = \frac{|H_\mathrm{R}|^2 + |H_\mathrm{I}|^2}{P} = \frac{2|H|^2}{P} \tag{6.2.13}$$

from the upper-left 2×2 submatrix is the familiar chi-square with two degrees of freedom. Equivalently, the envelope, $r = |H|$, is Rayleigh distributed, i.e. the pdf is

$$p_r(r) = \frac{2r}{P} \mathrm{e}^{-r^2/P}. \tag{6.2.14}$$

The distribution of u is found by conditioning the vector $\boldsymbol{H}'_\mathrm{RI} = [H'_\mathrm{R}\ H'_\mathrm{I}]$ by $\boldsymbol{H}_\mathrm{RI} = [H_\mathrm{R}\ H_\mathrm{I}]$. The conditional distribution is normal because it is derived from a jointly normal distribution, derived below.

6.2.1.1 Use of Gaussian conditioning theorem

The conditioning theorem (Anderson, 1958) gives the distribution of a set of variables conditioned by another set of variables. In particular, if vector $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, meaning that the pdf is distributed as a multivariate Gaussian (see Appendix C) with a vector of means $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$, has partitioning

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \tag{6.2.15}$$

where $\boldsymbol{Y}_1$ is $p \times 1$ and $\boldsymbol{Y}_2$ is $q \times 1$, say, the means are in the vectors $\boldsymbol{\mu}_1$ $(p \times 1)$ and $\boldsymbol{\mu}_2$ $(q \times 1)$ respectively, and the covariance matrices are $\boldsymbol{\Sigma}_{11}$ $(p \times p)$, $\boldsymbol{\Sigma}_{22}$ $(q \times q)$, and $\boldsymbol{\Sigma}_{12}$ $(p \times q)$, then the distribution of $\boldsymbol{Y}_1 \mid \boldsymbol{Y}_2$ is $\mathcal{N}$ $(\boldsymbol{\mu}_{1\cdot2}, \boldsymbol{\Sigma}_{1\cdot2})$, where $\boldsymbol{\mu}_{1\cdot2}$ $(p \times 1)$ and $\boldsymbol{\Sigma}_{1\cdot2}$ $(p \times p)$ are given by

$$\boldsymbol{\mu}_{1\cdot2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\, \boldsymbol{\Sigma}_{22}^{-1}\, (\boldsymbol{Y}_1 - \boldsymbol{\mu}_2) \tag{6.2.16}$$

and

$$\boldsymbol{\Sigma}_{1\cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\,\boldsymbol{\Sigma}_{22}^{-1}\,\boldsymbol{\Sigma}_{21}^{-1}. \tag{6.2.17}$$

6.2.1.2 Student-T distribution for phase derivatives

Applying the conditioning theorem for the matrix $\boldsymbol{Y}_1 = \boldsymbol{H}'_{\mathrm{RI}}$ and $\boldsymbol{Y}_2 = \boldsymbol{H}_{\mathrm{RI}}$, we get the distribution of $\boldsymbol{H}'_{\mathrm{RI}} \mid \boldsymbol{H}_{\mathrm{RI}}$ to have mean

$$\boldsymbol{\mu}_{1\cdot 2} = [-\bar{u}H_{\mathrm{I}},\ \bar{u}H_{\mathrm{R}}] \tag{6.2.18}$$

and covariance

$$\boldsymbol{\Sigma}_{1\cdot 2} = \begin{bmatrix} \frac{1}{2}Ps_u^2 & 0 \\ 0 & \frac{1}{2}Ps_u^2 \end{bmatrix} \tag{6.2.19}$$

where

$$s_u^2 = \overline{u^2} - \bar{u}^2 \tag{6.2.20}$$

is the variance of the spatial Doppler frequency, i.e. s_u is the angular spread of the power of the incoming waves. So the distribution of the single variable H'_{R} conditioned by $H = H_{\mathrm{R}} + \mathrm{j}H_{\mathrm{I}}$ for example, is $\mathcal{N}\left(-\bar{u}H_{\mathrm{I}},\ (1/2)P\overline{s_u^2}\right)$. Now applying these Gaussian distributions to the equation for the phase derivative u above, then the distribution is

$$u \mid H \sim \mathcal{N}\left(\bar{u},\ \frac{1}{2}\frac{P}{|H|^2}s_u^2\right). \tag{6.2.21}$$

The analogous result for the group delay is

$$\tau_{\mathrm{g}} \mid H \sim \mathcal{N}\left(\overline{\tau_{\mathrm{g}}},\ \frac{1}{2}\frac{P}{|H|^2}s^2\right) \tag{6.2.22}$$

and s^2 is the variance of the delay time (s is the delay spread).

Conditioning by the power gives

$$u \mid |H|^2 \sim \mathcal{N}\left(\overline{u},\ \frac{s_u^2}{|H|^2}\right), \tag{6.2.23}$$

and the explicit form is

$$p_u(u \mid |H|^2) = \frac{|H|}{\sqrt{2\pi}s_u}\mathrm{e}^{-(u-\bar{u})^2|H|^2/(2s_u^2)}. \tag{6.2.24}$$

This is used below to give expressions for the angular power spectrum.

Now the normalised, central moment for the spatial Doppler frequency is

$$v = \frac{(u-\bar{u})}{\sqrt{\frac{1}{2}Ps_u^2/|H|^2}} = \frac{(u-\bar{u})}{s_u/\sqrt{\eta}} \tag{6.2.25}$$

so we have, for a given $\eta = 2|H|^2/P$, that

$$v \mid \eta \sim \mathcal{N}(0,1), \tag{6.2.26}$$

and since η is fixed and independent of v, u (without the conditioning), it must be distributed as a student-T with two degrees of freedom. The student-T pdf is

$$p_u(u) = \frac{1}{2} \frac{s_u^2}{\left(s_u^2 + (u - \bar{u})^2\right)^{3/2}} \tag{6.2.27}$$

and the cdf is

$$P(u \leq U) = \frac{1}{2} + \frac{1}{2} \left(\frac{(U - \overline{u})}{\sqrt{s_u^2 + (U - \overline{u})^2}} \right). \tag{6.2.28}$$

A significant point is that, under the assumptions of the many-path model, the distributions of the phase derivatives are dependent only on the mean and spread of the distributions, and not the details of the form of the distributions. So a practical implication is that simulations for the effect of the random FM or group delay on a channel can be performed using a simple model, for example, the two Rayleigh-fading branches are required (the dynamic two-path model, discussed below).

6.2.2 Group delay distribution

From above, the cdf for the group delay is a student-T distribution with two degrees of freedom (Andersen *et al*., 1990)

$$P_{\tau_g}(\tau_g) = \frac{1}{2} \left(1 + \frac{\tau_g - \overline{\tau_g}}{\sqrt{s^2 + \left(\tau_g - \overline{\tau_g}\right)^2}} \right), \tag{6.2.29}$$

where s is the delay spread and $\overline{\tau_g}$ is the mean delay. The distribution depends only on the moments of the profile, the mean and the delay spread, s. The cdf for the group delay is given in Figure 6.2.1, along with results for the static two-path model.

For the heavy line, the ordinate is τ_g/s, which is the general result. The lighter lines are for the static two-path model, and for these, the ordinate is τ_g/τ_2, with a_2 as the parameter. The curves illustrate the large values of instantaneous delay spread, which occur only about the fades. As an example from the general Rayleigh case, the group delay will exceed 20 times the delay spread for 0.06 per cent of the time. A delay spread of 1 μs (typical urban outdoor) implies that a cw receiver with an omnidirectional antenna may need to handle synchronisation swings of up to ±20 μs in order to track for 99.4 per cent of the time.

6.2.3 Group delay excursions with mobile position

Near a deep fade, interesting phenomena occur in the two dimensions which are not apparent from the one-dimensional model (Andersen *et al*., 1990). For the fade at (z_0, ω_0), the transfer function can be Taylor expanded as

$$H(\omega, z) = H(\omega_0, z_0) + H'_\omega(\omega_0, z_0)\Delta\omega + H'_z(\omega_0, z_0)\Delta z. \tag{6.2.30}$$

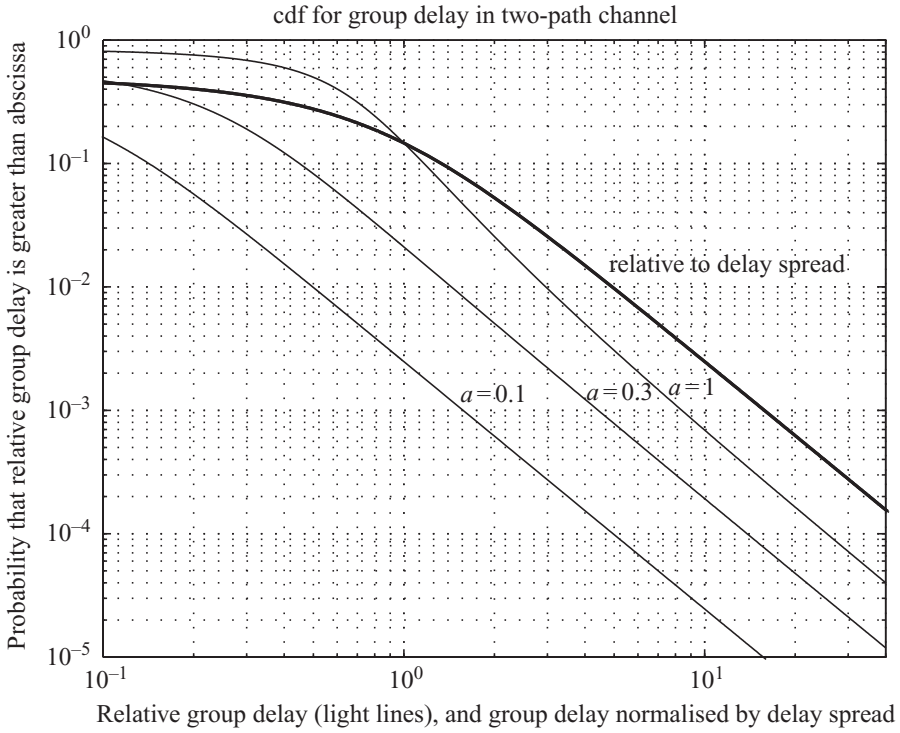


Figure 6.2.1 *The cdf for the absolute group delay for the general Rayleigh case (heavy line) with ordinate τ_g/s and for the static two-path case (light lines) with ordinate τ_g/τ_2. The results also apply, by analogy, to the random FM of the channel*

The group delay, from above, can be written

$$\begin{aligned} \tau_g &= \frac{c_1 + c_2 \Delta z}{|H(\omega_0, z_0)|^2} \\ &\approx \frac{c_2 \Delta z}{|H(\omega_0, z_0)|^2} \qquad (c_1 \ll c_2)\,, \end{aligned} \tag{6.2.31}$$

where

$$c_1 = \mathrm{Im}\{H(\omega_0, z_0)\} \cdot \mathrm{Re}\{H'_\omega(\omega, z)\} - \mathrm{Re}\{H(\omega_0, z_0)\} \cdot \mathrm{Im}\{H'_\omega(\omega, z)\}, \tag{6.2.32}$$

$$c_2 = \mathrm{Im}\{H'_z(\omega, z)\} \cdot \mathrm{Re}\{H'_\omega(\omega, z)\} - \mathrm{Re}\{H'_z(\omega, z)\} \cdot \mathrm{Im}\{H'_\omega(\omega, z)\} \tag{6.2.33}$$

are constants which can be positive or negative. Since c_1 is small near a fade, it is clear that there is a rapid group delay variation, from a large value to one side of z_0, to one of opposite sign on the other side of z_0. This is shown in Figure 6.2.2.

This behaviour has ramifications for the timing synchronisation of a moving receiver, where, as the terminal moves through a deep fade, the symbol position in time takes a potentially large excursion.

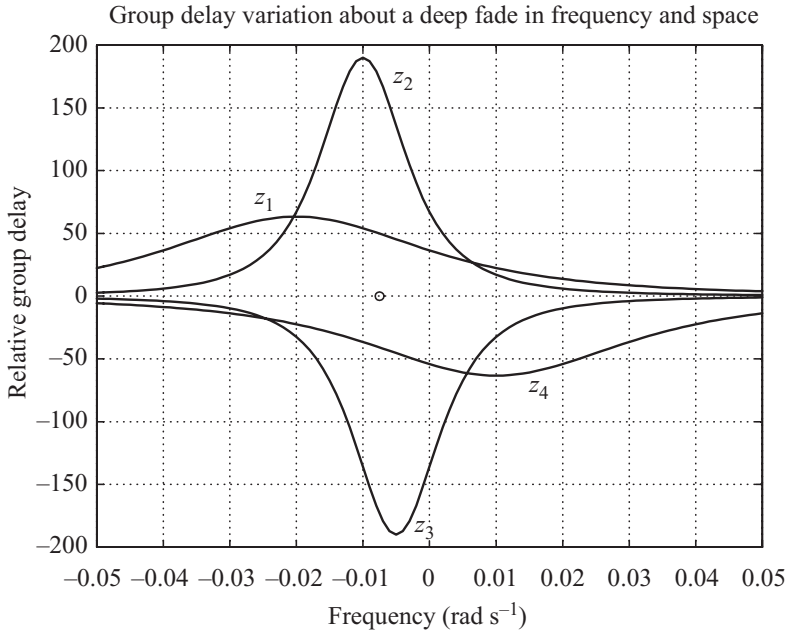


Figure 6.2.2 The variation of group delay as a function of three receiver positions z_1, z_2, z_3, which traverse a deep fade in the mobile channel transfer function. The group delay changes sign with large excursions

In the two-path model, the above equation for the group delay being proportional to the change in position breaks down, with $c_2 = 0$ and c_1 giving a $\Delta\tau = \tau_2$ dependency (as opposed to the Δz dependency for the general case above). This is a situation where the two-path case does not represent an important behaviour trend of the real-world, many-path channel.

The general behaviour in a static many-path model can also be shown on a three-dimensional plot of the group delay against distance and frequency about a fade (Figure 6.2.3). The group delay has large excursion over a small areal cell either side of the fade. The group delay is well behaved in non-fading regions.

6.2.3.1 Doppler frequency excursions

By the symmetry of the two-dimensional transfer function, the same behaviour occurs for the Doppler frequency shift. Large Doppler excursions of opposing polarity, the random FM, are found for two different frequencies in the neighbourhood of a fade, and are experienced by a mobile terminal as it moves through the region of the fade.

6.2.3.2 Fade depth in the 1D and 2D models

The behaviour about a 'deep fade' of this section applies to all fades in the 2D distribution which is Rayleigh-distributed in magnitude. The fades are always real in the sense that they go to zero at a point in the 2D plane. For cuts in the 2D transfer function, for example for fixed frequency or fixed distance, most zeros are complex;

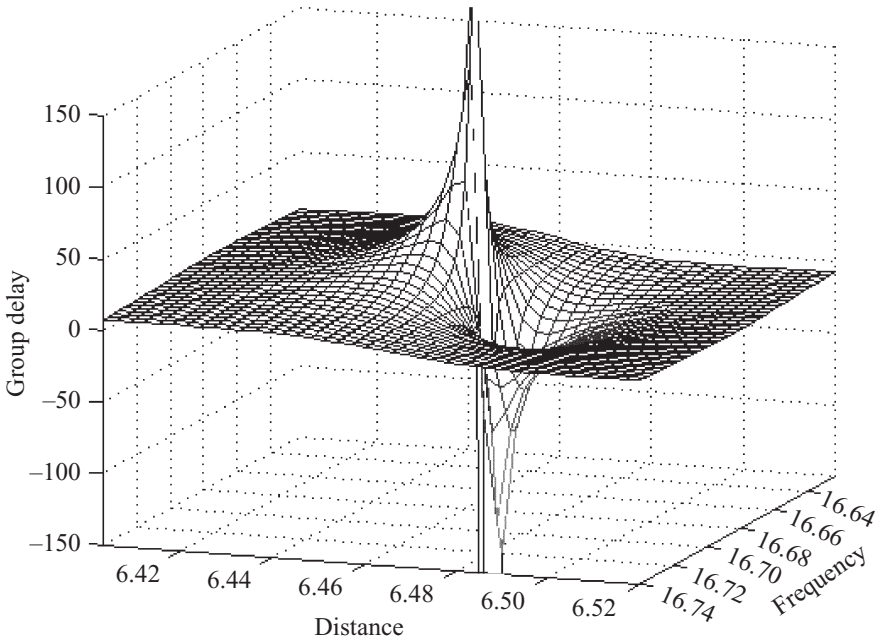


Figure 6.2.3 The group delay about a fade in the position and frequency plane. Two large spikes of opposite sign occur at the sides of the fade (Andersen et al., 1990)

that is, a fade (i.e. a minimum) in a 1D transfer function (distance or frequency) is seldom truly zero.

6.2.4 Spatial gradient of power

The spatial gradient of the power (Eggers *et al*., 2000) can also be found using the above application of the conditioning theorem. Using γ to donate the instantaneous power, $\gamma = |H|^2$,the gradient is

$$\gamma' = \left(H_{\mathrm{R}}^2 + H_{\mathrm{I}}^2\right)' = 2H_{\mathrm{R}}H_{\mathrm{R}}' + 2H_{\mathrm{I}}H_{\mathrm{I}}'. \tag{6.2.34}$$

This gradient can be with respect to distance or frequency. The distance gradient is of more interest because it provides information regarding the rate of power control, used in many systems, required for the fast fading. Using the known Gaussian distributions for $H_{\mathrm{R}}'|H$ and $H_{\mathrm{I}}'|H$ then the distribution for the power gradient, conditioned by the channel H, i.e. $\gamma'|H$, is

$$\gamma'|H \sim \mathcal{N}(0, 2Ps^2\gamma), \tag{6.2.35}$$

i.e. zero mean Gaussian with variance $2Ps^2\gamma$, where P is the total average power and s is the appropriate Doppler spread. The unconditional distribution is found from

$$p_{\gamma'}(\gamma') = \int_0^\infty p_{\gamma'}(\gamma'|\gamma)\, p_\gamma(\gamma)\, \mathrm{d}\gamma \tag{6.2.36}$$

and using the exponential distribution for the power,

$$p_\gamma(\gamma) = \frac{1}{P}\mathrm{e}^{-\gamma/P} \tag{6.2.37}$$

the result is

$$p_{\gamma'}(\gamma') = \frac{1}{2Ps}\mathrm{e}^{-|\gamma'|/Ps}, \tag{6.2.38}$$

which as a double-sided exponential, can be called a Laplacian distribution.

The cdf for the Laplacian distribution is

$$\begin{aligned} \mathrm{P}\left(\gamma' < \gamma_0'\right) &= \tfrac{1}{2}\mathrm{e}^{-|\gamma'|/Ps}, \qquad \gamma' \le 0 \\ &= 1 - \tfrac{1}{2}\mathrm{e}^{-|\gamma'|/Ps}, \qquad \gamma' \ge 0 \end{aligned} \tag{6.2.39}$$

and is plotted in Figure 6.2.4.

It is straightforward to find that the distribution of the derivative of the log of the spatial derivative is also student-T. The decibel power gradient is $\gamma'_{\mathrm{dB}} = (10\log_{10}\gamma)'$

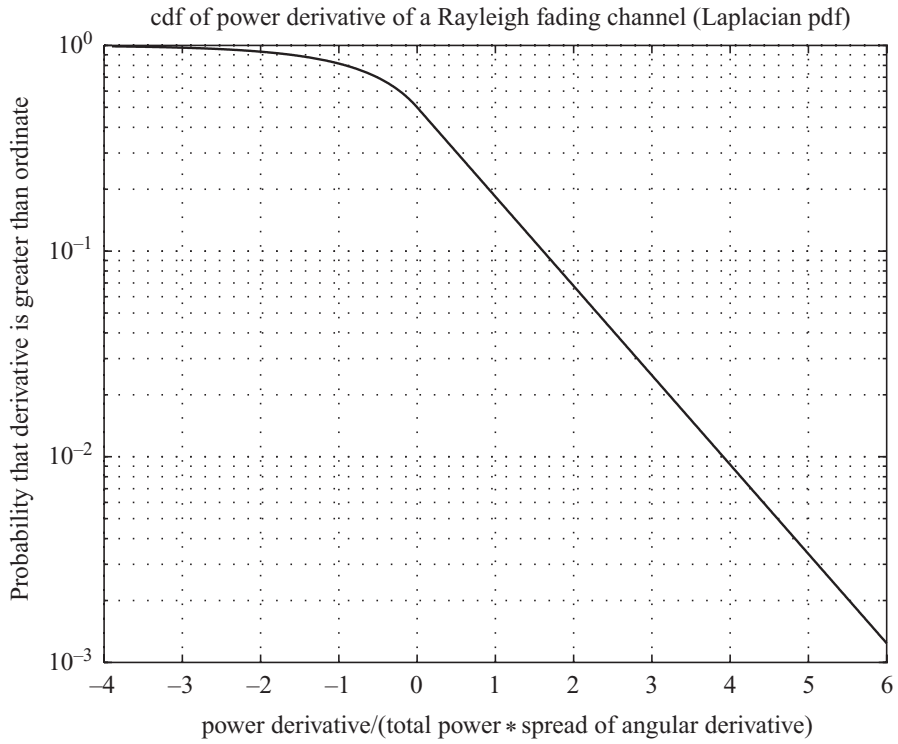


Figure 6.2.4 The cdf for the gradient of the power of a Rayleigh fading channel. The pdf is symmetric about zero, so the probabilities from both sides of the distribution can be found from the plot. On the abscissa the derivative is normalised by the total power and the appropriate Doppler spread

and the student-T pdf is

$$p_{\gamma_{\mathrm{dB}}}(\gamma_{\mathrm{dB}}) = \frac{1}{2}\frac{s_{\gamma\mathrm{dB}}^2}{\left(s_{\gamma\mathrm{dB}}^2 + \gamma_{\mathrm{dB}}'^2\right)^{3/2}}, \tag{6.2.40}$$

where $s_{\gamma\mathrm{dB}} = 2s_u 10/\ln 10\,\mathrm{dB\,m^{-1}}$ is the spread. The result fits with the same distribution of the phase derivative in the sense that the two quantities are the real and imaginary parts of the derivative of the log of the channel, namely

$$(\log H)' = \left(\tfrac{1}{2}\log(\gamma)\right)' + \mathrm{j}\phi'. \tag{6.2.41}$$

The relations between the envelope and phase are discussed in Section 7.7.

6.2.5 Dynamic two-path model

In the static model, the parameters a_i and τ_i were fixed. If these parameters become statistical quantities, varying in time, rather than constants, then the model is referred to as dynamic. This would be the situation in the above section where the delay and Doppler bins contributed sinc pulses with Rayleigh-distributed amplitudes and uniformly distributed phases, which change independently with time and mobile position. While this is referred to as the two-(dynamic) path model, the behaviour clearly belies that there are many sources contributing to the transfer function behaviour.

The dynamic two-path model normally refers to both paths being time-varying, i.e. different to the static two-path case, which is scaled to have one path fixed. We no longer have the relation linearly linking distance and position, $z = Vt$. For the two-path, a_1, a_2 and τ_1, τ_2 from the static model can be viewed as local, instantaneous values in the dynamic two-path model. The source positions (u_1, u_2) can be considered constant but the source amplitude, phase and delay change with time. The channel is written

$$\begin{aligned} h^{(2\mathrm{D})}(\tau, z, t) = {} & a_1(\tau(t,z))\mathrm{e}^{\mathrm{j}\alpha_1(t,z)}\mathrm{e}^{\mathrm{j}u_1 z}\delta(\tau - \tau_1(t,z)) \\ & + a_2(\tau(t,z))\mathrm{e}^{\mathrm{j}\alpha_2(t,z)}\mathrm{e}^{\mathrm{j}u_2 z}\delta(\tau - \tau_2(t,z)). \end{aligned} \tag{6.2.42}$$

The Fourier transform only has meaning at a fixed instant t_0;

$$h^{(2\mathrm{D})}(\tau(t = t_0), z) \Leftrightarrow H(\omega(t = t_0), z) \tag{6.2.43}$$

and so

$$\begin{aligned} H^{(2\mathrm{D})}(\omega, z, t_0) = {} & a_1(\tau(t_0, z))\,\mathrm{e}^{\mathrm{j}\alpha_1(\omega,z)}\mathrm{e}^{\mathrm{j}u_1 z}\mathrm{e}^{-\mathrm{j}\omega\tau_1(t_0,z)} \\ & + a_2(\tau(t_0, z))\,\mathrm{e}^{\mathrm{j}\alpha_2(\omega,z)}\mathrm{e}^{\mathrm{j}u_2 z}\mathrm{e}^{-\mathrm{j}\omega\tau_2(t_0,z)}. \end{aligned} \tag{6.2.44}$$

The statistics are typically a Rayleigh pdf for a_i over narrow spectral bins and unknown for τ_i in the sense that it relies on a given scenario. Specific models of the two-path profile, such as the GSM profiles, assign a specfic value for τ_2 of course. Averaging over (slow) time, t, or distance, z, the impulse response can be written using the power delay profile

$$P(\tau) = \langle |h^{(2D)}(\tau, z)|^2 \rangle_z = |h^{(2)}(\tau)|^2 = \sigma_1^2 \delta(\tau) + \sigma_2^2 \delta(\tau - \tau_2), \tag{6.2.45}$$

where τ_1 has been set to zero. Also, the variances of the Rayleigh envelopes, σ_1^2, σ_2^2, and the time delay difference are assumed as time invariant. The delay spread is then

$$s^{(2)} = \frac{c\tau_2}{1 + c^2} \tag{6.2.46}$$

as in the static scenario, and here $c = \sigma_1^2/\sigma_2^2$.

For computer simulations and visualisation, a possible scenario with a uniform angular distribution, which corresponds to the dynamic two-path model, comprises two rings of many time-varying sources with random phases (Clarke scenarios) located about the receiver position, depicted in Figure 6.2.5. The radial location of the rings is given by the delays τ_2 and τ_1. There are many physical paths that are summed for each delay to give the Rayleigh statistics at each of the two delay values. The receiver locus is small relative to the delay difference $\Delta\tau = \tau_2 - \tau_1$ in order to maintain an accurate two-path model with respect to the simulation bandwidth.

An example of the resulting channel is given in the 2D transfer function plot of Figure 6.2.6, in which $\sigma_1^2 = \sigma_2^2$, meaning that the coherence bandwidth is a minimum. The envelope follows the two-path periodic behaviour in frequency (because there

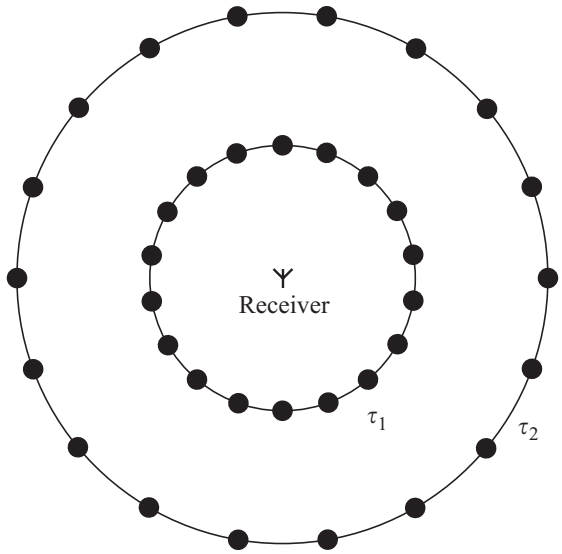


Figure 6.2.5 *The dynamic two-path model can be viewed as a scenario of two rings of many random sources about the receiver position. At a given time and position the receiver sees random signals at the two delay times, τ_1 and τ_2*

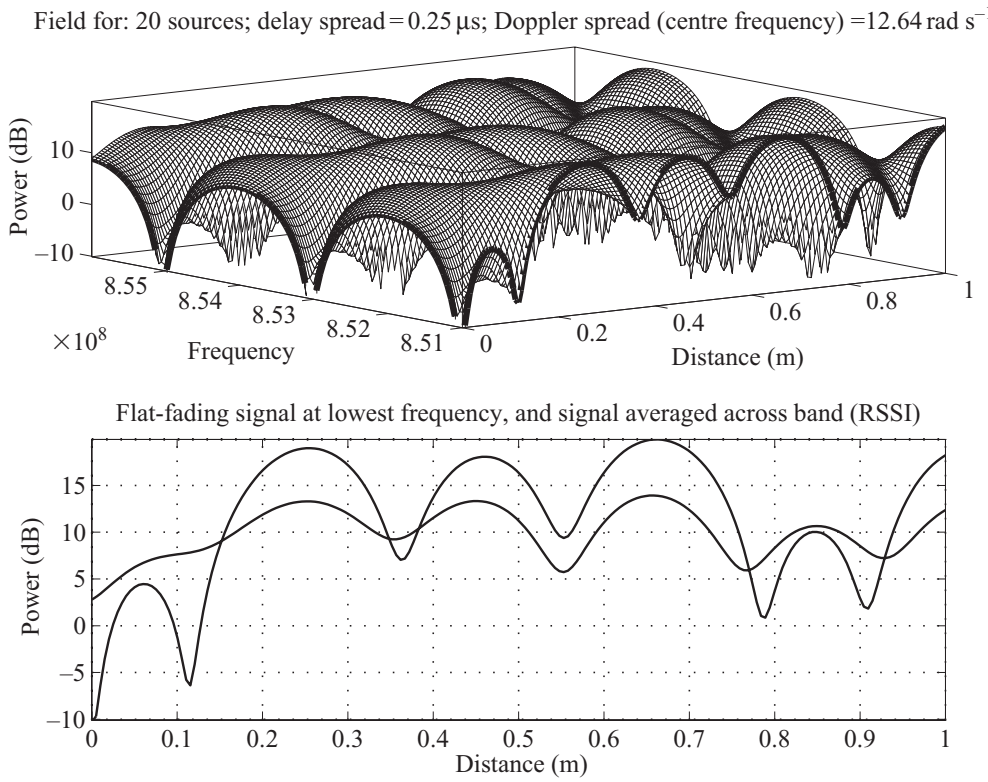


Figure 6.2.6 The envelope of the transfer function for a dynamic two-path model. The time delay difference is 0.5 μs giving the 2 MHz frequency period, and the envelope is Rayleigh in the spatial coordinate

are only two signal positions in delay time), and is Rayleigh distributed in distance. The sources were static for generating the plot, and the difference in the two delays is again 0.5 μs, giving the frequency periodicity in the transfer function of 2 MHz.

6.3 Doppler moments

The quasi-periodicity of the Rayleigh-distributed envelope, for a fixed frequency, is related to the Doppler spread. The spread is affected by the distribution of the effective sources, i.e. the product of the antenna pattern and the distribution of incoming waves. In the Clarke scenario which is used in the example, the angular distribution of sources is uniform,

$$p_\theta(\theta) = \frac{1}{\pi}, \quad |\theta| \leq \pi, \tag{6.3.1}$$

and with $\theta = \cos^{-1}(u/k)$, the pdf for the spatial Doppler has the same form as that for the two-path magnitude,

$$p_u(u) = \frac{1}{\pi\sqrt{k_C^2 - u^2}}, \quad |u| \leq k_C, \tag{6.3.2}$$

and this is the normalised form of the quantity $P(u) = \langle |a(u)|^2 \rangle = |a(u)|^2$ from Chapter 2. (No averaging is required since both delay bins have the same angular distribution in this model.) The cdf follows as

$$P_u(u \le u_0) = \int_{-k_C}^{u_0} p_u(u)\,\mathrm{d}u = 1 - \frac{1}{\pi}\cos^{-1}\left(\frac{u_0}{k_C}\right). \tag{6.3.3}$$

6.3.1 Basic definition

The sign convention is depicted in Figure 6.3.1. The angular range in the scenario is $[\theta_L, \theta_U]$. For $u_U = k_C \cos\theta_U$ and $u_L = k_C \cos\theta_L$, the range in the Doppler domain is written $[u_U, u_L]$, i.e. the positive range is $\Delta u = u_L - u_U$.

To illustrate the Doppler moments, the two-path model is recalled. For two waves arriving from u_1 and u_2, here with the same amplitude, the pdf is

$$p_u^{(2)} = \tfrac{1}{2}\left(\delta(u - u_1) + \delta(u - u_2)\right), \tag{6.3.4}$$

which gives a mean of

$$\bar{u}^{(2)} = \frac{u_1 + u_2}{2} \tag{6.3.5}$$

and a spread of

$$s_u = \frac{1}{2}\sqrt{u_1 - u_2}. \tag{6.3.6}$$

The expression for the mean shows that it is defined only in the half-plane as expected. (For modelling of both planes, the sources in one plane are reflected onto the other plane, and the mean is defined in that plane.) The mean is zero for symmetrically

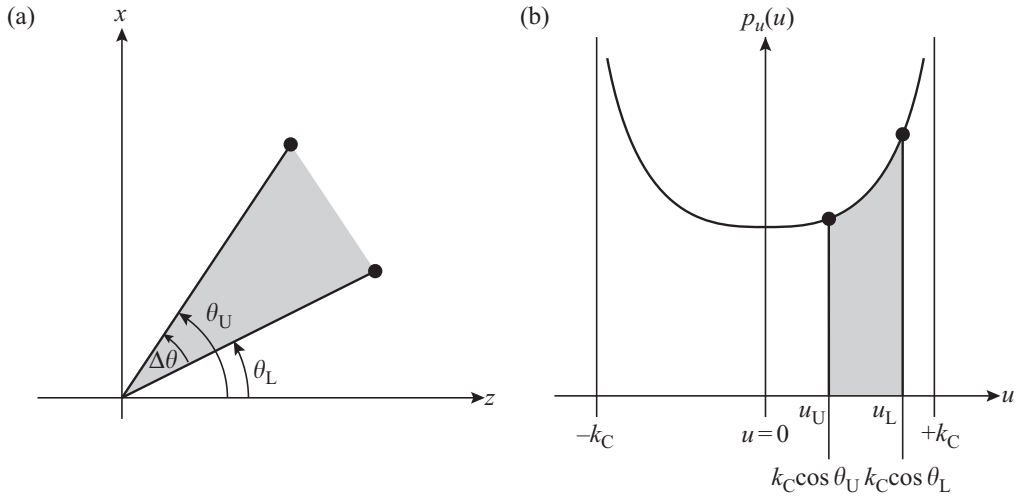


Figure 6.3.1 Sign conventions for the spatial and Doppler distributions. The mean direction is kept less than half of the angular range in order to confine the directions to be between 0 and π

placed sources, $u_1 = -u_2$. The spread is zero when there is only a single point source, i.e. when $u_1 = u_2$, and the maximum value of the spread occurs when the two static scenario sources are in the forward and backward directions, $u_1 = -u_2 = k_C$, and here the spread is k_C.

For the uniform scenario, the mean spatial Doppler shift is zero and so the Doppler spread is equal to just the square root of the mean square Doppler frequency,

$$\begin{aligned} s_u^{(U)} &= \left(\int_{-k_C}^{k_C} u^2 p_u(u)\,\mathrm{d}u - \left(\int_{-k_C}^{k_C} u p_u(u)\,\mathrm{d}u \right)^2 \right)^{1/2} \\ &= \left(\frac{1}{\pi} \int_{-k_C}^{k_C} \frac{u^2}{\sqrt{k^2 - u^2}} \mathrm{d}u \right)^{1/2} \\ &= \frac{k_C}{\sqrt{2}}. \end{aligned} \tag{6.3.7}$$

If the Doppler spectrum is truncated, i.e. only incident power from an angular range $[\theta_L, \theta_U]$ is considered, then the Doppler spread can be related to the angular quantities as follows. By defining a mean directional angle,

$$\theta_0 = \frac{\theta_U + \theta_L}{2} \tag{6.3.8}$$

and the angular range

$$\Delta\theta = \theta_U - \theta_L \tag{6.3.9}$$

$$s_u = \tfrac{1}{2}\sqrt{u_1 - u_2}, \tag{6.3.10}$$

and confining the range such that $\theta_0 \leq \Delta\theta/2$ then the moments are

$$\langle u \rangle = \int_{u_U}^{u_L} u p_u(u)\,\mathrm{d}u = \frac{2k_C}{\pi} \cos\theta_0 \sin\left(\frac{\Delta\theta}{2}\right), \tag{6.3.11}$$

and

$$\langle u^2 \rangle = \int_{u_U}^{u_L} u^2 p_u(u)\mathrm{d}u = \frac{k_C^2}{2\pi}(\cos 2\theta_0 \sin\Delta\theta + \Delta\theta), \tag{6.3.12}$$

and the spatial Doppler spread is

$$s_u^{(T)} = \frac{k_C}{\sqrt{2}} \left(\frac{1}{\pi} \cos 2\theta_0 \sin\Delta\theta + \frac{\Delta\theta}{\pi} - \frac{8}{\pi^2} \cos^2\theta_0 \sin^2\left(\frac{\Delta\theta}{2}\right) \right)^{1/2}. \tag{6.3.13}$$

This corresponds to the case where the incident angular power is confined to be uniformly distributed between the limiting angles $[\theta_L, \theta_U]$ and this could approximate a base station being illuminated by the scatterers about a transmitting mobile. Similarly, it could be from an idealised directional 'sector' antenna in the presence of an omnidirectional scenario. The Doppler spread, normalised by the delay spread of the uniform scenario (this is the square root term in the equation) is given in Figure 6.3.2.

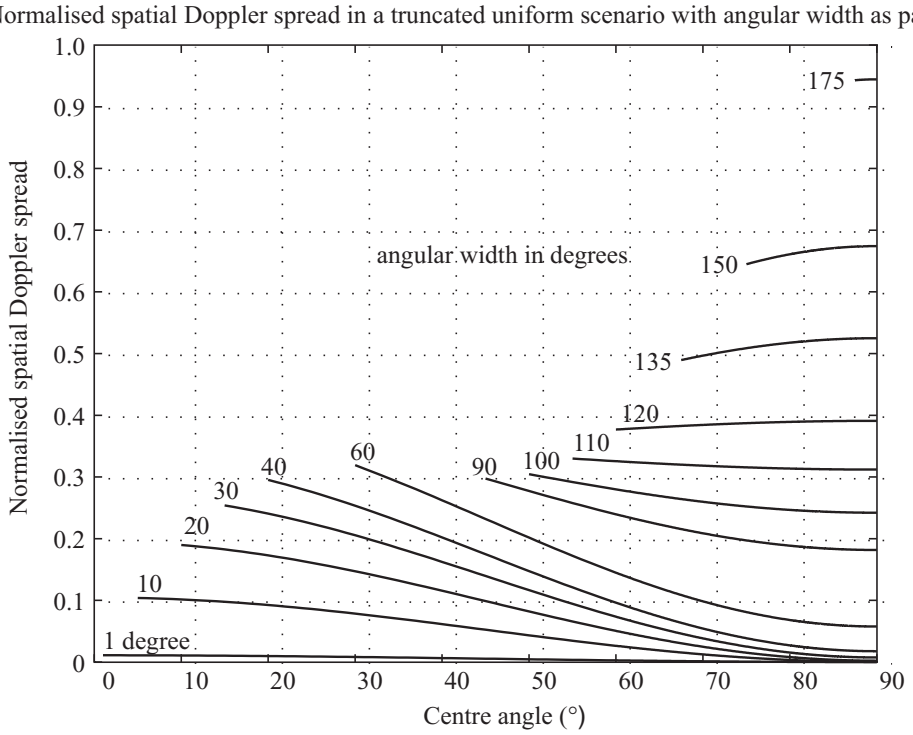


Figure 6.3.2 The normalised Doppler spread in a sectorised angular range against the mean angle. The mean angle is constrained to be within half of the angular range. The normalisation is with respect to the Clarke (uniform) scenario

For the case where the angular range is greater than 180°, then the moment defining relations alter owing to the circular nature of the support of the pdfs. If the range is now, in linear-angle terms, $\Delta\theta = [\theta_{\mathrm{U}}, \theta_{\mathrm{L}}] > \pi$, then the u-range *is* expressed as $\Delta u = [-k_{\mathrm{C}}, k_{\mathrm{C}} \cos\theta_{\mathrm{U}}] + [k_{\mathrm{C}} \cos\theta_{\mathrm{L}}, k_{\mathrm{C}}]$. This range is the non-shaded part in Figure 6.3.1. The moments are then expressed as, for example for the mean angle,

$$\begin{aligned}\langle u \rangle &= \int_{-k_{\mathrm{C}}}^{k_{\mathrm{C}} \cos\theta_{\mathrm{U}}} u \; p_u(u)\, \mathrm{d}u + \int_{k_{\mathrm{C}} \cos\theta_{\mathrm{L}}}^{k_{\mathrm{C}}} u \; p_u(u)\, \mathrm{d}u \\ &= \int_{-k_{\mathrm{C}}}^{k_{\mathrm{C}}} u \; p_u(u)\, \mathrm{d}u + \int_{u_{\mathrm{L}}}^{u_{\mathrm{U}}} u \; p_u(u)\, \mathrm{d}u, \end{aligned} \tag{6.3.14}$$

or

$$\langle \theta \rangle = \int_{0}^{\pi} \theta p_\theta(\theta)\, \mathrm{d}\theta + \int_{\theta_{\mathrm{U}}}^{\theta_{\mathrm{L}}} \theta p_\theta(\theta)\, \mathrm{d}\theta. \tag{6.3.15}$$

The Doppler spread is plotted against the mean angle with the angular range as a parameter, in Figure 6.3.2.

6.3.1.1 Spatial periodicity of transfer function

The quasi-periodicity of the Rayleigh distributed envelope in the space domain is found from the autocorrelation of the envelope, $R_{|H|}(\Delta z)$, given via the Fourier transform of the angular power distribution (Appendix B, where the notation $S_H(\Delta z)$ is used)

$$R_H(\Delta z) \Leftrightarrow P(u). \tag{6.3.16}$$

This gives the well-known result for the autocorrelation coefficient function of the power (envelope squared), $|H(\Delta z)|^2$, as $J_0^2(k_C \Delta z)$ for the Clarke scenario. Substituting for the Clarke scenario Doppler spread, $s_u^{(U)} = k_C/\sqrt{2}$, this can be written

$$\rho_{|H|}(\Delta z) \approx J_0^2\left(\sqrt{2} s_u \Delta z\right) \qquad \text{(Clarke scenario)}. \tag{6.3.17}$$

The correlation distance to the first maximum of J_0^2 away from the origin is $k_C \Delta z \approx 3.84$ and this gives a quasi-periodicity of the Rayleigh signal as $3.84/(2\pi) \sim 0.6\lambda$. For large correlation displacements, the quasi-periodic form of $J_0^2(k_C \Delta z)$ approaches $\cos^2(k_C \Delta z)$, i.e. the periodicity becomes $\Delta z \approx 0.5\lambda$. The delay spread in the above example is 12.6 μs, and adjacent fades are separated on average by about $0.6\lambda = 0.21$ m at 851 MHz. Figure 6.3.3 depicts the Doppler spectrum and correlation coefficient functions for the two-path and Clarke models.

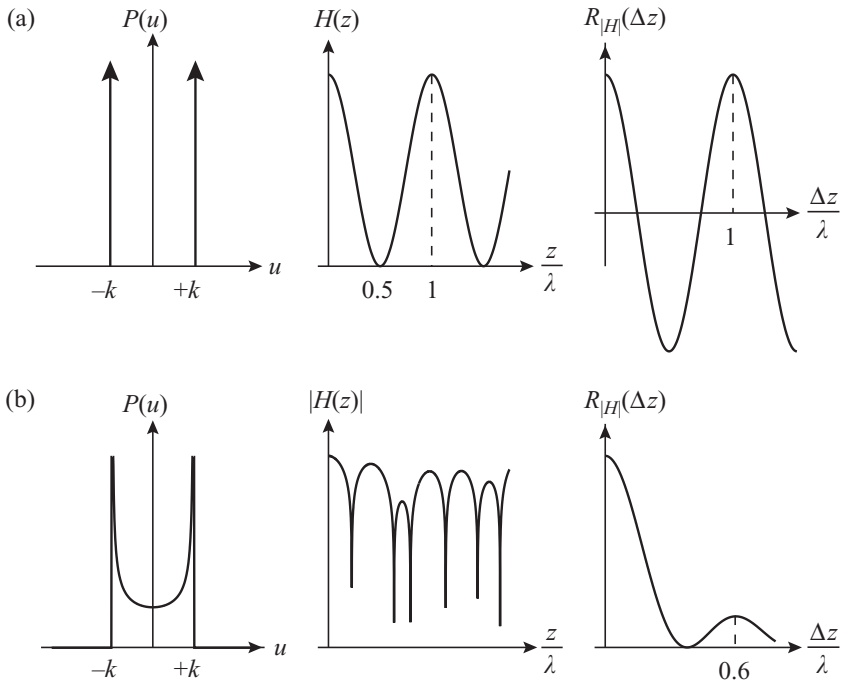


Figure 6.3.3 Examples of the Doppler power profile, the transfer function and the spatial correlation function. (a) A two-path model with $a_1 = a_2, \theta_1 = 0, \theta_2 = \pi$; (b) the uniform scenario

6.3.2 Angle of arrival from a cluster of scatterers

The phase gradient statistics for the random FM and the group delay have another application, that of the angle-of-arrival (AOA) at an array. The AOA has become of interest for adaptive base stations, where the beamforming allows increased gain and interference reduction between users. Instead of classical point-to-point situations, the signals arrive from many points, or clusters of scatterers. The detection of the AOA from the cluster can be used for directing the downlink beam towards the cluster associated with the wanted terminal. The statistics of the detected AOA, from an array which cannot resolve all the components of the cluster, are therefore of interest for the system performance.

The situation is depicted in Figure 6.3.4. The cluster of scatterers has a nominal centre angle of α_C, and this is measured from the broadside of the array antenna, following convention. The beamwidth of the antenna pattern is denoted θ_B. When θ_B is much smaller than the spread of the cluster, s_u, and depicted as θ_C in the figure, then the antenna can resolve 'subclusters' of the scatterers. The situation here of interest, however, is when $\theta_B \gg \theta_C$, so the cluster appears in the beam almost as a Rayleigh fading point source.

At a given position of the mobile a snapshot is taken and the array sees a plane wave with a given amplitude and a linear phase gradient across the aperture. The phase variation is assumed linear and any algorithm for finding the AOA will use this gradient. As the mobile terminal moves, the scatterers within the cluster undergo random variations. After several snapshots a distribution of the AOA can be built up and a mean direction found. The AOA distribution will be different from the incident

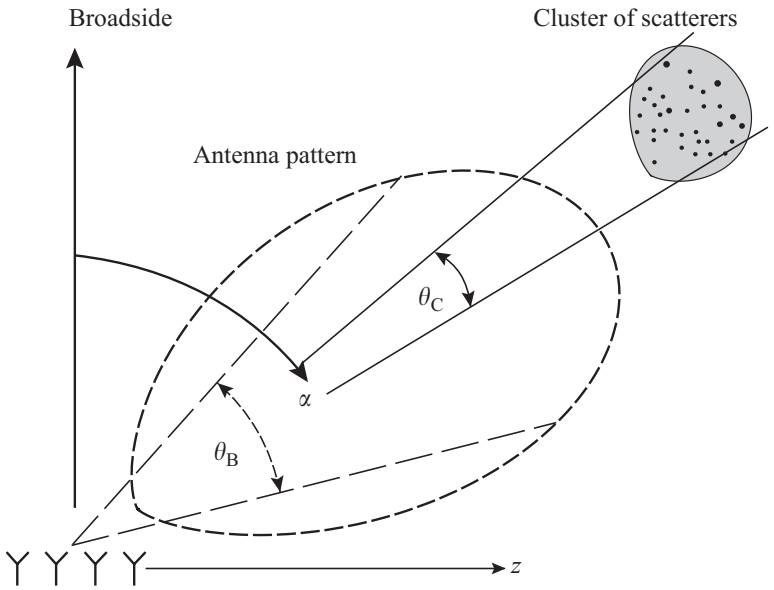


Figure 6.3.4 A cluster of scatterers illuminating an array antenna. The spread of the cluster is shown as θ_C. The nominal centre of the cluster is measured from broadside, as α_C

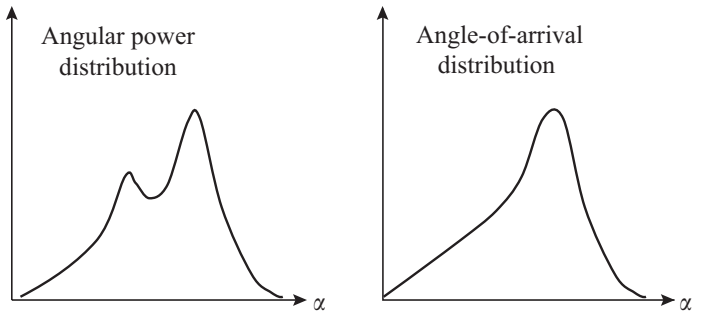


Figure 6.3.5 An angular power spectrum of the incident waves at the antenna will be different from the calculated AOA distribution

power distribution, as depicted in Figure 6.3.5. The azimuth power spectrum depends on the environment, and in principle can take many different forms. As an example, Pedersen *et al.* (1987) have approximated the distribution as Laplacian. Only for an array with infinite resolving power will the AOA distribution be the same as the angular power spectrum. For the limited resolution of a finite aperture array, the AOA distribution is a filtered version of the angular power distribution.

The incident waves will vary along the array in the z-direction as $e^{\mathrm{j}k_C z \sin\alpha}$, so the phase gradient along the array is $k_C \sin\alpha$. Thus we can use the results for the phase derivatives in the previous sections. From the student-T distribution for $u = k_C \cos\theta$ (here, θ is the angle from the z-axis, so $\cos\theta = \sin\alpha$), recalled as

$$p_u(u) = \frac{1}{2} \frac{s_u^2}{(s_u^2 + (u - \bar{u})^2)^{3/2}} \tag{6.3.18}$$

transforming to a function of α gives

$$p_\alpha(\alpha) = \frac{1}{2} \frac{s_u^2 \cos\alpha}{(s_u^2 + (\sin\alpha - \overline{\sin\alpha})^2)^{3/2}} \tag{6.3.19}$$

but this needs a normalisation to the real angles within $\pm\pi/2$, which results in

$$p_\alpha(\alpha) = \frac{\frac{1}{2} s_u^2 \cos\alpha / (s_u^2 + (\sin\alpha - \overline{\sin\alpha})^2)^{3/2}}{((1 - \overline{\sin\alpha})/(s_u^2 - (1 - \overline{\sin\alpha})^2)^{1/2} + (1 + \overline{\sin\alpha})/(s_u^2 - (1 + \overline{\sin\alpha})^2)^{1/2})}. \tag{6.3.20}$$

The normalisation factor in the denominator is slightly less than 1. This arises because the pdf in u-space actually features a small probability that the phase gradient is so large that it corresponds to $|\sin\alpha| > 1$.

In a similar way to the other phase gradients for the channel (random FM and the group delay), the distribution of the instantaneous AOA depends only on the mean and

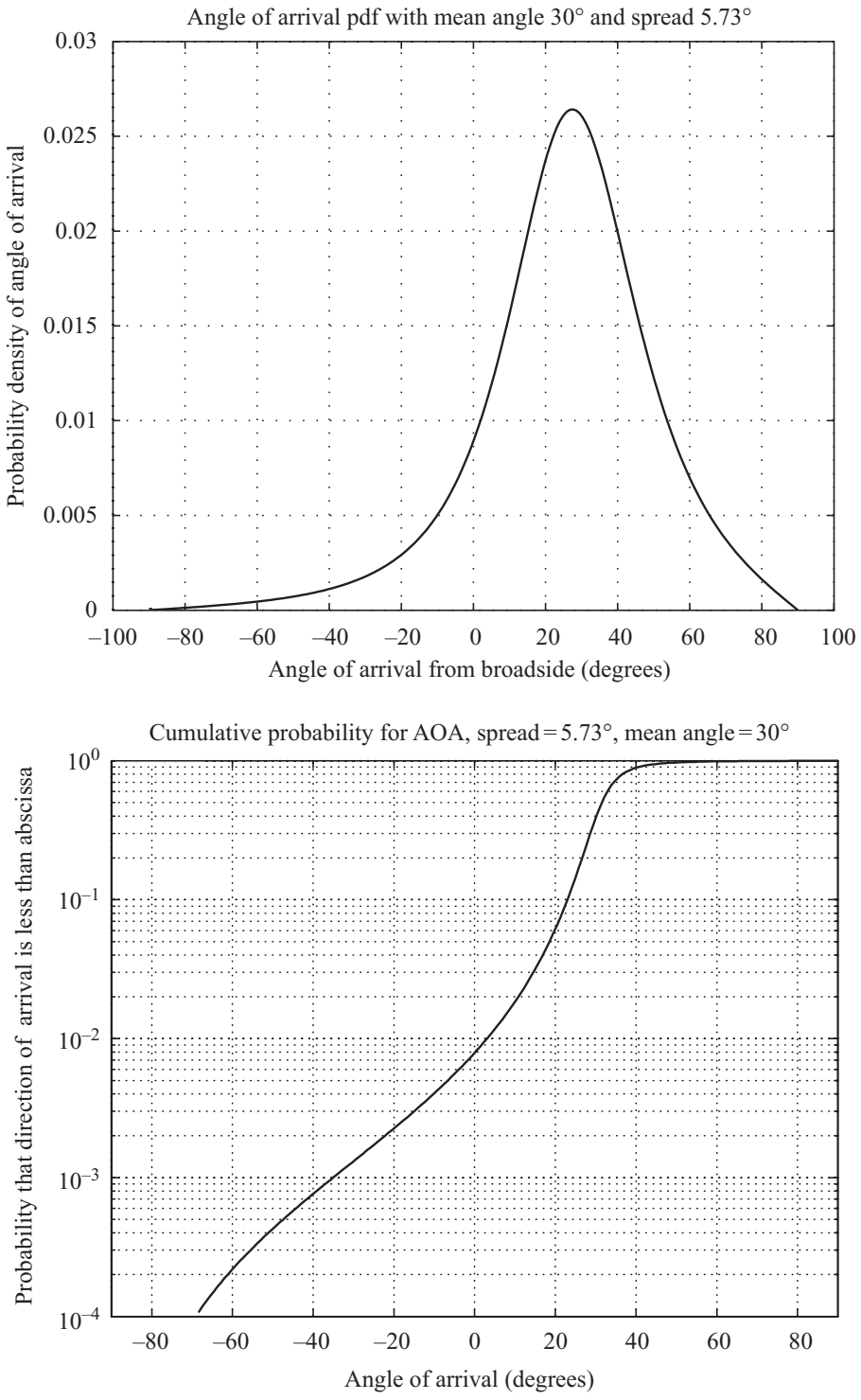


Figure 6.3.6 *(a) The pdf of the instantaneous angle of arrival for a power spectrum of incident waves with spread $s_u = 0.1$ (5.7°) and a mean angle $\alpha_C = 30°$. (b) The cdf*

variance of the power spectrum of the incident waves, and not on the power spectrum shape. An example of the pdf is given in Figure 6.3.6.

The student-T distribution of instantaneous AOAs has long tails depending on the angular spread. The finite probability that $|\sin\alpha| > 1$, corresponding to angles outside of real space, is $0.5s_u^2$. Averaging over several samples is necessary to avoid large errors in the AOA.

6.3.2.1 Conditional probabilities

The probability of the AOA, u, conditioned by its power is from above (repeating equation (6.2.24)),

$$p_u\left(u\middle||H|^2\right) = \frac{|H|}{\sqrt{2\pi}s_u}\mathrm{e}^{-((u-\bar{u})^2|H|^2)/(2s_u^2)}, \tag{6.3.21}$$

and integrating this over the exponential distribution for the observed power gives $p_u(u)$ above. Transforming (see Appendix B), the power conditioned by the AOA is (Andersen *et al.*, 2001)

$$p\left(|H|^2\,\middle|\,u\right) = \frac{|H|\,\left((u-\bar{u})^2+s_u^2\right)^{3/2}}{\sqrt{2\pi}s_u^3}\exp\left[-\frac{(u-\bar{u})^2|H|^2}{(2s_u^2)}-\frac{|H|^2}{2}\right], \tag{6.3.22}$$

from which the mean power conditioned on AOA, the observed power spectrum, is

$$\langle|H|^2\ \ |\ u\rangle = \frac{s_u^2}{(u-\bar{u})^2+s_u^2}. \tag{6.3.23}$$

Similarly to the unconditioned observed quantities, the conditional observed power spectrum depends only on the mean, $\bar{u}$, and the spread, s_u, of the incident power distribution, $P(u)$. The power quantity is often more convenient for measurements. Experimental results are given in Chapter 4.

6.4 Correlation spacings

The correlation spacing in the frequency domain is traditionally referred to as the 'correlation spacing' or 'correlation bandwidth', whereas in the spatial domain, the terms 'decorrelation distance' and 'correlation distance' are used.

6.4.1 Correlation distance–Doppler spread product

In Section 5.2, the coherence bandwidth in hertz (B)–delay spread product was quoted as following Fleury's relation,

$$B\cdot s \geq \frac{1}{2\pi}\cos^{-1}C_C \tag{6.4.1}$$

where C_C is the value of the correlation coefficient defining the coherence bandwidth. An analogous relation occurs for the correlation distance–Doppler spread product.

6.4.1.1 Correlation distance

Similar to the coherence bandwidth, the *correlation distance* d_d is a rather loosely defined quantity. It often refers to where the correlation coefficient function of the envelope drops to 0.7. For the uniform scenario, for example, the definition is from

$$\rho^2_{|H|^2}(\Delta z = d_d) = |\rho_H(\Delta z = d_d)|^2 = J_0^2\,(k_{\mathrm{C}} d_{\mathrm{d}}) = 0.7 \Rightarrow d_d \approx 0.13\lambda. \quad (6.4.2)$$

This definition for d_d arises from spatial diversity considerations for omnidirectional antennas in a uniform scenario. (It corresponds to about a 1 dB loss in the average SNR of the combined two-branch signal, relative to the average SNR of the combined uncorrelated ($\rho = 0$) branches (Pierce and Stein, 1960).) The correlation distance depends on the angular distribution of sources $P(u)$ since this defines (via Fourier transform) the autocorrelation function. For signals received by an antenna, $R_H(\Delta z)$ refers to the radio channel instead of the propagation channel, and the function is modified by the antenna pattern, which in turn modifies the correlation distance.

6.4.1.2 Uncertainty principle

A form of correlation distance can also be defined by the spread of the spatial correlation. The product of this and the angular spread is then lower bounded from the uncertainty principle. If d_d is taken as the spatial spread (the standard deviation of $\rho_{|H|^2}(\Delta z)$, namely, define here $d_d = \sigma_{\Delta z}$), then the usual uncertainty principle, for real functions and unbounded support, is (Papoulis, 1977)

$$\sigma_{\Delta z} \cdot s_u \geq \tfrac{1}{2}, \quad (6.4.3)$$

where s_u is the spread of the Doppler profile with the equivalence holding for Gaussian forms.

6.4.1.3 A pattern for minimum spatial correlation function spread

The Doppler profile $P(u)$ has bounded support, given by $[-k_{\mathrm{C}}, k_{\mathrm{C}}]$, in a static environment. The uncertainty principle in this case takes a form given by (Papoulis, 1977) which is

$$k_{\mathrm{C}} \cdot \sigma_{\Delta z} \geq \frac{\pi}{2} \quad (6.4.4)$$

with the equality holding for

$$P(u) = p_u(u) = K \cos\left(\frac{\pi u}{2k_{\mathrm{C}}}\right), \quad |\theta| \leq \frac{\pi}{2}; \quad (6.4.5)$$

where K is the normalising constant. So with k_{C} fixed, this Doppler profile provides the minimum spread of the spatial correlation function, and this is

$$\sigma^{(\min)}_{\Delta z} = \frac{\lambda_{\mathrm{C}}}{4}. \quad (6.4.6)$$

The angular power distribution, as a function of angle θ, follows as

$$p_\theta(\theta) = p_u(u)\left|\frac{\partial u}{\partial \theta}\right| = K\ k_C \cos\left(\frac{\pi}{2}\cos\theta\right)|\sin\theta|, \quad |u| \le k_C; \tag{6.4.7}$$

which is a dipole-like pattern (the power pattern of a small dipole is in fact $\sin^2\theta$a; see Chapter 8) with its nulls along the direction of motion. Recall that since this represents the power incident from the effective sources, the distribution can be interpreted as an antenna power pattern in the presence of an omni scenario, or as a directional scenario with an omni antenna pattern, or as a product of the pattern and the scenario of sources.

Recall that the spread is not the same as the correlation distance, but is similar in the sense that both parameters represent a measure of the extent of the correlation function. The above pattern gives a minimum spread, but does not give the minimum correlation distance as defined by a displacement distance at which the correlation coefficient function drops to a given value. The minimum such correlation distance would result from two narrow beams at $u = \pm k_C$, i.e. along the directions of motion, virtually the opposite to the pattern above! This pattern is like the two-path model, and while the central lobe of the spatial correlation function is of minimal width (and so the correlation distance is minimum) the spatial correlation is also periodic and the spread is interpreted as being infinite. In such a situation, the periodicity offers an alternative measure for the spreading of the correlation function.

6.4.1.4 Spatial periodicity

An alternative parameter for the spreading of the spatial correlation function, which can be convenient for certain scenarios, is to make d_d the periodicity (cf. that resulting from the two-path model) or quasi-periodicity (cf. the Clarke scenario) of the spatial correlation behaviour of the fields.

The maximum value of the Doppler spread is $s_u^{(\max)} = k_C$, occurring for the two-path model where $a_1 = a_2$, $u_1 = -u_2 = k_C$. This results in the minimum periodicity, defined here as d_d, which is $\lambda/2$ for the first peak in the periodic correlation coefficient function. So for this two-path model, $d_d \cdot s_u^{(2)} = k_C \cdot \lambda/2 = \pi$. For the uniform Clarke scenario, this form of the product is $d_d \cdot s_u^{(U)} = 0.6\lambda \cdot k_C/\sqrt{2} = 2.67$.

6.4.1.5 Summary

The spread, or standard deviation (as defined in the above discussions), correlation distance and periodicity are different measures of the spread of correlation-type functions. These measures minimise for different forms of the function as seen above. For diversity considerations, the correlation distance, as long as it is clearly defined (see below), is the most useful since it is a direct measure of the average spacing required for 'uncorrelated' channels. However, even this form can represent a poor or ambiguous definition for certain function behaviour of certain scenarios, and the chosen correlation values, as discussed in the following section.

6.4.2 Correlation distance for directional antennas or scenarios

The spatial correlation distance for a given incident angular distribution of effective sources can be found directly from the Fourier relation between the correlation function $R_H(\Delta z)$ of the channel $H(z)$ and the Doppler power profile,

$$R_H(\Delta z) \Leftrightarrow P(u). \tag{6.4.8}$$

It is emphasised that the Doppler profile is for the effective sources, i.e. the antenna pattern is included. The relationship can therefore be used for solving diversity spacing requirements of omnidirectional antennas in the presence of directional scenarios (e.g. vehicular mobile cases) or for directional antennas in the presence of a constant angular power density, or a combination of these (e.g. a base-station case).

For either directional scenarios or directional antennas it is convenient to use a circular function, and the $\cos^n\theta$ form, originally proposed by Lee (1973), is the simplest. For numerical problems with large n, the angle can be divided by a factor and the range of the argument correspondingly reduced to allow n to remain a manageable size. For very narrow beams, the asymptotic form for large n can be found from using the small θ approximation $\ln(\cos\theta) \approx \ln\left(1-\theta^2/2\right) \approx -\theta^2/2$, which results in $\cos^n\theta \approx \exp\left(-n\theta^2/2\right)$, i.e. the cosine form becomes similar to Gaussian.

The incident angular power density is written

$$p_\theta(\theta) = Kg^2(\theta)\cos^{2n}\left(\frac{\theta-\theta_0}{2}\right), \quad |\theta-\theta_0| \le \pi, \tag{6.4.9}$$

where the cos form is the angular power density (directional scenario) and $g(\theta)$ is the receiving amplitude pattern (or vice versa), K is a scaling constant, and θ_0 is the mean (i.e. peak) angle of power angular density. θ and θ_0 are zenith angles measured from the direction of travel. From now on $g(\theta)$ is taken as omnidirectional and the $\cos^n$ pattern represents the power density from the effective sources. If $g(\theta) = 1$, then

$$\begin{aligned} K &= \frac{1}{\pi}\frac{(2n)!!}{(2n-1)!!} = \frac{\Gamma(n+1)}{\sqrt{\pi}\Gamma(n+1/2)}, \\ &= \frac{1}{\pi}, \quad n = 0. \end{aligned} \tag{6.4.10}$$

An interpretation of the situation is given in Figure 6.4.1.

For an omni antenna operating in the directional scenario (or vice versa), the Doppler profile is

$$\begin{aligned} P(u) = p_u(u) &= p_\theta(\theta)\left|\frac{\partial\theta}{\partial u}\right| \\ &= \frac{K}{\sqrt{k_C^2-u^2}}\cos^{2n}\left(\frac{\cos^{-1}(u/k_C)-\cos^{-1}(u_0/k_C)}{2}\right), \end{aligned} \tag{6.4.11}$$

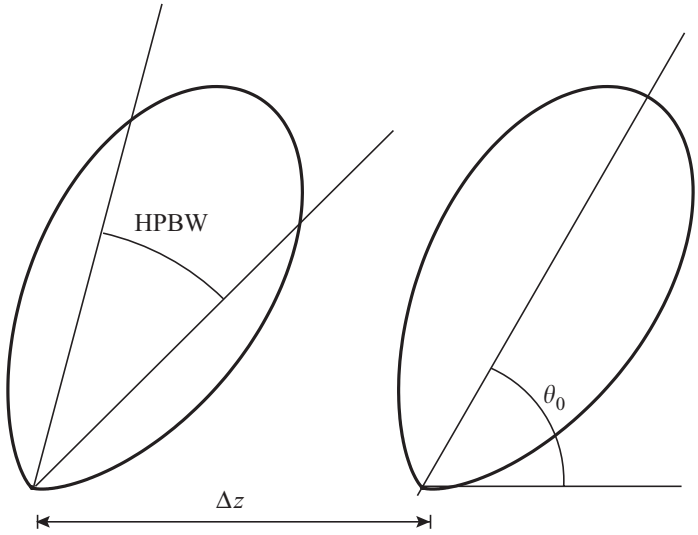


Figure 6.4.1 *An example of the* $\cos^{2n}((\theta - \theta_0)/2)$ *angular power density, depicted as a directional antenna pattern with the z-axis the direction of travel and* θ_0 *the mean angle measured from the z-axis*

where again K is a scaling constant, $u_0 = k_C \cos\theta_0$, and k_C is recalled as the wavenumber for the carrier frequency. Once the angular power density $p_\theta(\theta)$ is established, the spatial correlation function is found from the Fourier transform

$$R_H(\Delta z) = \int_{-k_C}^{k_C} \frac{K}{\sqrt{k_C^2 - u^2}} \cos^{2n}\left(\frac{\cos^{-1}(u/k_C) - \cos^{-1}(u_0/k_C)}{2}\right) e^{-ju\Delta z}\, du. \tag{6.4.12}$$

Singularities at $u = \pm k_C$ (only occurring for $n = 0$) can be handled by their analytic integration, and numeric integration is not required for $n = 0$ or for any n with $\theta_0 = \pi/2$. The correlation is purely real when the beam maximum angle $\theta_0 = \pi/2$ or $n = 0$ (see below).

In summary, the above equation is the correlation between the voltages from antennas whose patterns are common and given by $g(\theta)$, but are separated by a distance Δz.

Another way to find the correlation function is to state directly the correlation between the spaced antenna patterns for a given angular power distribution. The pattern from an identical antenna spaced at a distance Δz (see Figure 6.4.1) can be written

$$g_2(\theta) = g_1(\theta) e^{jk\Delta z \cos\theta}. \tag{6.4.13}$$

Then the correlation between the voltages for the 2D case can be written

$$R_H(\Delta z) = \langle g_1(\theta) g_2(\theta) \rangle = \int_0^{2\pi} p_\theta(\theta) g_1(\theta) g_2^*(\theta)\, d\theta, \tag{6.4.14}$$

which gives the same results as above. This approach is more general, however, for diversity, because the patterns do not need to be the same. This is taken further in Chapter 7.

For measured patterns and most modelled patterns, the integral is solved numerically. However, the choice of pattern modelling function can be arranged to give simplified results. For example, in the special case where the patterns are broadside ($\theta_0 = \pi/2$), and using a pattern function of $g(\alpha) = \cos^n(\alpha)$, where $\alpha = \theta - \pi/2$ is the angle from broadside, over the half-space only, gives the standard integral

$$\begin{aligned} R_H(\Delta z) &= \int_{-\pi/2}^{\pi/2} \cos^{2n}(\alpha) e^{\mathrm{j}k_C \Delta z \sin\alpha}\, \mathrm{d}\alpha \\ &= \sqrt{\pi} 2^n \Gamma\,(n + 1/2)\, \frac{J_n(k_C \Delta z)}{(k_C \Delta z)^n}. \end{aligned} \tag{6.4.15}$$

However, for finding the correlation spacing for antennas, as given below, the inverse of the normalised form of R_H is required, and this has to be undertaken numerically. So closed form solutions for correlation distances are not normally possible.

In general, the correlation coefficient of the Gaussian channel is complex, denoted

$$\rho_H(\Delta z) = \frac{R_H(\Delta z) - \langle H \rangle^2}{R_H(0) - \langle H \rangle^2} = \frac{R_H(\Delta z)}{R_H(0)} = \rho_{II}(\Delta z) - \mathrm{j}\rho_{IQ}(\Delta z). \tag{6.4.16}$$

As noted above, an imaginary part arises when the mean angle is non-zero, i.e. when the Doppler spectrum is asymmetric, and this phenomenon is clear from the Fourier transform relationship. The relative forms of the real and imaginary components produces the form of the power correlation. The correlation coefficient of the power signal is

$$\rho_{|H|^2} = |\rho_H|^2 = \rho_{II}^2 + \rho_{IQ}^2. \tag{6.4.17}$$

A comparision of the $\rho_{|H|^2}$ and the $|\rho_H|^2$, calculated over finite samples, can be found by simulation, and an example is given below in Figure 6.4.3. M. Leonard-Taylor is acknowledged for assistance with this simulation (Leonard-Taylor, 1997). An example of the modelled Doppler and correlation functions is given in Figure 6.4.2 where ρ_{II}, ρ_{IQ} and $\rho_{|H|^2} = |\rho_H|^2$ are given for a directional scenario with half power beamwidths (HPBWs) of the $\cos^n$ amplitude pattern, given by

$$HPBW = 4\cos^{-1}\left(2^{-1/2n}\right). \tag{6.4.18}$$

A result from simulation of the scenario is given in Figure 6.4.3, which includes the estimates of the correlation components, the power signal correlation coefficient, $\rho_{r^2} = \rho_{|H|^2}$, and the magnitude squared of the signal correlation coefficient, $|\rho_H|^2$. The calculation was over a distance of 2000 wavelengths. Figure 6.4.3 shows a typical difference between the two correlation coefficients, which is small for large correlations but increases for small correlations at the larger displacements. The power signal correlation estimate can be negative whereas $|\rho_H|^2$ must always be positive

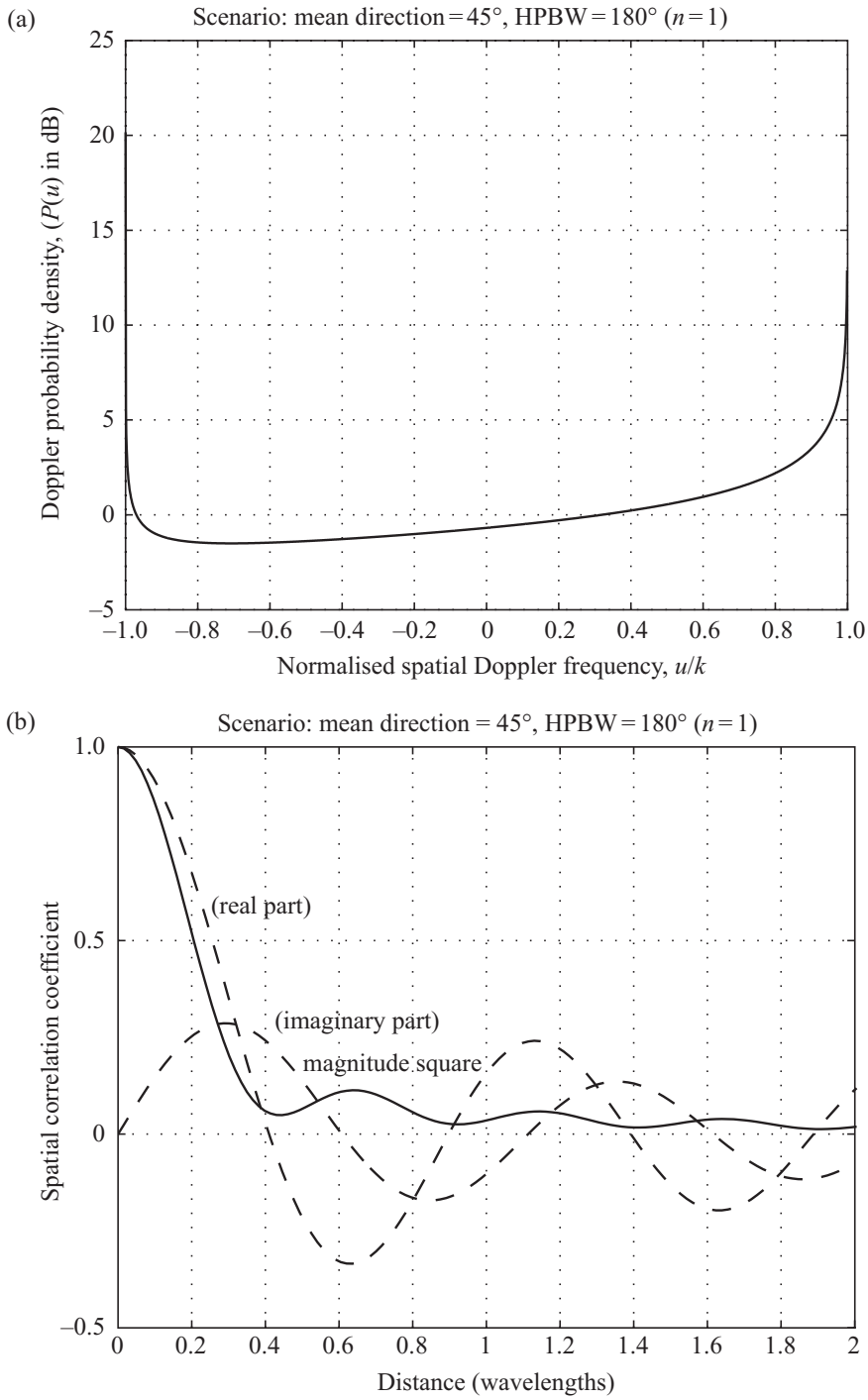


Figure 6.4.2 *(a) The spatial Doppler frequency pdf and (b) its Fourier transform, i.e. the spatial power correlation coefficient including the real and imaginary components for a directional scenario. The example is similar to that of the pattern of Figure 6.4.1: the mean angle is* $\theta_0 = 45°$*, but here* $n = 1$ *which is a scenario* $HPBW = 180°$*. The function is well behaved (monotonic) for correlations above about 0.1, and the 0.5 correlation distance is well defined*

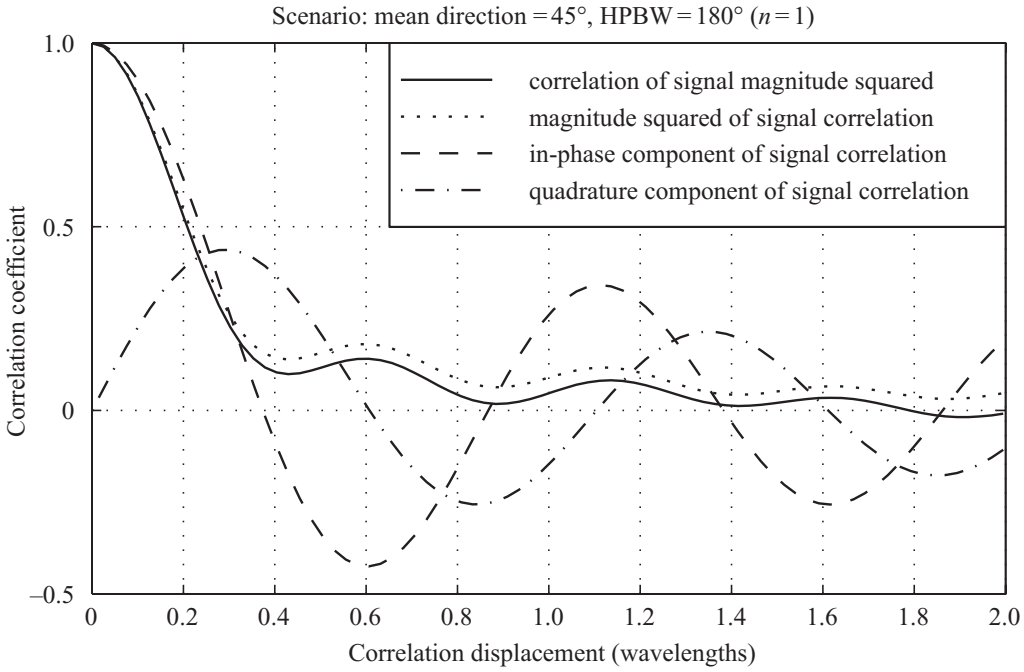


Figure 6.4.3 The correlation coefficients as calculated from a computer simulation with data length 2000 wavelengths

of course. An ensemble average of correlations from many simulations sees both correlation coefficients converge to be the same as that in Figure 6.4.2.

The real and imaginary components hold their form and periodicity, with their relative amplitudes being the variation, as the scenario parameters (the mean direction and HPBW) are varied. For monotonic (single beam) scenarios, the correlation is well behaved in the sense that a $\rho_{|H|^2} = 0.5$ correlation distance is well defined. This distance is plotted in Figure 6.4.4 (Vaughan, 2000a) and can be used as a guideline for spatial correlation distances.

The omnidirectional scenario is given by the $HPBW = 360°$ giving $d_d = 0.13\lambda$ from $J_0^2(k_C d_d) = 0.5$. As the directionality increases, the correlation distance increases, with the rate of increase greater for mean angles approaching 0, i.e. when the mean direction is along the spacing axis.

For the region where the HPBW is less than twice the mean angle, the straight line graphs demonstrate that the HPBW is inversely proportional to the correlation distance (cf. the uncertainty principle, where the product of the spreads of a Fourier pair are bounded, $\sigma_u \sigma_{\Delta z} \geq 1/2$.) From a least mean square solution, the relationship can be expressed, with the HPBW and mean angle in degrees, approximately as

$$HPBW \cdot \frac{\Delta z}{\lambda} = F(\theta_0), \quad HPBW < \sim\theta_0, \tag{6.4.19}$$

where

$$F(\theta_0) = \frac{2200}{(\theta_0 + 1)^{1.3}} + 11, \quad 10° \leq \theta_0 \leq 90°. \tag{6.4.20}$$

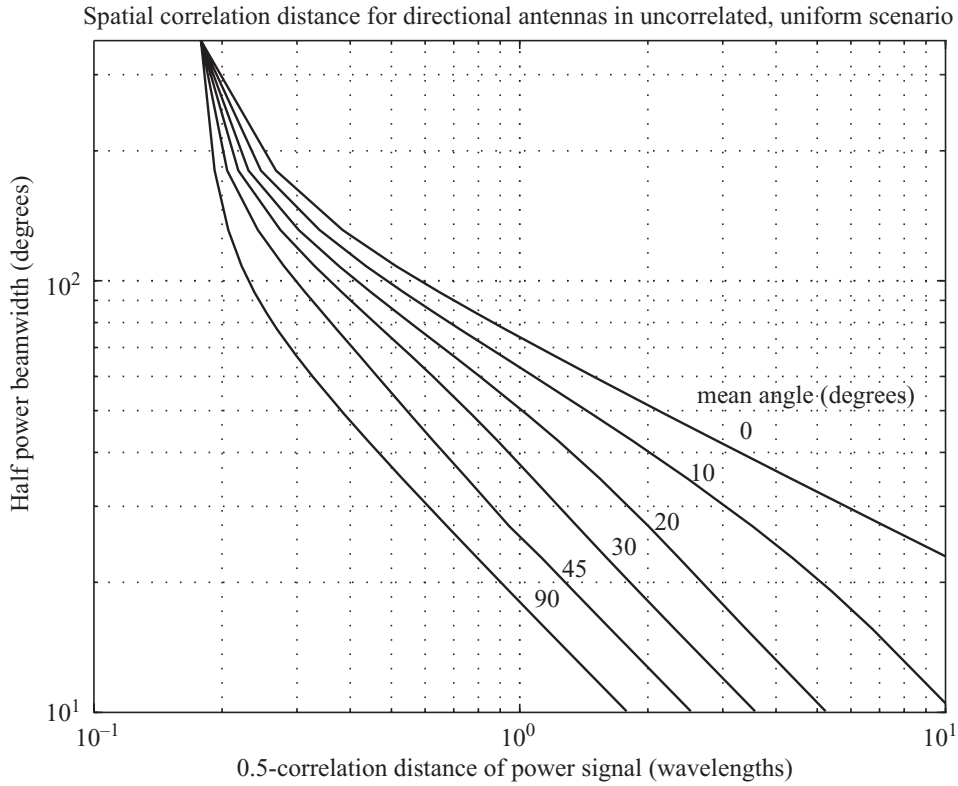


Figure 6.4.4 The correlation distance for a directional scenario. The directionality can be caused by the incident angular power distribution at the antenna, or the directionality of the antenna, or both. The curves are approximately straight when the HPBW is less than the beam direction angle

For scenarios or antennas with more complicated directionality, problems can arise with the definition of the correlation distance. The correlation distance can become poorly defined when the function is not monotonic. From the Fourier transform behaviour, this occurs, for example, when the majority of energy of the Doppler spectrum is near the support edges, $u = \pm k_C$ and the Fourier transform starts to take on a periodic nature. In another limiting case, we have a single ray, which is a delta function in the Doppler spectrum, and this has a correlation function which is always of magnitude one and the correlation distance does not exist. An example is for a double-lobed directional scenario, obtained for example by using a dipole-like pattern in an omnidirectional scenario. The correlation coefficient is given in Figure 6.4.5, which shows the potential confusion for the value of the 0.5 correlation distance[1].

[1] Mathematically, the correlation functions can always have a properly defined correlation distance, but in an engineering sense, the interpretation of results must include a recall that there are considerable assumptions in the model, and that this combined with the oscillatory nature of the function create uncertainty in the practical correlation distance.

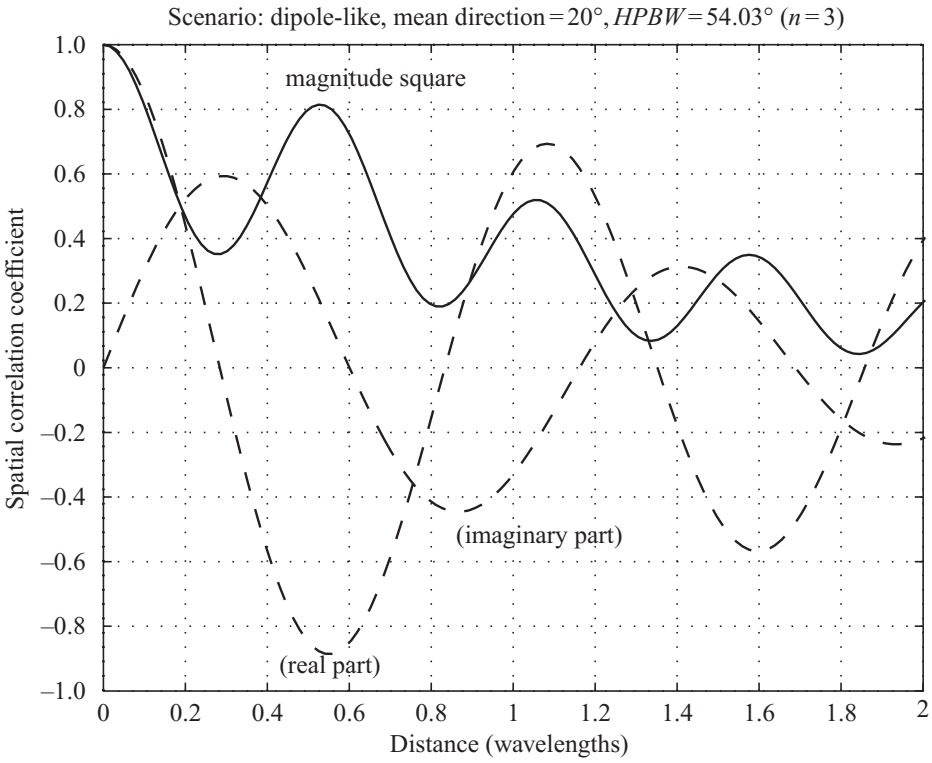


Figure 6.4.5 The correlation for double-lobed antenna power pattern (or scenario), given by $\cos^{2n}(\theta - \theta_0)$*, for the example* $n = 2, \theta_0 = 20°$

6.4.3 Frequency correlation

In an analogous way to the distance correlation, the power delay profile gives the correlation frequency (known as the coherence bandwidth), through the relationship $P(\tau) \Leftrightarrow R_H(\Delta\omega)$. For example, the exponential profile $P(\tau)$ and the corresponding correlation function are

$$P(\tau) = \frac{1}{s}\,\mathrm{e}^{-\tau/s} \quad \Leftrightarrow \quad \frac{s}{1 + \mathrm{j}\Delta\omega} = R_H(\Delta\omega), \tag{6.4.21}$$

respectively, where s is the delay spread. This gives the 0.5 coherence bandwidth, which is where the power correlation coefficient falls to 0.5, i.e. when

$$\rho_H(\Delta\omega) = \left|\frac{C_H(\Delta\omega)}{C_H(0)}\right|^2 = \left|\frac{R_H(\Delta\omega)}{R_H(0)}\right|^2 = 0.5, \tag{6.4.22}$$

which is $\Delta\omega^{(0.5)} = 1/s$.

6.4.4 Combined space and frequency correlation

The relation $R_H(\Delta z, \Delta\omega) \Leftrightarrow P(u, \tau)$ can be used to study the combination of both the frequency diversity and the space diversity. The spatial and frequency correlation

coefficients separate (Kalkan and Clarke, 1997)

$$|\rho_H(\Delta\omega, \Delta z)|^2 = |\rho(\Delta\omega)|^2 \cdot |\rho(\Delta z)|^2 \tag{6.4.23}$$

only if the angle of arrival and delay time of the waves are independent. Some geometric-based models for urban and suburban propagation (Liberti and Rappaport, 1996; Nørklit and Andersen, 1998) would not support true independence between the delay time and angles of arrival; however, at the mobile the independence seems a reasonable assumption. If the assumption can hold, then for a given scenario, the distance and frequency spacing for a given correlation can be found. As an example, the exponential delay profile and $\cos^n(\theta - \theta_0)$ directional scenario, the $|\rho_H|^2 = 0.5$ spacings are from

$$0.5 = \frac{1}{1 + (\Delta\omega)^2} \cdot K \int \frac{1}{\sqrt{k^2 - u^2}} \cos^{2n}\left(\frac{\cos^{-1}(u/k) - \cos^{-1}(u_0/k)}{2}\right) \mathrm{e}^{-\mathrm{j}u\Delta z}\, \mathrm{d}u. \tag{6.4.24}$$

A sample solution is plotted in Figure 6.4.6 (Vaughan, 2000a).

When the frequency spacing is zero, the distance spacing (about two wavelengths) corresponds to purely space diversity. Similarly, for spatially coincident beams, the

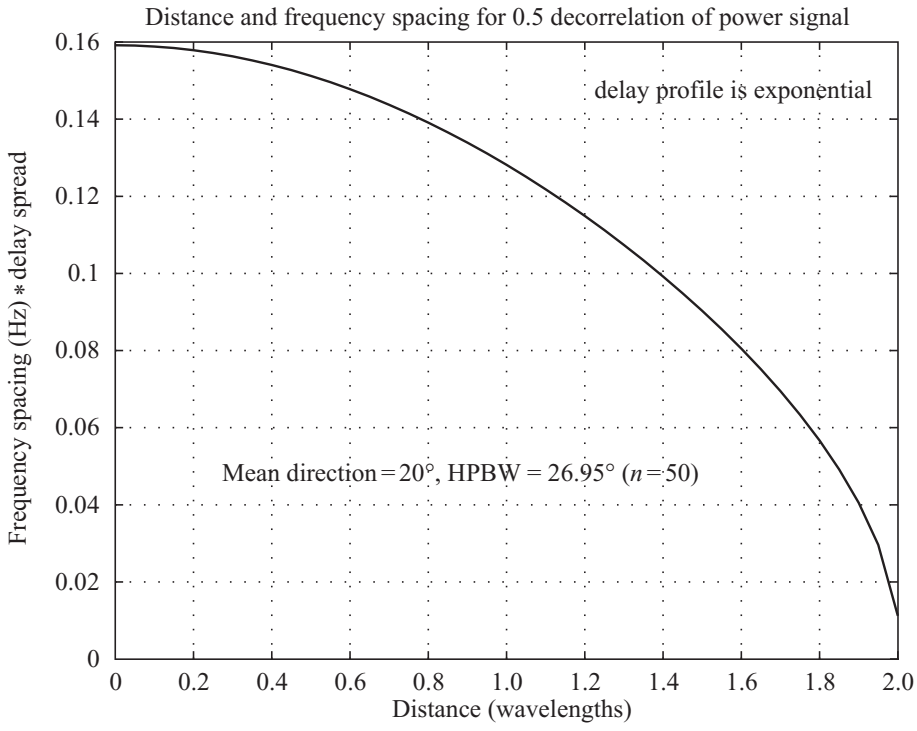


Figure 6.4.6 *The angular frequency (normalised by the delay spread) and distance (in wavelengths) spacings for a directional antenna with* $\cos^{2n}$ *power pattern with HPBW of* $27°$ $(n = 50)$ *and mean angle of* $\theta_0 = 20°$. *It is assumed that the angle of arrival and delay times of the scenario are independent*

required pure frequency diversity is a spacing equal to the inverse of the delay spread s. A trade-off example is to have a distance spacing of one wavelength and a frequency spacing of $0.8/s$.

Such curves need to be derived for any particular scenario to get the appropriate results. The algorithm is simple: for a range of spacings Δz_i calculate a set of distance correlation coefficients $\rho(\Delta z_i)$ from the term containing the integral and then the corresponding frequency spacings are given by

$$\Delta\omega_i = \sqrt{\frac{1}{\Gamma}\rho(\Delta z_i) - 1}, \tag{6.4.25}$$

where Γ is the correlation coefficient value ($\Gamma = 0.5$ in the above examples).

The trend of the design curves for different patterns is as intuitively expected. As the mean angle approaches broadside (to the axis of antenna spacing), i.e. $\theta_0 = 90°$, the curve moves to intercept the space axis closer to the origin. For example, for $\theta_0 = 70°$, the intercept is at 0.7 wavelengths, with the angular frequency axis intercept unchanged. A larger beamwidth results in a similar trend.

6.4.4.1 Summary

The many-path model can make use of Fourier transform methods which are convenient and insightful. It is not possible, in general, to relate the complete physical scattering of the environment to the effective scattering distribution. This is because the polarisation information of the physical environment is reduced to a scalar by the action of the antenna. The transform relations can be used with models of the scenario for diversity design and examples are given for spaced directive antennas and space–frequency diversity. The assumptions of the scenario models for the results presented are: wide-sense stationarity; Rayleigh fading; a smooth-lobed directive antenna pattern; a uniform angular distribution of incident power; an exponential distribution of power decay with delay time; and the multipath angle-of-arrival and delay time uncorrelated.

6.5 Integrating the transfer function over finite bandwidths and distances

6.5.1 Effect of integrating power over finite bandwidth

The Rayleigh-distributed envelope in the frequency domain undergoes deep fades in a similar manner to the envelope in the distance coordinate. Following the development for the two-path model, the fade depth of the integrated power reduces for increasing bandwidth. The Rayleigh distributed envelope averaged over a bandwidth is

$$\overline{r}(B) = \frac{1}{B}\int_{\omega-B/2}^{\omega+B/2} r(\omega)\,\mathrm{d}\omega. \tag{6.5.1}$$

Finding the pdf of $\overline{r}(\omega)$ seems a difficult problem, but the reducing variance with increasing bandwidth B can be quantified by dealing with the power quantities.

6.5.1.1 Power signal formulation

The power signal, $r^2 = |H|^2$, is dealt with since the statistical solution is exact, and the envelope behaviour is always within 10 per cent of the power result (see Appendix B). The power is often available in practice as the received signal strength indicator (RSSI) signal. The integration over frequency is typically an (analogue) implementation, i.e. the integral is implemented as a continuous, rather than sampled, averaging process.

The variance of the mean power signal is given by (Appendix B)

$$\sigma^2_{\overline{r^2}} = \frac{2}{B}\int_0^B \left(1 - \frac{\Delta\omega}{B}\right) C_{r^2}(\Delta\omega)\,\mathrm{d}\Delta\omega, \tag{6.5.2}$$

where the frequency covariance function is

$$C_{r^2}(\Delta\omega) = R_{r^2}(\Delta\omega) - \mu^2_{r^2} \tag{6.5.3}$$

and $R_{r^2}(\Delta\omega)$ is the autocorrelation function of the channel power with μ_{r^2} the true mean of the channel power. Assigning an exponential decay for the power delay profile, here expressed for normalised energy,

$$P(\tau) = \frac{1}{s}\mathrm{e}^{-\tau/s}, \tag{6.5.4}$$

where s is the delay spread, then the correlation for the transfer function is (Appendix B)

$$\begin{aligned} R_H(\Delta\omega) &= \int_0^\infty \frac{1}{s}\mathrm{e}^{-t/s}\mathrm{e}^{-\mathrm{j}\Delta\omega t}\mathrm{d}t, \quad t \geq 0, \\ &= \frac{s}{1 + \mathrm{j}\Delta\omega\, s}. \end{aligned} \tag{6.5.5}$$

Since the mean of the transfer function is zero (the channel model assumes zero-mean, Gaussian components), the correlation coefficient is just the normalised correlation,

$$\rho_H(\Delta\omega) = \frac{1}{1 + (\Delta\omega\, s)^2} - \mathrm{j}\frac{\Delta\omega\, s}{1 + (\Delta\omega\, s)^2}, \tag{6.5.6}$$

and for the power signal, this is (Appendix B)

$$\rho_{r^2}(\Delta\omega) = \rho_{|H|^2}(\Delta\omega) = |\rho_H(\Delta\omega)|^2 = \frac{1}{1 + (\Delta\omega\, s)^2}. \tag{6.5.7}$$

From the definition of the correlation coefficient, the covariance for the power can be written

$$R_{r^2}(\Delta\omega) - \mu^2_{r^2} = \left(R_{r^2}(0) - \mu^2_{r^2}\right)\rho_{r^2}(\Delta\omega). \tag{6.5.8}$$

Now using a property of Rayleigh signals in the normalising term

$$R_{r^2}(0) = \overline{r^4} = 2\left(\overline{r^2}\right)^2 = 2\mu^2_{r^2}, \tag{6.5.9}$$

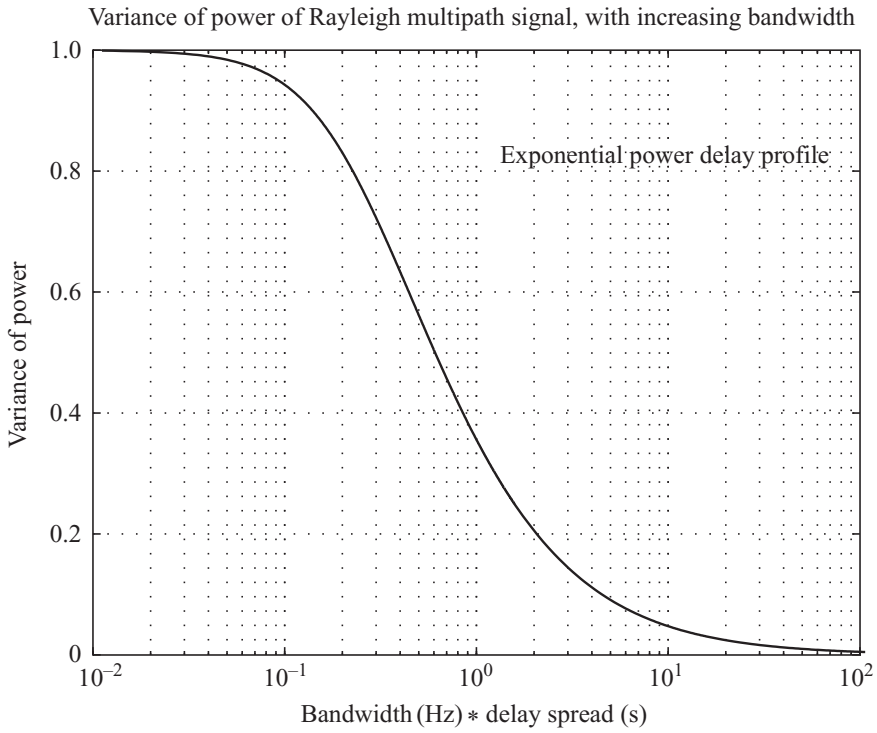


Figure 6.5.1 The reduction in variance of the power in the Rayleigh-envelope channel, as a function of the bandwidth. The curve is for an exponential delay profile with delay spread s

we get

$$R_{r^2}(\Delta\omega) - \mu_{r^2} = \mu_{r^2}^2 \left(\frac{1}{1 + (\Delta\omega\, s)^2} \right). \tag{6.5.10}$$

Using this result, the variance of the mean power is, here normalised,

$$\frac{\sigma_{\overline{r^2}}^2(B)}{\mu_{r^2}^2} = 2\frac{\tan^{-1} Bs}{Bs} - \frac{\ln\left((Bs)^2 + 1\right)}{(Bs)^2}. \tag{6.5.11}$$

For zero bandwidth (cw case), the normalisation is by its variance since for Rayleigh envelopes

$$\sigma_{\overline{r^2}}(B) = \mu_{r^2}, \quad B \to 0. \tag{6.5.12}$$

The normalised variance is plotted in Figure 6.5.1 (Vaughan, 2000a,b) in which the bandwidth B is expressed in Hz instead of radians as in the formulation above.

The variance begins to drop quickly when $B_{\text{Hz}}s \approx 0.1$ and begins to flatten out at $B_{\text{Hz}}s \approx 3$. So the bandwidth needs to be larger that about $0.1/s$ Hz to reduce the fading significantly, in the sense of significantly reducing the variance of the power. In the zero bandwidth limit, the power has an exponential pdf (Rayleigh distributed envelope) and the variance of the power is the energy squared. In the limiting wideband case the variance is small with the envelope pdf approaching Gaussian

since it is equivalent to a summation of essentially uncorrelated frequency samples. (Note that the correlation coefficient never reaches zero for finite bandwidth, so the 'uncorrelated' frequency samples always retain a finite correlation and are never truly independent.) For a finite frequency band, the pdf cannot be truly Gaussian since the envelope cannot be negative. This suggests a log-normal-like form for the many-path pdf. For the cases in between the extreme wideband and the cw, the envelope pdf has a rapidly changing (with bandwidth) pdf, analogous to the few-path model discussed above.

From the correlation coefficient similarity (within 10 per cent) for the envelope and power (Appendix B), the normalised envelope variance is similar to the normalised power variance,

$$\frac{\sigma_{\overline{r}}^2}{\sigma_r^2} \approx \frac{\sigma_{\overline{r^2}}^2}{\sigma_{r^2}^2}, \tag{6.5.13}$$

where $\sigma_r^2 = (1 - \pi/4)\overline{r^2}$ so Figure 6.5.1 can also be used as an approximation for the envelope, although the effect of the integration over the band has not been quantified. The envelope correlation coefficient is always smaller than the power correlation coefficient, so we can at least state that the normalised variance of the envelope will be smaller than that for the power.

6.5.1.2 Spread about mean

A useful presentation of the above result is the spread which can be written, here for $\pm\beta$ standard deviations about the mean, as

$$S_{\overline{r^2}}^{(\beta\sigma)} = \frac{\mu_{r^2} + \beta\sigma_{\overline{r^2}}}{\mu_{r^2} - \beta\sigma_{\overline{r^2}}}, \tag{6.5.14}$$

where μ_{r^2} is the true mean. This can be expressed in dB and as a function of the bandwidth B as

$$S_{\overline{r^2}}^{(\beta\sigma)}(B), \text{(in dB)} = 10\ \log_{10}\left(\frac{1 + \beta(\sigma_{\overline{r^2}}/\mu_{r^2}(B))}{1 - \beta(\sigma_{\overline{r^2}}/\mu_{r^2}(B))}\right), \tag{6.5.15}$$

and this is given in Figure 6.5.2. Without knowing the pdf, the interpretation of this result should be tempered with care. For small bandwidths $B_{\text{Hz}}s \ll 1$, the pdf will be similar to a lognormal form and because it cannot extend to negative values of the argument, the quantity $(\mu_{r^2} - \beta\sigma_{\overline{r^2}})$ (the denominator in the spread) will not exist for higher values of β. For example, this occurs for $\beta \geq 1$ for the one-sided exponential distribution in which the mean is equal to the standard deviation. For larger bands, however, the distribution of the integrated power approaches Gaussian from the central limit theorem. An example from Figure 6.5.2 is that for a given measurement, a signal which fully occupies a 'brick wall' channel with an exponential power delay profile of width $B_{\text{Hz}}s = 10$, then the 1-σ spread is 2 dB. If it can be assumed that the distribution of power is Gaussian, then this means that some 64 per cent ($\mu \pm 1\sigma$) of single measurements will be within a range of 2 dB, and that some 95 per cent ($\mu \pm 2\sigma$) of measurements will be within a 4 dB range.

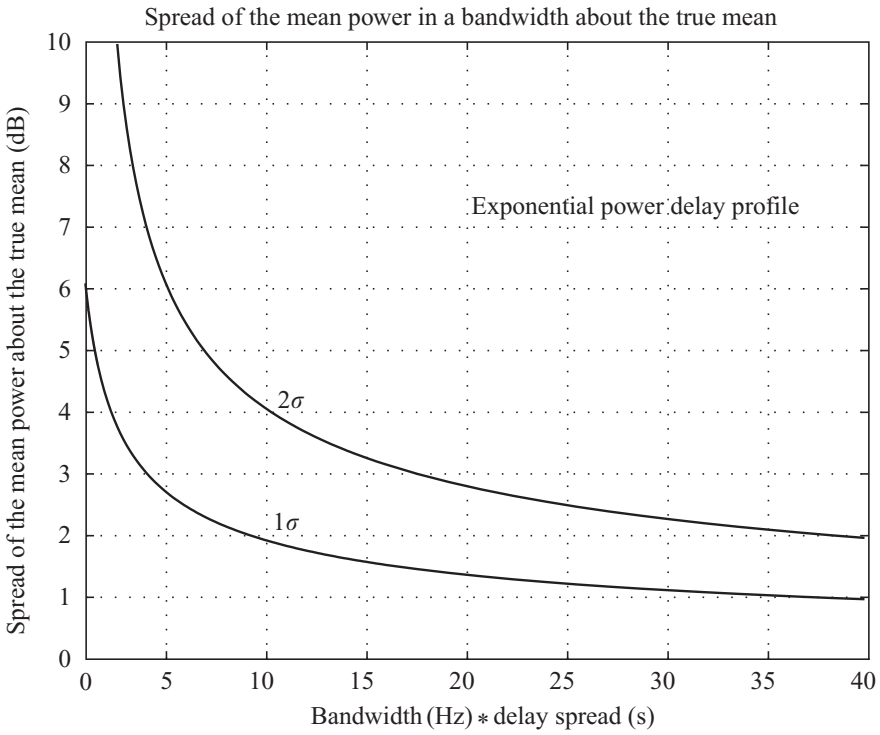


Figure 6.5.2 *The* 1-σ *and* 2-σ *spread for the estimate of the mean power of a channel occupying a band of B Hz and having an exponential power delay profile with delay spread s*

6.5.2 Effect of integrating power over a finite distance

6.5.2.1 Integrated power

The spatial analogy to integrating power over a bandwidth is to integrate the power over a distance. This is used for estimating the mean power level in coverage measurements. The variance of the power is found the same way as for the finite bandwidth case above.

For example, the omnidirectional Doppler spectrum gives the covariance function,

$$C_{r^2}(z) = R_{r^2}(z) - \mu_{r^2}^2 = \sigma_{r^2} J_0^2 (k_C z)\,, \tag{6.5.16}$$

where k_C is the wavenumber and z is the distance. The normalised variance of the integrated power signal is then (using Appendix B)

$$\frac{\sigma_{\overline{r^2}}^2(L)}{\sigma_{r^2}^2} = \frac{2\lambda}{L}\int_0^{L/\lambda} J_0^2(k_C z)\,\mathrm{d}z - J_0^2(k_C L) - J_1^2(k_C L). \tag{6.5.17}$$

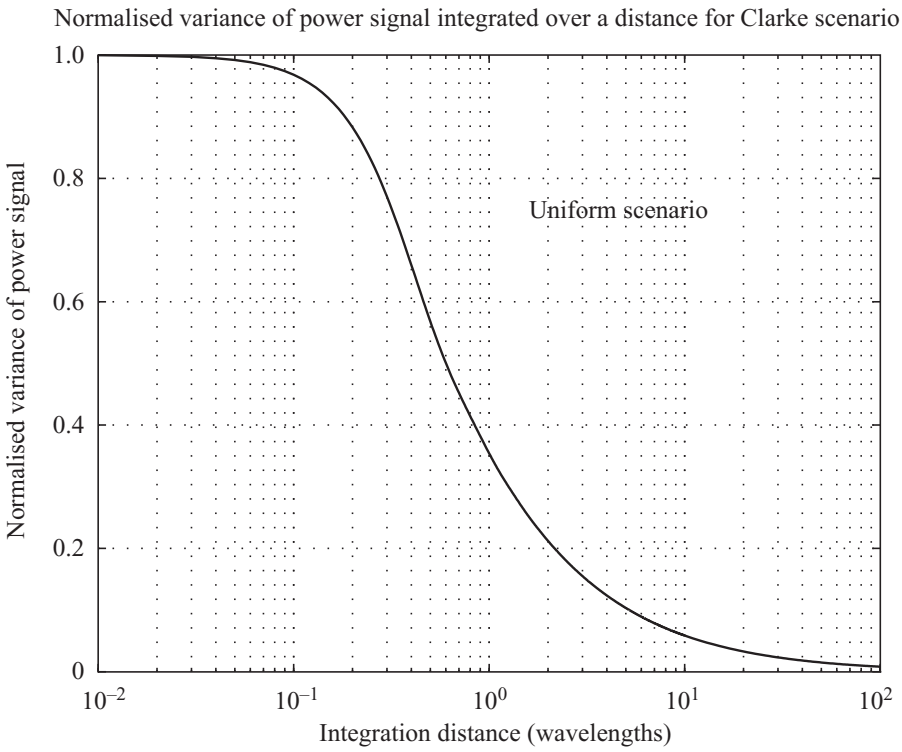


Figure 6.5.3 The variance of the power of the signal which is averaged over distance in a uniform scenario. The variance is normalised to the variance of the power (envelope squared). The behaviour of the envelope variance is similar

For computation, the limiting value is found using

$$\mathrm{Lim}_{\epsilon \to 0} \left\{ \frac{1}{\epsilon} \int_0^{\epsilon} J_0^2(kx)\, \mathrm{d}x \right\} = 1. \tag{6.5.18}$$

The normalised variance is plotted in Figure 6.5.3 (Vaughan, 2000b) and is almost identical to that of the bandwidth relation in Figure 6.5.1. This is because the covariance functions used here, although having different forms, have similar rates of decay. The variance drops quickly after a few tenths of wavelengths of averaging distance, but the reduction levels out at around two to ten wavelengths, and then the decrease becomes very slow with increasing distance. As with the frequency integration case above, the envelope case has not been explicitly calculated, but it can again be stated that the normalised envelope variance will be smaller than the normalised power variance since the envelope correlation coefficient is smaller than that of the power signal.

6.5.2.2 Spread about mean

The spread for $\pm\beta$ standard deviations about the mean, is presented in decibel form in Figure 6.5.4 for $\beta = 1, 2$.

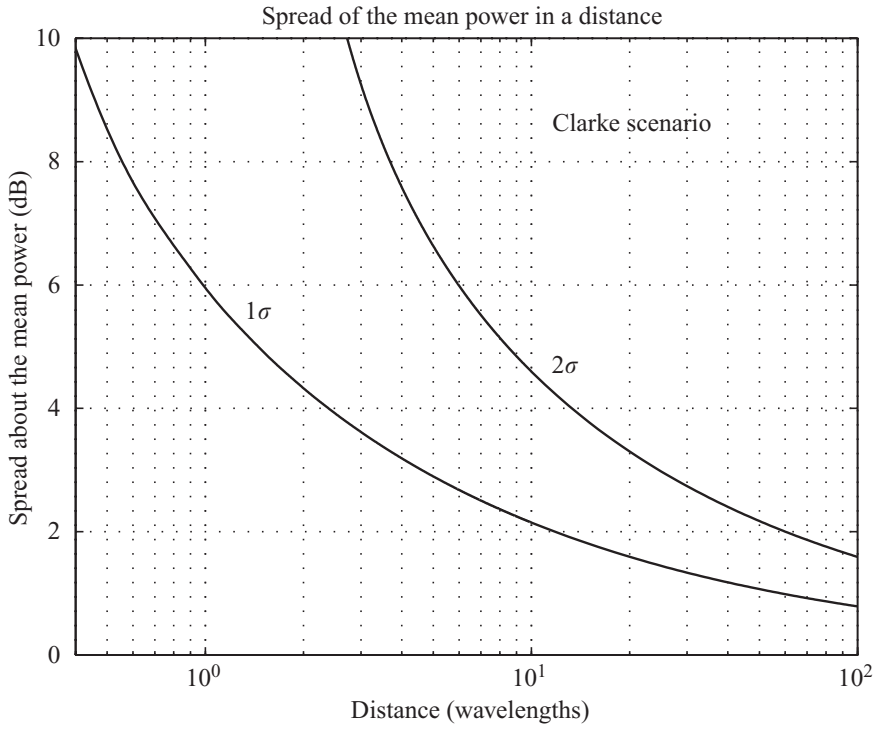


Figure 6.5.4 The one- and two-standard deviation spread of the power signal integrated over a distance within a uniform scenario. The four-wavelength distance corresponds to about ten uncorrelated samples (see Figure 6.5.5)

The interpretation is that the proportion of measurements occupying the pdf between $(\mu - 2\sigma)$ and $(\mu + 2\sigma)$ (95 per cent if the pdf is Gaussian) would lie within a range given by the spread. For example, for 95 per cent of measurements to be within a ± 1 dB range, a distance of some 60 wavelengths is required for the measurement. In practice, it would be fortuitous to find a stationary signal over this distance, although it can be born in mind that the distance need not be lineal. For example, a 60-wavelength serpentine-like path, where the adjacent traces are separated by a half-wavelength, fills an area of less than six by five wavelengths.

In implementing this measurement approach, the true (analogue) integral is replaced by dense samples. These samples are highly correlated. An approach also used is to deal with independent samples, as follows.

6.5.2.3 Sampled power

It is of interest to treat the same problem from a sampled signal approach, since this is an established way to estimate from measurements. The sampling requirements are that the samples are independent. Some experimenters sample at a spatial interval where the correlation coefficient is zero, and this spacing depends on the Doppler spectrum. For a uniform scenario, the first zero of the correlation coefficient function is at about 0.38λ, and this is the correct average spacing for uncorrelated adjacent

samples, although this value is not critical since for the envelope, the correlation coefficient function $J_0^2(kz)$ has zero gradient here. In practice, the required spacing is usually larger owing to the scenario not being the same as the models. For 25 independent samples, say, the distance required would be ten wavelengths, or for a real sampling, a square of two by two wavelengths. However, these samples have low correlations, and are not truly independent.

The interpretation of independent samples occurring at a fixed correlation spacing is, strictly speaking, incorrect. This is because there is no justification in assuming that individual samples from the same scenario are independent even if their mutual correlations are zero. The proper interpretation of the sampled measurement being independent is that each sample must be taken from a different statistical realisation or 'sample' of the same scenario. In computer simulations, this is straightforward to implement, for example by assigning new random phases to the discrete sources. Obviously, such sampling cannot be directly equated to a given distance in a particular scenario realisation. Although it may be possible to find spacings for independent samples, it is not clear what the spacings are. However, an equivalent spacing can be estimated by comparing the results of the continuous and sampled cases.

The power (envelope squared), here denoted γ, is sampled. Its mean is denoted Γ. Its distribution is exponential,

$$p_\gamma(\gamma) = \frac{1}{\Gamma} \mathrm{e}^{-\gamma/\Gamma} \tag{6.5.19}$$

and the variance, σ_γ^2, and mean are identical, i.e.

$$\sigma_\gamma^2 = \Gamma^2. \tag{6.5.20}$$

The sample estimates are denoted, using the usual formulas for N samples, for the mean

$$\bar{\gamma} = \frac{1}{N} \sum_{i=1}^{N} \gamma_i, \tag{6.5.21}$$

and the variance,

$$\overline{\sigma^2}_\gamma = \frac{1}{1-N} \sum_{i=1}^{N} (\gamma_i - \bar{\gamma})^2 = \frac{1}{N-1} \left(\sum_{i=1}^{N} \gamma_i - N\bar{\gamma}^2 \right)^2. \tag{6.5.22}$$

If the samples are drawn from identical distributions (which seems reasonable at some spacing within short distances of tens of wavelengths in the same physical environment), and are considered independent, then the sample variance of the mean estimate is simply

$$\sigma_{\bar{\gamma}}^2 = \frac{\sigma_\gamma^2}{N}. \tag{6.5.23}$$

The central limit theorem gives $\bar{\gamma}$ as Gaussian distributed if γ is either Gaussian distributed for any N, or for any other distribution of γ in the large sample limit

($N \to \infty$). The case here deals with a finite number of samples and a non-Gaussian distribution for γ. So the mean power will not be Gaussian distributed (as can be observed from the fact that the mean cannot be negative since the samples are non-negative), and the pdf is unknown. But for a reasonable number of samples, say more than ten, then the Gaussian distribution, with the exception of negative values at the lower tail, is accurate enough for the rules-of-thumb presented here. From the approximation that the mean power is normally distributed with the above parameters, denoted $\mathcal{N}\left(\Gamma, \sigma_\gamma/\sqrt{N}\right)$, then by the usual transformation

$$z = \frac{\bar{\gamma} - \Gamma}{\left(\sigma_\gamma/\sqrt{N}\right)} \tag{6.5.24}$$

the distribution of z is $\mathcal{N}(0,1)$. The confidence interval for z is

$$\Pr(-X < z < X) = \operatorname{erf}\left(\frac{X}{\sqrt{2}\sigma_z}\right) \tag{6.5.25}$$

and this is written for the power variables as

$$\Pr\left(1 - \left(\frac{X\sigma_\gamma}{\sqrt{N}\Gamma}\right) < \frac{\bar{\gamma}}{\Gamma} < 1 + \left(\frac{X\sigma_\gamma}{\sqrt{N}\Gamma}\right)\right) = \operatorname{erf}\left(\frac{X}{\sqrt{2}\sigma_z}\right). \tag{6.5.26}$$

If we now introduce the reduction in variance resulting from a finite bandwidth from the previous section, then the effect of bandwidth (integration over frequency) can be included with the effect of integrating over distance. The reduction in the standard deviation of the power due to the increasing bandwidth, is denoted by α. (Note the ordinate of Figure 6.5.1 is α^2.) Examples from the curve of Figure 6.5.1 are (B is in Hz): $Bs = 0, \alpha = 1$ (i.e. single frequency, or cw case); $Bs = 0.2, \alpha \approx 0.9$; $Bs = 0.5, \alpha \approx 0.7$; $Bs = 1, \alpha \approx 0.6$; $Bs = 5, \alpha \approx 0.3$. The confidence interval becomes

$$\Pr\left(1 - \left(\frac{\alpha X}{\sqrt{N}}\right) < \frac{\bar{\gamma}}{\Gamma} < 1 + \left(\frac{\alpha X}{\sqrt{N}}\right)\right) = \operatorname{erf}\left(\frac{X}{\sqrt{2}\sigma_z}\right) \tag{6.5.27}$$

and as an example, a value of $X = 2\sigma_z$ gives a 95 per cent confidence level, i.e. the limits in this equation will hold for 95 per cent of the measurements taken. Note that the variance reduction from the finite bandwidth has an equivalent effect to changing the confidence level. Figure 6.5.5 gives the limits in dB, $10 \log_{10}(1 \pm 2\beta/\sqrt{N})$, for a couple of examples of finite bandwidth as well as the cw (where the bandwidth approaches zero) case (Vaughan, 2000b). The dB difference between the limits is the spread. Recall that these curves are approximations only, especially for the smaller numbers of samples.

For larger sample sizes, especially if taken linearly, the sampling distances will probably encompass some shadow-fading effects, and the spatial mean power estimate would then include this shadow-type fading. Here, the assumption in the formulation that the underlying distribution being sampled is unchanging (resulting in the variance

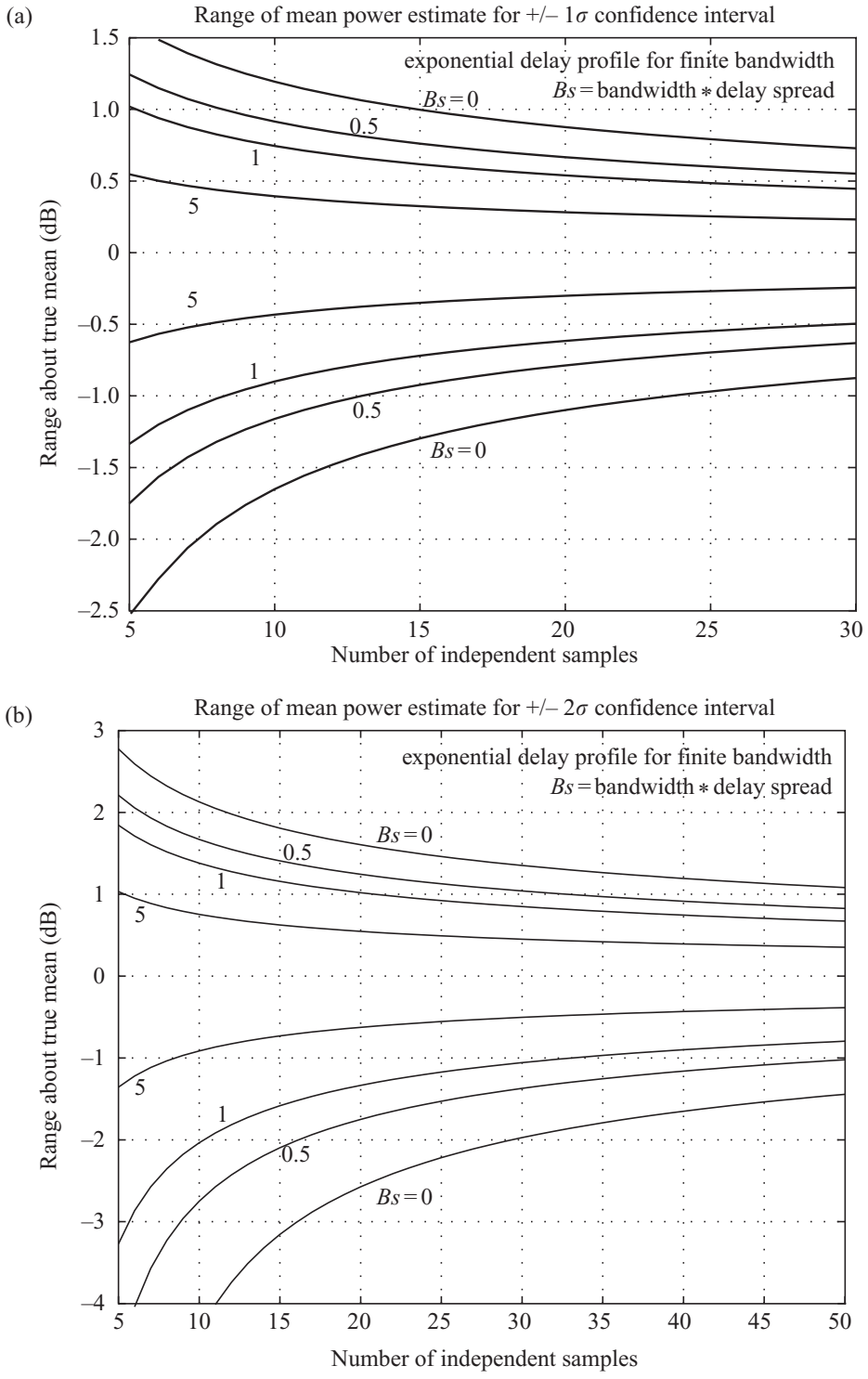


Figure 6.5.5 An approximation for the range of the mean power calculated from independent samples of the power, with the effect of bandwidth as a parameter. For finite bandwidths, the power delay profile is taken as exponential. Bs is the bandwidth (Hz)–delay spread (s) product. (a) The $\pm 1\sigma$ spread and (b) the $\pm 2\sigma$ spread

reduction by a factor of N), becomes violated and the variance formulation breaks down. The wider bandwidth has the effect of making the mean power measurements more accurate, or equivalently, implies that fewer samples need to be taken for a given accuracy. For example, 20 independent samples of a cw measurement ($Bs = 0$) gives the 95 per cent confidence limits as about -2.5 dB to about 1.5 dB. Using a wideband channel signal with $Bs = 5$ gives these limits as about ± 0.2 dB. All the curves are very slowly converging after about 30 samples.

For comparison with the results from the continuous integration, the spread from the sampled approach is given in Figure 6.5.6 and is on the same scale as Figure 6.5.4. The abscissa is here the number of samples times 0.5, which corresponds to distance in wavelengths (cf. Figure 6.4.6) if we take a sample every 0.5 wavelengths. The curves are similar but not identical. The 'proper' average spacing for uncorrelated adjacent samples is 0.38λ for the uniform scenario, and this multiplier moves the sampled spread to be smaller than the integrated spread of Figure 6.5.4. This is because the sampled approach assumes independent samples, whereas samples spaced by their distance for which the correlation is zero, are not necessarily independent, as discussed above. The effect of the finite bandwidth on the spread is also given in Figure 6.5.6 which contains the same information as in Figure 6.5.5.

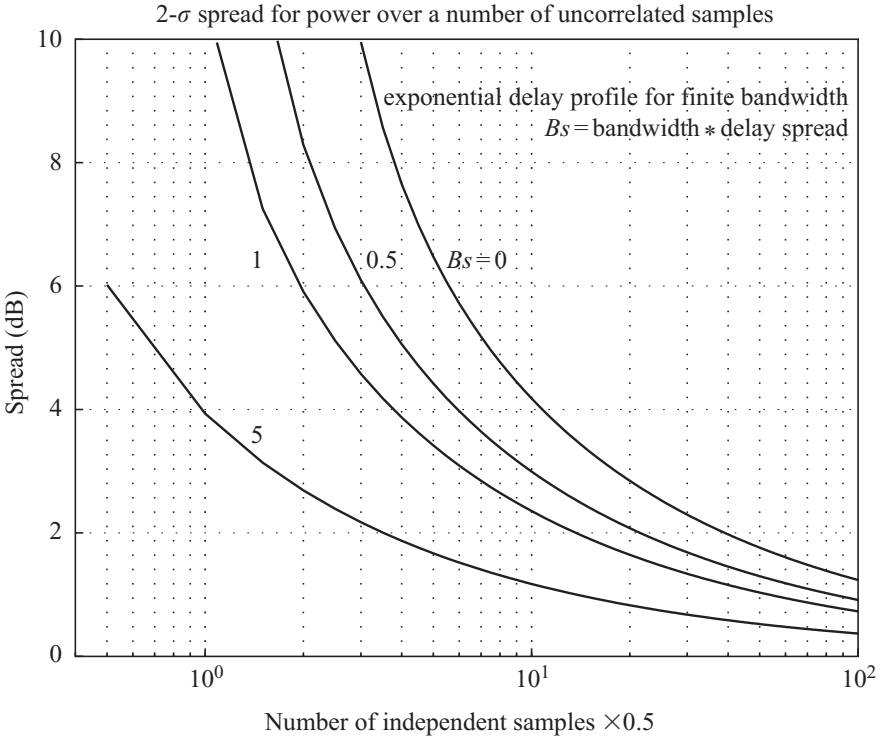


Figure 6.5.6 The spread of the mean power calculated from independent samples of the power, with the effect of bandwidth as a parameter. For finite bandwidths, the power delay profile is taken as exponential. Bs is the bandwidth (Hz)–delay spread (s) product. The figure has its abscissa scaled to allow comparison with Figure 6.5.4

The practical accuracy limit for mean signal strength is sometimes perceived as being the distance required for obtaining sufficient independent samples being larger than the distance over which the signal statistic (mean) is stationary. By using a wider bandwidth signal, this distance limit is reduced according to the bandwidth size.

For systems that are constrained to be narrowband, an alternative is to use antenna diversity and utilise the fact that the combined signal has reduced variance and as with the finite bandwidth case, therefore requires a smaller sampling (or integrating) distance. In implementing a measurement system, note that a signal combiner is not necessarily required – the M multiple antennas (say M dipoles using space diversity) give M times as many samples for a given distance covered, and the distance required for a given accuracy is M times smaller than the spread-vs.-distance curves indicate.

Related to the estimation of the mean using a finite distance or number of samples, is the topic of estimating the correlation coefficient and this is covered in Chapter 7.

6.5.3 Continuous form: space and frequency integration

The two-dimensional form for finding the mean signal strength is

$$\overline{r^2} = \langle r^2 \rangle_{B,L} = \frac{1}{BL} \int_{-B/2}^{B/2} \int_{-L/2}^{L/2} |H(\omega, z)|^2 \, \mathrm{d}z \, \mathrm{d}\omega. \tag{6.5.28}$$

The two-dimensional correlation function is $C_{r^2}(\Delta\omega, \Delta z)$ and the variance of the mean estimate $\overline{r^2}$ is

$$\begin{aligned} \sigma^2_{\overline{r^2}} &= \frac{1}{BL} \int_{-B}^{B} \int_{-L}^{L} C_{\overline{r^2}}(\Delta\omega, \Delta z) \left(1 - \frac{|\Delta\omega|}{B}\right) \left(1 - \frac{|\Delta z|}{L}\right) \mathrm{d}\Delta z \, \mathrm{d}\Delta\omega \\ &= \frac{2}{BL} \left(\int_0^B C_{\overline{r^2}}(\Delta\omega) \left(1 - \frac{\Delta\omega}{B}\right) \mathrm{d}\Delta\omega \int_0^L C_{\overline{r^2}}(\Delta z) \left(1 - \frac{\Delta z}{L}\right) \mathrm{d}\Delta z \right), \end{aligned} \tag{6.5.29}$$

with the second line holding if the covariance function is separable.

6.6 Examples of modelled scatter distributions

To fix ideas regarding the relationship between the scenario and the effective scatter distribution, some examples and their properties are presented. Although the Clarke model can demonstrate realistic short-term complex fading behaviour, there can be unrealistic results which are artifacts of unrealistic aspects of the model. Continuous and discrete formulas are discussed in order to assist with the implications of the properties of scenario models for the mobile transfer function. To promote clarity, no normalisation has been applied here (the distributions below cannot be used necessarily as pdfs), the 2π factor (see below) in the transfer function is dropped, and the scenarios are confined to $\pi > \theta > 0$.

The multipath signals received by a moving receiver are often modelled using a deterministic scenario such as the Clarke (1969) scenario. The receiver is placed at

the centre of the ring and is moved in space. The antenna acts to weight and sums the source contributions to form the scalar received signal.

Recall that the formulation requires the condition that the receiving antenna is in the far-field of the sources and that the source amplitudes and phases are constant for changes in the receiver position. The condition holds if the physical scatterers (trees, buildings, etc.) are many wavelengths from the low-gain mobile receiving antenna, and their amplitude and phase do not change much over the trajectory of the mobile. The physical scattering model features vector (i.e. polarised) incident fields with different phases and delays in 3D space. The effective scattering model is scalar and is in 2D space with the effective scatterers being the scalar values of the incident fields weighted by the receiving antenna pattern. The term 'effective scatterers', or just scatterers in this section, is used because these scalar representations do not necessarily correspond to the physical scatterers 'visible' to the mobile antenna. The effective scattering distribution, $a(\tau, u)$, and the transfer function are a Fourier transform pair, and are recalled as

$$\begin{aligned} H(\omega, z) &= \frac{1}{2\pi} \int_0^{\infty} \int_{-k_C}^{k_C} a(\tau, u) \mathrm{e}^{-\mathrm{j}\omega\tau} \mathrm{e}^{\mathrm{j}uz} \, \mathrm{d}u \, \mathrm{d}\tau \\ a(\tau, u) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(\omega, z) \mathrm{e}^{\mathrm{j}\omega\tau} \mathrm{e}^{-\mathrm{j}uz} \, \mathrm{d}\omega \, \mathrm{d}z. \end{aligned} \tag{6.6.1}$$

6.6.1 Single source

This simplest case is

$$\begin{aligned} a(\tau, u) &= a_1 \mathrm{e}^{\mathrm{j}\phi_1} \delta(u - u_1) \delta(\tau - \tau_1) \\ H(\omega, z) &= a_1 \mathrm{e}^{\mathrm{j}\phi_1} \mathrm{e}^{-\mathrm{j}\omega\tau_1} \mathrm{e}^{\mathrm{j}u_1 z}, \end{aligned} \tag{6.6.2}$$

showing how the transfer function is constant amplitude with a phase affected by (i) the position of the receiver, z, and the spatial Doppler shift of the source u_1; and (ii) the frequency ω and the delay time to the effective scatterer τ_1. (Note here the interpretation of the delta function in the integral with finite limits for the transfer function.) The single source case is illustrated in Figure 6.6.1.

6.6.2 Distributed line source distribution with constant amplitude and phase

The function $\tau(u)$ describes a constant amplitude and phase curve, so (for unity amplitude and zero phase)

$$\begin{aligned} a(\tau, u) &= \delta(\tau(u)) \\ H(\omega, z) &= \int_{-k_C}^{k_C} \mathrm{e}^{\mathrm{j}(uz - \omega\tau(u))} \, \mathrm{d}u. \end{aligned} \tag{6.6.3}$$

For multiple lines, the transfer function is a summation of the Fourier transforms of each line. A depiction of the line source is given in Figure 6.6.2.

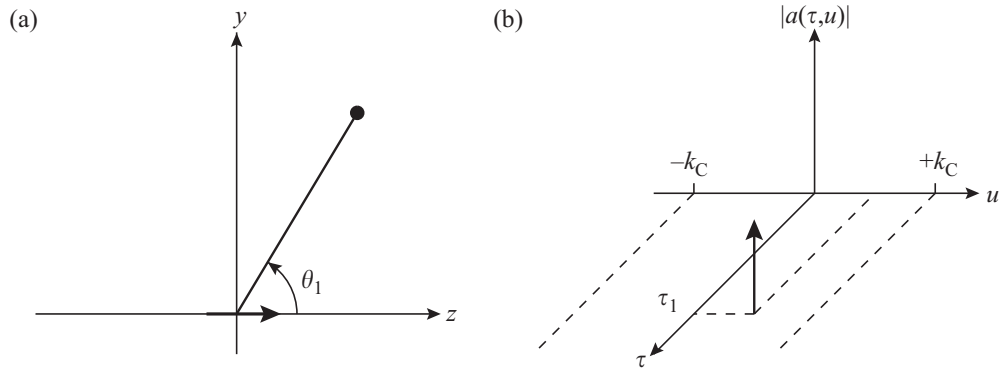


Figure 6.6.1 *(a) The spatial representation of the one-sided ($y > 0$) single source with the mobile antenna at the orgin moving along the positive z-axis and (b) the effective scattering distribution*

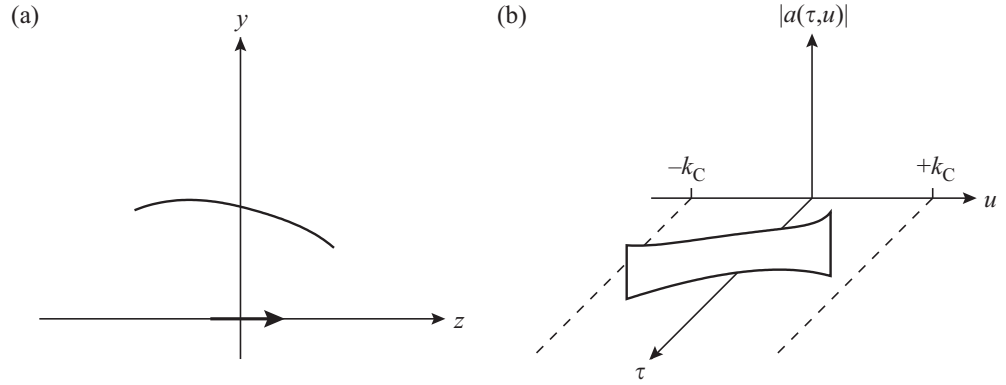


Figure 6.6.2 *(a) The spatial representation of the constant amplitude and phase line source and (b) the effective scattering distribution*

6.6.3 Distributed line source distribution with random phase

In order to randomise the phase of a continuous effective scatter distribution, the amplitude is sampled and the samples assigned a random phase, ϕ_i. The formulas become (here for individual amplitudes a_i as well)

$$\begin{aligned} a(\tau, u) &= \sum a_i \mathrm{e}^{\mathrm{j}\phi_i} \delta(\tau - \tau(u)) \delta(u - u_i) \\ H(\omega, z) &= \sum a_i \mathrm{e}^{\mathrm{j}\phi_i} \int_{-k_C}^{k_C} \mathrm{e}^{\mathrm{j}(uz - \omega\tau(u))} \delta(u - u_i)\, \mathrm{d}u \\ &= \sum a_i \mathrm{e}^{\mathrm{j}\phi_i} \mathrm{e}^{-\mathrm{j}\omega\tau_i} \times 2k_C \frac{\sin k_C z}{k_C z}, \end{aligned} \tag{6.6.4}$$

where $\tau_i = \tau(u_i)$. In this case the delta functions are interpreted as requiring infinite limits on the integral with the result that the transfer function becomes expressed as a

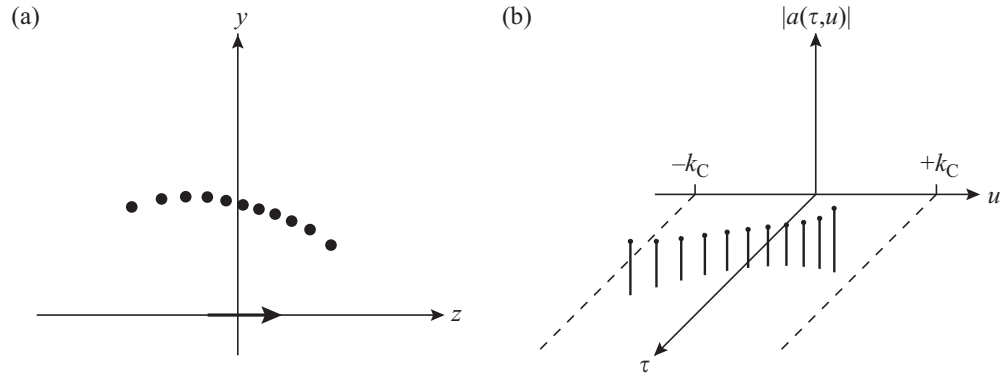


Figure 6.6.3 (a) The spatial representation of the discretised constant amplitude and phase line source and (b) the effective scattering distribution

sum of sinc functions in z (cf. the single-source case, above, where the sinc function is omitted). The transfer function now features Rayleigh fading as long as there are sufficient sources (say $> \sim 10$ with similar amplitudes). See Figure 6.6.3.

6.6.4 Distributed circular line source with constant amplitude, phase and delay

This is similar to the Clarke scenario, except that the sources are continuous and in phase. The relations are

$$\begin{aligned} a(\tau, u) &= \frac{1}{\sqrt{k_C^2 - u^2}} \delta(\tau - \tau_1), \quad -k_C < a < k_C, \\ &= 0 \quad \text{elsewhere;} \\ H(\omega, z) &= \mathrm{e}^{-\mathrm{j}\omega\tau_1} \cdot 2J_0(k_C z), \end{aligned} \tag{6.6.5}$$

i.e. the magnitude of the transfer function is proportional to $|J_0(k_C z)|$, because of the radial symmetry of the scenario; z can be interpreted as the receiver's (spatial) azimuthal position, i.e. the transfer function is constant for fixed radial positions in space. A change in frequency (or delay time, τ_1, which corresponds to the radius of the ring source) involves a linear phase change on the transfer function.

The amplitude distribution term $(k^2 - u^2)^{-1/2}$ does not have a major bearing on the transfer function. In this sense, the support of $a(\tau, u)$ and its phase (which is constant in this example) comprises the more important information of the distribution. For example, if we drop the amplitude term above, i.e. the $(k^2 - u^2)^{-1/2}$ term, the transfer function becomes

$$H(\omega, z) = \mathrm{e}^{-\mathrm{j}\omega\tau_1} \cdot 2\frac{\sin k_C z}{k_C z}, \tag{6.6.6}$$

which has similar form to the $J_0(k_C z)$ form (see Figure 6.6.4).

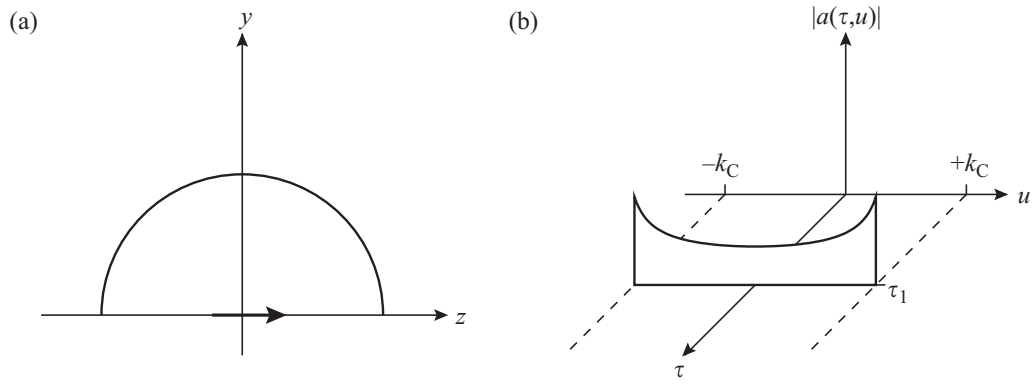


Figure 6.6.4 (a) The spatial representation of the constant amplitude and phase circular line source and (b) the effective scattering distribution

6.6.5 Discrete uniform scenario, uniform delay (Clarke scenario)

The classical scenario is expressed

$$a(\tau, u) = \sum \frac{1}{\sqrt{k_C^2 - u_i^2}} e^{j\phi_i} \delta(u - u_i)\delta(\tau - \tau_i), \quad -k_C < u_i < k_C,$$
$$= 0 \quad \text{elsewhere;} \tag{6.6.7}$$

where the summation denotes an even spacing between the u_i. An alternative, from the previous section, is to use a constant amplitude for $a(\tau, u)$, but space the delta function components corresponding to constant spacing between θ_i

$$a(\tau, u) = \sum e^{j\phi_i} \delta(k_C \cos\theta - k_C \cos\theta_i)\delta(\tau - \tau_i), \quad |\theta| \leq \pi, \tag{6.6.8}$$

and the summation here denotes an even spacing in θ_i.

The transfer function is

$$H(\omega, z) = \sum e^{-j\omega\tau_i} \frac{1}{\sqrt{k_C^2 - u_i^2}} \cdot 2k_C \frac{\sin k_C z}{k_C z}, \tag{6.6.9}$$

which in the z-direction is the summation of $(\sin k_C z)/z$ functions weighted by an amplitude term with random phase. The result is classical Rayleigh flat-frequency fading. A change in frequency imposes a linear phase change with frequency which is proportional to the delay time to the ring of scatterers (see Figure 6.6.5).

6.6.6 Clarke scenario with single phase wave delay

A plane wave illuminates the ring of scatterers, which reradiate (without a common phase change) which in turn results in a delay profile. With these considerable physical assumptions, the model is shown in Figure 6.6.6.

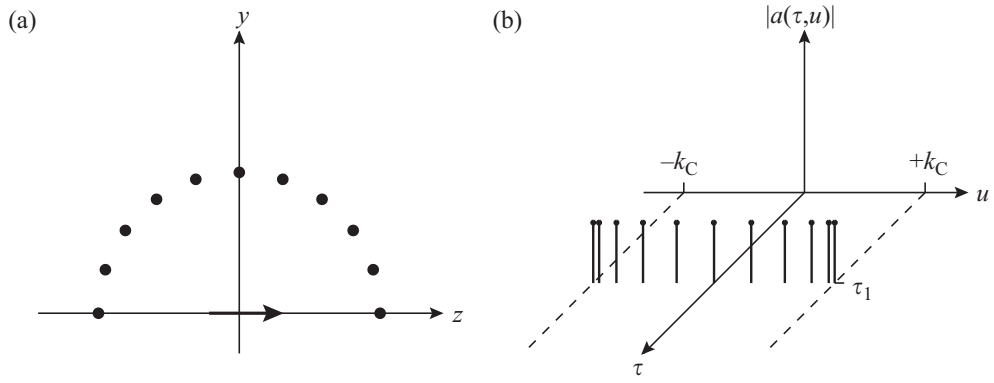


Figure 6.6.5 (a) The spatial representation of the Clarke scenario and (b) the effective scattering distribution

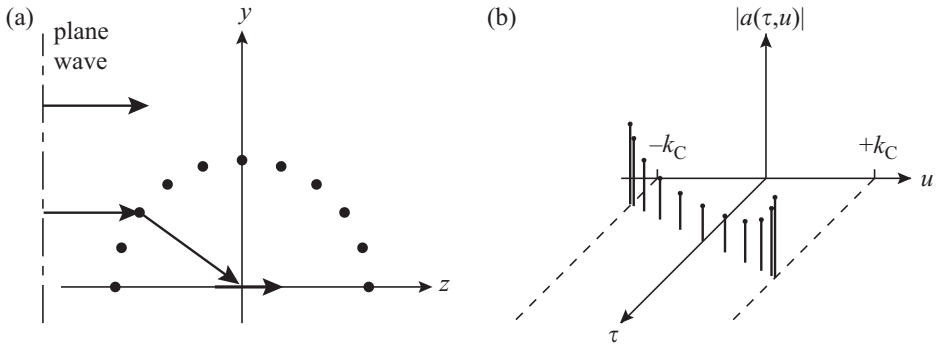


Figure 6.6.6 (a) The spatial representation of the Clarke scenario with single wave illumination and (b) the effective scattering distribution

The effective scattering distribution is

$$a(\tau, u) = \sum \frac{1}{\sqrt{k_C^2 - u_i^2}} e^{j\phi_i} \delta(\tau - \tau(u))\delta(u - u_i), \quad -k_C < u_i < k_C, \\ = 0 \quad \text{elsewhere;} \tag{6.6.10}$$

where

$$\tau(\theta) = \tau_1(2 + \cos\theta), \tag{6.6.11}$$

i.e.

$$\tau(u) = \tau_1\left(2 + \frac{u}{k_C}\right). \tag{6.6.12}$$

The effective sources thus occupy a cardioid when the radial coordinate is delay time, and the effective scattering distribution $a(\tau, u)$ has undergone a linear skew owing to the plane wave illumination. Multiple plane wave illumination from different

angles would produce multiple scatter distributions comprising elliptic-like skews. The transfer function, for the single wave illumination is

$$H(\omega, z) = \mathrm{e}^{-\mathrm{j}\omega 2\tau_1} \sum \mathrm{e}^{-\mathrm{j}\omega\tau_i} \frac{1}{\sqrt{k_\mathrm{C}^2 - u_i^2}} \mathrm{e}^{\mathrm{j}\phi_i} \cdot 2k_\mathrm{C} \frac{\sin(k_\mathrm{C} z + \omega\tau_1)}{(k_\mathrm{C} z + \omega\tau_1)}, \tag{6.6.13}$$

and the effect of the plane wave illumination is to shift the transfer function in the (ω, z) domain, and to impose a linear phase change with frequency.

The magnitude of the transfer function has the shifting property

$$|H(\omega, z)| = |H(\omega + \Delta\omega, z + \Delta z)|, \tag{6.6.14}$$

where

$$\Delta\omega = \frac{\Delta z \tau_1}{k_\mathrm{C}}. \tag{6.6.15}$$

This phenomenon was first observed numerically by Amoroso (1996), whose parameters are: spatial radius to sources $= 140\lambda$, frequency $= 1.98\,\mathrm{GHz}$, $\Delta z = 0.84\lambda$, and $\Delta\omega \approx 75\,\mathrm{MHz}$. In a real situation, this shifting property in the transfer function will not occur – it is an artifact of the scenario model.

6.6.7 *Continuous uniform scenario with uniform phase and exponential delay profile*

A non-zero delay spread is introduced by incorporating an exponential delay profile with the same delay spread, s, for each source direction. With constant phase, the continuous form of the distribution is

$$a(\tau, u) = \left(\frac{1}{s}\mathrm{e}^{-(\tau-\tau_1)/s}\right) \cdot \frac{1}{\sqrt{k_\mathrm{C}^2 - u^2}}, \quad \tau \geq \tau_1, \tag{6.6.16}$$

and this is the bathtub function which decreases exponentially along the delay axis. It gives the transfer function as

$$H(\omega, z) = \left(\frac{-1 + \mathrm{j}\omega s}{1 + (\omega s)^2}\right) \mathrm{e}^{-\mathrm{j}\omega\tau_1} \cdot 2J_0(k_\mathrm{C} z). \tag{6.6.17}$$

6.6.8 *Discrete uniform scenario with uniform phase and exponential delay profile*

In discretising the above example, the amplitudes for a given delay time are all taken as the same. The discrete sources are then

$$a(\tau, u) = \frac{1}{s}\mathrm{e}^{\frac{\tau_1}{s}} \sum_{m\ (\text{in } \tau)} \sum_{n\ (\text{in } \theta)} \mathrm{e}^{\mathrm{j}\phi_{mn}} \mathrm{e}^{-\frac{\tau_m}{s}} \delta(\tau - \tau_m)\delta(k_\mathrm{C}\cos\theta - k_\mathrm{C}\cos\theta_n) \tag{6.6.18}$$

and the $u = k_\mathrm{C}\cos\theta$ summation is over even spacings in θ_n. The effective source distribution is depicted in Figure 6.6.7. The transfer function has a delay spread of s

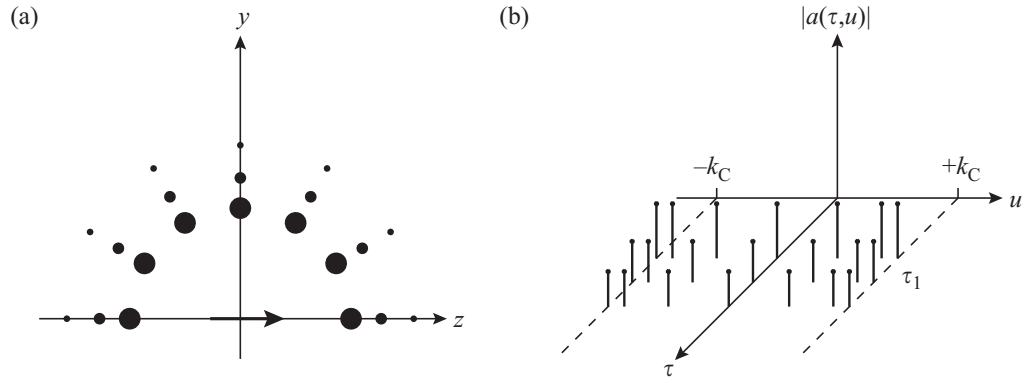


Figure 6.6.7 (a) The spatial representation of the discrete uniform scenario with exponential delay profile and (b) the effective scattering distribution

as long as the receiver is concentric to the source positions, and there is Rayleigh fading. With a finite delay spread, the translation properties of the transfer function do not occur.

6.7 Averaged scenario models

6.7.1 Basic formulation

The location of the radiating sources (scatterers) around a receiver is a statistical quantity in mobile communications. Models for the time-averaged distributions of these source bearings relative to mobile antenna coordinates, the *averaged scenario*, is useful for the design of the antennas since the pattern of the receiving antenna should be 'matched' as well as possible to this distribution to maximise the time-averaged gain of fixed beam antennas. Few experimental data are available for the averaged scenarios. Instantaneously, it is possible that only a few sources are present. While the principles of propagation can be used to infer such instantaneous scenarios, some guesswork and even wishful thinking has gone into developing the averaged scenarios which should be kept simple and tractable.

Recall from Chapter 2 that the incident fields are written

$$\boldsymbol{E}(\theta, \phi, t) = E_\theta(\theta, \phi, t)\hat{\theta} + E_\phi(\theta, \phi, t)\hat{\phi}, \tag{6.7.1}$$

where $\boldsymbol{E}$, E_θ, and E_ϕ here have units of volt m^{-1} steradian^{-1}. The time, t, and the scaled position, $z = Vt$, can be interchanged for the moving receiver in a static scenario. Averaging over the time here means averaging over the fading of the channel. So *uncorrelated signals*, or *uncorrelated scatterers*, means that the fading of the waves coming from different directions or delays is uncorrelated. The signals which are borne by the incident waves are correlated since they are replicas of the same signal.

A time- (or space-) averaged linear polarisation matrix can be defined which contains the angular correlations of the polarisation components of the incident fields.

It characterises and models the time-averaged incident field distribution and is written (cf., Chapter 4, Collin and Zucker, 1969)

$$\mathbf{\Gamma}'(\theta_1, \phi_1; \theta_2, \phi_2) = \begin{bmatrix} \mathbf{\Gamma}'_{\theta\theta} & \mathbf{\Gamma}'_{\theta\phi} \\ \mathbf{\Gamma}'_{\phi\theta} & \mathbf{\Gamma}'_{\phi\phi} \end{bmatrix}, \tag{6.7.2}$$

where the elements are of the form

$$\Gamma'_{\theta\phi}(\theta_1, \phi_1; \theta_2, \phi_2) = \left\langle E_\theta(\theta_1, \phi_1, t)\, E^*_\phi(\theta_2, \phi_2, t) \right\rangle_t. \tag{6.7.3}$$

The use of the prime is to separate it from the polarisation matrix of diversity antennas, which is denoted similarly, but without the prime. The receiving antenna polarisation matrix is discussed in Chapter 8. Other orthogonal polarisations can be used here of course, in particular the circular polarisations.

If the waves of each polarisation are uncorrelated with each other, or independent, then the off-diagonal terms of $\mathbf{\Gamma}'$ are zero, i.e. $\mathbf{\Gamma}'_{\theta\phi} = \mathbf{\Gamma}'_{\phi\theta} = 0$. If it is also assumed that the incoming waves from different angles are uncorrelated, then the polarisation matrix reduces to

$$\mathbf{\Gamma}'(\theta_1, \phi_1; \theta_2, \phi_2) = \delta(\theta_1 - \theta_2)\delta(\phi_1 - \phi_2) \begin{bmatrix} \left\langle |E_\theta(\theta_1, \phi_1)|^2 \right\rangle_t & 0 \\ 0 & \left\langle |E_\phi(\theta_2, \phi_2)|^2 \right\rangle_t \end{bmatrix}. \tag{6.7.4}$$

This polarisation matrix, describing the incident waves, can be used in an integral equation with antenna-receiving patterns for deriving antenna diversity parameters such as gain and correlation coefficients. Interpreting the matrix as a stand-alone quantity (instead of appearing in an integral equation) means that the 'delta' functions in the equation should be interpreted as finite-width, finite-energy types of function. The angular decorrelation angle is then given by the width of the function. For most modelling, however, the delta functions are considered as Dirac functions.

Denote the polarised distributions of incoming waves as $\left\langle |E_\theta(\theta, \phi)|^2 \right\rangle = P_1 p_\theta(\theta, \phi)$, where $p_\theta(\theta, \phi)$ is the pdf of angular power per steradic square and P_1 is the total power in the θ-polarisation, and simlarly for P_2 in the ϕ-polarisation. Now normalising the total power in both polarisations so that $\mathrm{Tr}\left(\mathbf{\Gamma}'^{(N)}\right) = 1$,

$$\begin{aligned} &\mathbf{\Gamma}'^{(N)}(\theta_1, \phi_1; \theta_2, \phi_2) \\ &\quad = \delta(\theta_1 - \theta_2)\delta(\phi_1 - \phi_2) \frac{1}{P_1 + P_2} \begin{bmatrix} P_1 p_\theta(\theta_1, \phi_1) & 0 \\ 0 & P_2 p_\phi(\theta_2, \phi_2) \end{bmatrix}. \end{aligned} \tag{6.7.5}$$

Finally, if the distribution of incident power is the same for each polarisation, $S(\theta, \phi)$, then the polarisation matrix can be written (Vaughan and Andersen, 1987),

$$\mathbf{\Gamma}'(\theta_1, \phi_1; \theta_2, \phi_2) = S(\theta, \phi)\delta(\theta_1 - \theta_2)\delta(\phi_1 - \phi_2) \begin{bmatrix} XPD & 0 \\ 0 & 1 \end{bmatrix}, \tag{6.7.6}$$

and the cross-polar discrimination of the incident waves from the scenario

$$XPD = \frac{\mathbf{\Gamma}'_{\theta\theta}}{\mathbf{\Gamma}'_{\phi\phi}} = \frac{P_1}{P_2} \tag{6.7.7}$$

is the ratio of the averaged power in each polarisation. It is discussed further below. The elements of the polarisation matrix are now for the horizontally polarised power and the vertically polarised power, i.e.

$$\mathbf{\Gamma}'_{\phi\phi}(\theta_1, \phi_1; \theta_2, \phi_2) = S(\theta, \phi)\delta(\theta_1 - \theta_2)\delta(\phi_1 - \phi_2), \qquad \text{(horizontal)} \qquad (6.7.8)$$

$$\mathbf{\Gamma}'_{\theta\theta}(\theta_1, \phi_1; \theta_2, \phi_2) = XPD \cdot S(\theta, \phi)\delta(\theta_1 - \theta_2)\delta(\phi_1 - \phi_2), \qquad \text{(vertical)}. \qquad (6.7.9)$$

In summary, we assume here the same angular distribution of incoming waves for both polarisations after averaging in a given environment, and that the fading in each polarisation is mutually uncorrelated. The waves in each polarisation are also assumed uncorrelated for changing incident angles. The total incoming power in the vertical polarisation is *XPD* times that of the horizontal polarisation.

6.7.2 Cross-polar discrimination (XPD)

The average ratio of power in the polarisations is often set to unity for convenience; however, experimental results often indicate that this is not the case. From the rough surface scattering treatment in Chapter 3, single reflections from rough surfaces give some ensemble average values for a single reflection. These are in the order of 10 dB for linear copolar to cross-polar ratios but this figure depends on several parameters. After the multiple reflections of urban/suburban environments, measurements by Lee and Yeh (1972), Kozono *et al*. (1984), Vaughan (1991), and Eggers *et al*. (1993) give some numbers for XPD, which range from 0 dB to 18 dB, with 6 dB often found in urban environments. These measurements are referred to 'geographical' coordinates, since they are measured using geographically vertical and horizontal dipole type antennas. All the measurements include the antenna effect as well, of course, and the measured XPD is in terms of the radio channel signals, not purely from the waves of the propagation channel. The XPDs of the radio and propagation channels are different in general because of the effect of the antenna pattern. The measured XPD depends on the antenna pattern. The quantities can be summarised:

$$XPD(\theta, \phi) = \frac{\langle |E_\theta(\theta, \phi)|^2 \rangle}{\langle |E_\phi(\theta, \phi)|^2 \rangle} = \frac{P_1}{P_2} \qquad \text{(scenario)}, \qquad (6.7.10)$$

$$XPD = \frac{\left\langle \left| \int\int E_\theta(\theta, \phi) \cdot h^*_\theta(\theta, \phi) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \right|^2 \right\rangle}{\left\langle \left| \int\int E_\phi(\theta, \phi) \cdot h^*_\phi(\theta, \phi) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \right|^2 \right\rangle}$$

$$= \frac{P_1 \int S_\theta(\Omega) |h_\theta(\Omega)|^2 \, \mathrm{d}\Omega}{P_2 \int S_\phi(\Omega) |h_\phi(\Omega)|^2 \, \mathrm{d}\Omega} \qquad \text{(measured)}, \qquad (6.7.11)$$

where $h_{\theta,\phi}(\theta, \phi)$ are the polarised antenna patterns, Ω expresses (θ, ϕ) as the solid angle, and the integrals are over real space.

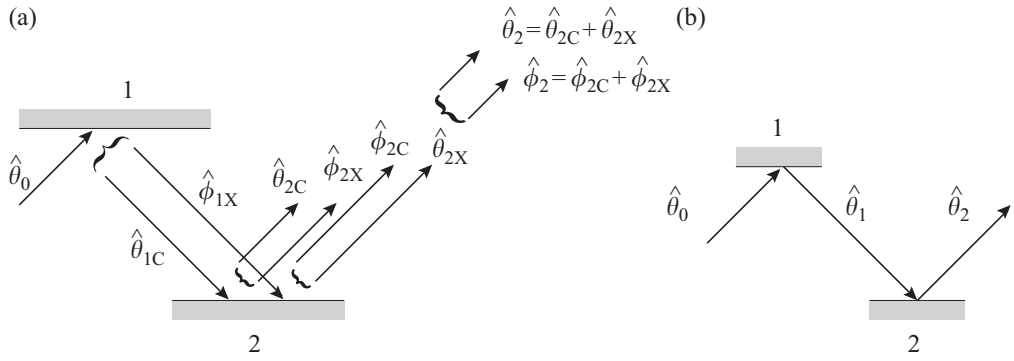


Figure 6.7.1 *(a) Elemental rays (components of the overall reflected wave, with single polarisation) from two scatterers with polarisation coupling for each reflection. (b) 'Composite' rays with no polarisation coupling on reflection*

6.7.2.1 Ray model from general scattering

A model for finding the XPD from multiple reflections is based on the scattering coefficients in a cascade network. The ray model is a physically appealing interpretation, and this is depicted in Figure 6.7.1 for the first two reflections. A ray, in the context here, occupies only one polarisation. There is one incident ray, denoted $\hat{\theta}_0$. The first reflection produces a co-polarised ray $\hat{\theta}_{1C}$ and a cross-polarised ray $\hat{\theta}_{1X}$. The second reflection produces $\hat{\theta}_{2C}$ and $\hat{\theta}_{2X}$ in the co-polarisation and $\hat{\phi}_{2C}$ and $\hat{\phi}_{2X}$ in the cross-polarisation. So after two reflections, there are four 'elemental' rays and after three reflections, eight elemental rays, and so on. These elemental rays are useful in computer ray tracing, but, assuming there is no appreciable delay difference between the polarisations from a reflection, they cannot be resolved by measurement. If the rays in the same polarisations are combined into single 'composite' rays, which is what would be observed, e.g.

$$\hat{\theta}_2 = \hat{\theta}_{2C} + \hat{\theta}_{2X}, \quad \hat{\phi}_2 = \hat{\phi}_{2C} + \hat{\phi}_{2X}, \tag{6.7.12}$$

then there are always two only rays per path, one for each polarisation.

After many (N) reflections with a variety of reflection coefficients, an 'output' elemental ray has power $P^{(\mathrm{O})}$, which is proportional to the product of the N power reflection coefficients, i.e.

$$P^{(\mathrm{O})} \propto \prod_{i=1}^{N} R^{(i)} \tag{6.7.13}$$

where $R^{(i)}$ are the power reflection coefficients. Successive path loss factors (see below) are omitted for now, but these factors also combine as products of numbers for the rays. If N is sufficiently large, then the powers in the 'output' rays for this case will be lognormally distributed, from the central limit theorem applied

to the product of positive random numbers (see Appendix C). It turns out, however, that very many factors are needed before the lognormal distribution is well approximated.

If a physical scatterer re-radiates (towards the receiving antenna) many incident rays that originate from independent ray paths, and their delay times are not resolvable, then the composite ray comprises a sum of random variables. The power in a general composite ray is therefore exponentially distributed (Rayleigh envelope), from the central limit theorem for the sum of random variables. If the scattering has no polarisation coupling, as indicated for the composite rays of Figure 6.7.1 (b), then such composite rays are the same as elemental rays.

So in the context of this cascade transmission line model, some information on the distributions of the reflected waves can be postulated for the power arriving from multipath scatterers. A signal received by a mobile antenna will normally be a sum of many multipath waves, so the power in the receiver can be expected to be always exponentially distributed, independent of the distribution of the powers of the incident rays, but assuming that the total power incident on the antenna is constant. If the total power incident on the antenna varies lognormally, then the envelope of the received signal follows the Suzuki distribution.

A single incident wave is denoted with having co-polar power $P_{\mathrm{C}}^{(\mathrm{I})}$ and cross-polar power $P_{\mathrm{X}}^{(\mathrm{I})}$, and a superscript of (O) is used for the output (reflected) power; see Figure 6.7.2, below. The co-polar to co-polar power coupling coefficient is denoted R_{CC} and the cross-polar to cross-polar coefficient is R_{XX}:

$$R_{\mathrm{CC}} = \frac{P_{\mathrm{CC}}^{(\mathrm{O})}}{P_{\mathrm{CC}}^{(\mathrm{I})}}, \quad R_{\mathrm{XX}} = \frac{P_{\mathrm{XX}}^{(\mathrm{O})}}{P_{\mathrm{XX}}^{(\mathrm{I})}}. \tag{6.7.14}$$

In general, the reflection coefficients are a function of angle, which is the direction of interest of the reflected wave. If the reflection is lossless and from a smooth, planar surface, then the cross- to co-polar reflection coefficient, and the co- to cross-polar reflection coefficient, respectively, are

$$R_{\mathrm{CX}} = 1 - R_{\mathrm{XX}}, \quad R_{\mathrm{XC}} = 1 - R_{\mathrm{CC}}. \tag{6.7.15}$$

A 'lossy' reflection (this could mean that some energy is absorbed by the scatterer but it is more likely that energy is 'lost' because it scatters in several different directions, and our interest is in only one direction) can be catered for by reducing the reflection coefficients, and the above energy conservation relations will no longer hold.

A reflection with polarisation coupling is written in terms of field strengths, and scattering parameters denoted

$$\begin{bmatrix} E_{\mathrm{C}}^{(\mathrm{O})} \\ E_{\mathrm{X}}^{(\mathrm{O})} \end{bmatrix} = \begin{bmatrix} S_{\mathrm{CC}} & S_{\mathrm{CX}} \\ S_{\mathrm{XC}} & S_{\mathrm{XX}} \end{bmatrix} \cdot \begin{bmatrix} E_{\mathrm{C}}^{(\mathrm{I})} \\ E_{\mathrm{X}}^{(\mathrm{I})} \end{bmatrix}, \tag{6.7.16}$$

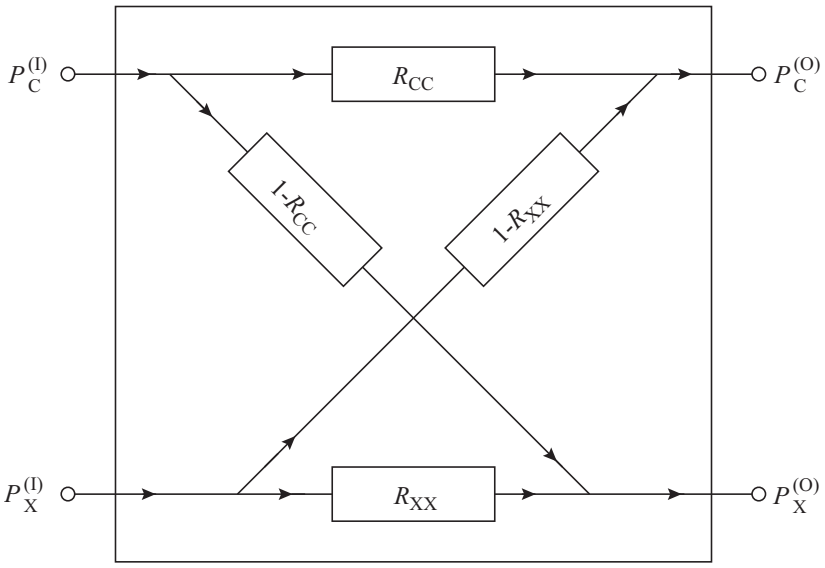


Figure 6.7.2 *A transmission matrix model for the polarisation coupling in a mean reflection of a wave from a smooth, planar surface where directional energy is conserved. R_{CC} is the mean co-polar power reflection coefficient and R_{XX} is the mean cross-polarisation power reflection coefficient*

where the scattering parameters are related to the power reflection coefficient by, for example, $|S_{CC}|^2 = R_{CC}$. The power relations for a 'mean reflection' can be stated

$$\begin{bmatrix} \left\langle |E_C^{(O)}|^2 \right\rangle \\ \left\langle |E_X^{(O)}|^2 \right\rangle \end{bmatrix} = \begin{bmatrix} \langle R_{CC} \rangle & \langle R_{CX} \rangle \\ \langle R_{XC} \rangle & \langle R_{XX} \rangle \end{bmatrix} \cdot \begin{bmatrix} |E_C^{(I)}|^2 \\ |E_X^{(I)}|^2 \end{bmatrix}, \tag{6.7.17}$$

which is denoted with power values by

$$\begin{bmatrix} P_C^{(O)} \\ P_X^{(O)} \end{bmatrix} = \begin{bmatrix} \langle R_{CC} \rangle & \langle R_{CX} \rangle \\ \langle R_{XC} \rangle & \langle R_{XX} \rangle \end{bmatrix} \cdot \begin{bmatrix} P_C^{(I)} \\ P_X^{(I)} \end{bmatrix}. \tag{6.7.18}$$

A section of the network model for the reflection is indicated in Figure 6.7.2.

Including with the reflection a path loss factor proportional to $d^{-\alpha}$, which is taken to be the same for each polarisation, results in

$$\begin{aligned} \begin{bmatrix} P_C^{(O)} \\ P_X^{(O)} \end{bmatrix} &= \begin{bmatrix} R_{CC} & R_{CX} \\ R_{XC} & R_{XX} \end{bmatrix} \begin{bmatrix} d^{-\alpha} & 0 \\ 0 & d^{-\alpha} \end{bmatrix} \begin{bmatrix} P_C^{(I)} \\ P_X^{(I)} \end{bmatrix} \\ &= d^{-\alpha} \begin{bmatrix} R_{CC} & R_{CX} \\ R_{XC} & R_{XX} \end{bmatrix} \begin{bmatrix} P_C^{(I)} \\ P_X^{(I)} \end{bmatrix}. \end{aligned} \tag{6.7.19}$$

Successive reflections can be expressed as a cascade of scattering matrices,

$$\begin{bmatrix} E_{\mathrm{C}}^{(\mathrm{O})} \\ E_{\mathrm{X}}^{(\mathrm{O})} \end{bmatrix} = \left(\left(d^{(1)} \right)^{-\alpha_1} \left(d^{(2)} \right)^{-\alpha_2} \cdots \right)^{1/2} \cdot \begin{bmatrix} S_{\mathrm{CC}} & S_{\mathrm{CX}} \\ S_{\mathrm{XC}} & S_{\mathrm{XX}} \end{bmatrix}^{(1)} \cdot \begin{bmatrix} S_{\mathrm{CC}} & S_{\mathrm{CX}} \\ S_{\mathrm{XC}} & S_{\mathrm{XX}} \end{bmatrix}^{(2)} \cdots \begin{bmatrix} E_{\mathrm{C}}^{(\mathrm{I})} \\ E_{\mathrm{X}}^{(\mathrm{I})} \end{bmatrix}. \tag{6.7.20}$$

This approach is the tracing of rays from a source, e.g. $E_{\mathrm{C}}^{(\mathrm{I})}$, to a destination, e.g. $E_{\mathrm{X}}^{(\mathrm{O})}$. In multipath propagation, there is no guarantee that the 'output' wave, which we consider is a wave incident on the receiving antenna, consists of only this ray. Other rays arriving from the same angle may be present as well, so that a given incident wave will, in general, comprise indistinguishable (by a receiver) contributions from many different ray paths. However, if the rays have resolvable delay time differences, then they can indeed be considered as separate effective scatterers (having the same direction, but different delays) by the receiver. As noted in Chapter 5, a non-zero bandwidth of the receiver acts to combine the rays from bins of delay time, and the non-zero angular resolution of the receiving antenna similarly acts to sum the ray contributions within angular bins.

In general scattering, there is no relationship between the reflection coefficients of each polarisation, and these are functions of the incident and scattering angle. However, for reflections from smooth, planar surfaces, the relationship $R_{\mathrm{CX}} = 1 - R_{\mathrm{CC}}$ can be used as follows.

6.7.2.2 Reflections from smooth, planar surfaces

After N lossless reflections from a smooth planar surface, where the reflection coefficients are independent, the output powers of each polarisation are proportional to (the space loss, assumed the same for each polarisation, is omitted here)

$$\begin{bmatrix} P_{\mathrm{C}}^{(\mathrm{O})} \\ P_{\mathrm{X}}^{(\mathrm{O})} \end{bmatrix} = \begin{bmatrix} R_{\mathrm{CC}} & 1 - R_{\mathrm{XX}} \\ 1 - R_{\mathrm{CC}} & R_{\mathrm{XX}} \end{bmatrix}^{N} \cdot \begin{bmatrix} P_{\mathrm{C}}^{(\mathrm{I})} \\ P_{\mathrm{X}}^{(\mathrm{I})} \end{bmatrix}, \tag{6.7.21}$$

in which the mean operator $\langle \cdot \rangle$ has been dropped. We are now effectively representing each reflection with a mean reflection coefficient value. This value can be drawn from a distribution (e.g. lognormal). This equation produces an XPD which converges to

$$\mathrm{Lim}\,|_{N \to \infty} XPD = \frac{P_{\mathrm{C}}}{P_{\mathrm{X}}} = \frac{1 - R_{\mathrm{XX}}}{1 - R_{\mathrm{CC}}}. \tag{6.7.22}$$

Figure 6.7.3 gives some examples of the powers in each polarisation for given power reflection coefficients, when the excitation is purely co-polarised, i.e. $\left[P_{\mathrm{C}}^{(\mathrm{I})}, P_{\mathrm{X}}^{(\mathrm{I})} \right] = [1, 0]$.

When the mean co- and cross-polar reflection coefficients are the same, then the power converges to be shared equally in each polarisation giving $XPD = 1$ after

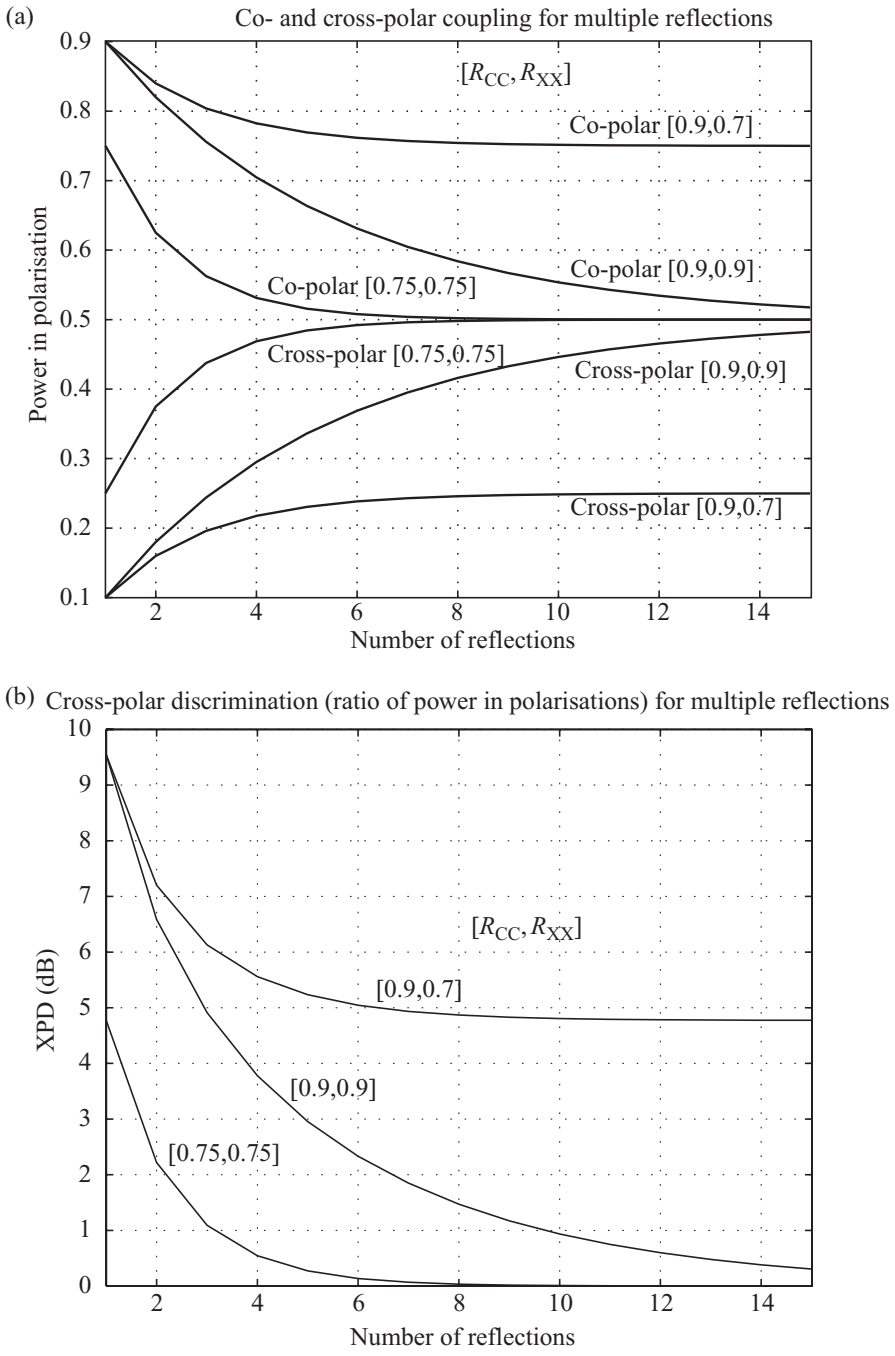


Figure 6.7.3 *(a) Points with straight-line interpolation for the co- and cross-polarised power after a number of lossless reflections with the mean reflection coefficients. (b) The XPD is the ratio of the co- and cross-polarised powers, converging to 0 dB if the co-polar and cross-polar reflection coefficients are the same and less than one. Values of the power reflection coefficients [R_{CC}, R_{XX}] examples are given on the graph*

several reflections. The convergence rate is determined by the size of the reflection coefficients. For the example $R_{XX} = R_{CC} = 0.75$, the XPD has essentially converged after eight reflections. When the co- and cross-polar reflection coefficients are different, then the power does not end up shared equally between polarisations in general. For the example $R_{CC} = 0.9$, $R_{XX} = 0.7$, the co-polar power converges to be 0.75 of the total power and the cross-polar power converges to 0.25, giving $XPD = 3$. When the power coefficients R_{XX} and R_{CC} are less than 0.5, the plot oscillates as the majority of power changes polarisation at each reflection.

6.7.3 Uncorrelated scatterers

Recall that, in general, the effective scatterers (the antenna effect is included) are a linear combination of physical scatterers' wave contributions. For a moving terminal, the time (i.e. distance z, for lineal trajectories) variable is the decorrelating factor for the physical and effective sources. The non-zero bandwidth is also a decorrelating factor. If there is non-zero bandwidth and a moving terminal, then both these factors act independently to decorrelate 'the scatterers'.

The formulation for a simple example helps to recall and fix ideas regarding the effective scatterers. Consider a physical scatterer denoted $\boldsymbol{A}_1$. This has a direction (θ_1, ϕ_1) relative to the receiving antenna which is at position z, and moving in the direction of z. This physical scatterer is assumed to be in the far field of the receiving antenna and vice versa. The distance to the scatterer is r_1, which corresponds to a delay $\tau_1 = r_1/c$. However, this 'physical delay' is not of direct importance: in general, there may be waves with different delay times emanating from the same physical scatterer. These waves are written (for m delays)

$$\boldsymbol{E}_1 = \sum_m \boldsymbol{E}_{1m}(\theta_1, \phi_1; \tau_m)\,. \tag{6.7.23}$$

For two delays, for example, the horizontally and vertically polarised waves are written

$$\begin{aligned} a(\theta, \phi; \tau) = {} & E_{\theta 11}(\theta_1, \phi_1; \tau_1) + E_{\phi 11}(\theta_1, \phi_1; \tau_1) \\ & + E_{\theta 12}(\theta_1, \phi_1; \tau_2) + E_{\phi 12}(\theta_1, \phi_1; \tau_2) \end{aligned} \tag{6.7.24}$$

where the (single-direction) form is, for example,

$$E_{\theta 11}(\theta_1, \phi_1; \tau_1) = E_{\theta 11}\delta(\theta - \theta_1)\delta(\phi - \phi_1)\mathrm{e}^{\mathrm{j}\omega(\tau - \tau_1)}. \tag{6.7.25}$$

The $\sin\theta$ term involved in integration over the spherical coordinates is embedded in the definition of the delta function in θ. The open circuit voltage of an ideal antenna is the sum of contributions weighted by the antenna pattern. So in this 'wave-gathering'

process, the vector form of the waves from the physical scatterers is converted into the scalar form of the impulse response. The open circuit voltage is proportional to

$$\begin{aligned} a(\theta, \phi; \tau) = & \sum_{n \text{ directions}} \sum_{m \text{ delays}} \{E_{\theta mn}(\theta_n, \phi_n; \tau_m) \cdot h(\theta_n, \phi_n) + E_{\phi mn}(\theta_n, \phi_n; \tau_m) \\ & \cdot h(\theta_n, \phi_n)\} \\ = & \sum_{n \text{ directions}} \sum_{m \text{ delays}} \left\{a_{\theta nm} + a_{\phi nm}\right\}, \end{aligned} \tag{6.7.26}$$

where any frequency dependence of the incident fields (other than that caused by the distribution of delays) and of the receiving pattern is implicit, and similarly the dependence of the incident fields on the receiver's position is implicit.

For the example of the single physical scatterer ($n = 1$) with $m = 2$ delays, signals comprising the impulse response are written

$$\begin{aligned} a(\theta, \phi; \tau) = & E_{\theta 11}(\theta_1, \phi_1; \tau_1) h_\theta(\theta_1, \phi_1) + E_{\phi 11}(\theta_1, \phi_1; \tau_1) h_\phi(\theta_1, \phi_1) \\ & + E_{\theta 12}(\theta_1, \phi_1; \tau_2) h_\theta(\theta_1, \phi_1) + E_{\phi 12}(\theta_1, \phi_1; \tau_2) h_\phi(\theta_1, \phi_1) \\ = & a_{\theta 11} + a_{\phi 11} + a_{\theta 12} + a_{\phi 12}. \end{aligned} \tag{6.7.27}$$

For a receiving antenna which is moving in the z-direction ($\theta = 0$), the ϕ variable integrates out to produce

$$a(\theta, \tau) = \int_{\phi=0}^{2\pi} a(\theta, \phi; \tau)\, \mathrm{d}\phi \tag{6.7.28}$$

and this is the equivalent of the effective scattering distribution $a(u, \tau)$ in which $u = k_C \cos\theta$, namely

$$a(u, \tau) = a(\theta, \tau) \cdot \frac{\partial \theta}{\partial u} = a\left(\cos^{-1}(u/k_C), \tau\right) \cdot \frac{1}{\sqrt{k_C^2 - u^2}}. \tag{6.7.29}$$

Note that the wave contributions (from the physical scatterers) which have a common (i.e. unresolvable) θ and τ are integrated in ϕ and so their ϕ resolution is irretrievably lost. An implication of this process is that components of the effective scattering distribution are, in general, sums of random contributions and their statistical distribution (Gaussian in the limit of many summations) may be different from the distribution of the random contributions.

6.7.3.1 Fourier resolution

The Fourier resolution gives a feel for the size of the summation bins. As an example, consider a trajectory of $z_0 = 5$ wavelengths. This gives a Fourier resolution of

$\Delta u \approx 1/z_0$, which corresponds to zenith angle bins $\Delta\theta$ ranging from about 2° in the direction of motion ($\theta = 0$) to about 14° in the broadside direction ($\theta = 90°$). Similarly for a bandwidth of say $\Delta f = 30\,\text{kHz}$, the Fourier resolution of the delay time is $\Delta\tau \approx 1/\Delta f = 33\,\mu\text{s}$, corresponding to a path length resolution of about $\Delta r \approx 10\,000\,\text{m}$. So waves arriving from, for example, all broadside angles (all ϕ) in a zenith angle bin $\Delta\theta \approx (83° < \theta < 97°)$ and a path difference bin $\Delta r \approx c\Delta\tau = 10\,000\,\text{m}$ are lumped together into a single effective scatterer in this example. (It is assumed that the antenna pattern is fixed and covers the field-of-view to the physical scatterers in this example.)

6.7.3.2 Measurement by time series correlation

An element of the polarisation matrix above can be estimated, assuming ergodic signals, by

$$\Gamma'^{(T)}_{\phi\phi}(\theta_1, \theta_2; \phi_1, \phi_2) = \int_0^{2\pi} \int_0^{\pi} \frac{1}{T} \int_0^{T} E_\phi(\theta_1, \phi_1, t)\; E^*_\phi(\theta_2, \phi_2, t)\; \mathrm{d}t\; \mathrm{d}\theta\; \mathrm{d}\phi. \tag{6.7.30}$$

This shows explicitly the cross-correlation (with zero time lag) of the time varying contents of the horizontally polarised rays for a duration of T. The time can be equivalently written as distance for a given speed and lineal trajectory of the receiving antenna. If the antenna is rotated, however, for example as can be imagined for a standing or sitting human using a handheld terminal, then the angular bearing, rather than distance, for the mobile antenna is the function of time. The 'uncorrelated scatterers' assumption is that this cross-correlation is always zero when the angle (θ_1, ϕ_1) is different to (θ_2, ϕ_2).

To confirm whether this is actually the case by cw measurement experiment is tricky. Using highly directive antennas at a fixed bearing angle separation can give the time waveforms as the antenna platform is moved. The high directivity is in order to resolve the effective scatterers in a single polarisation (defined by the antenna) as representing the two bearings. These waveforms can then be correlated to establish how uncorrelated the angularly separated signals are. The antenna arrangement may be workable for outdoor scenarios and high carrier frequencies (i.e. physically small antennas) but for indoor situations the physical size and co-location of the directive antennas becomes the limiting problem for such a measurement setup.

The measured waveforms are those belonging to the radio channel and the experimental correlation is over the received signals (flat-fading frequency transfer function)

$$r_i(t) = \int\int \boldsymbol{E}(\theta, \phi, t) \cdot \boldsymbol{h_i}(\theta, \phi)\; \mathrm{d}\theta\; \mathrm{d}\phi \tag{6.7.31}$$

where $\boldsymbol{h_i}$ is the ith antenna receiving pattern, i.e. the correlation is

$$R^{(T)}_{ij} = \frac{1}{T} \int_0^{T} r_i(t) r^*_j(t)\; \mathrm{d}t. \tag{6.7.32}$$

This corresponds to the element of the polarisation matrix, in the polarisation of the antenna, only if the incoming ray directions are properly resolved by the directive

antennas. For example, for a vertically polarised antenna, $i = \theta$, at bearing (θ_1, ϕ_1) and a horizontally polarised antenna at bearing (θ_2, ϕ_2),

$$R_{ij} = \Gamma_{\theta\phi} \tag{6.7.33}$$

if the incident waves are properly resolved by the antennas.

6.7.3.3 Correlation over a bandwidth

For increasing bandwidth, the signal borne by the ray becomes increasingly frequency selective from the cw case. Here, we can correlate the frequency-dependent behaviour of the signals borne by the rays. This can be written in exactly the same way as for the time correlation, here for a bandwidth of $2B$ rad

$$\Gamma'^{(B)}_{\phi\phi}(\theta_1, \theta_2; \phi_1, \phi_2) = \int_0^{2\pi} \int_0^{\pi} \frac{1}{2B} \int_{-B}^{B} E^*_{\phi}(\theta_1, \phi_1, \omega)\, E^*_{\phi}(\theta_2, \phi_2, \omega)\, \mathrm{d}\omega\, \mathrm{d}\theta\, \mathrm{d}\phi. \tag{6.7.34}$$

For very wideband signals, the postulate that waves arriving from different directions have arrived via different reflections allows us to scan in angle (as well as with distance, if required) for establishing experimentally the correlation values. Experiments would refer, as noted above, to the scatterers in the polarisation of the receiving antenna and the directive antennas would need to resolve the two ray directions. Also, the antenna is taken as independent of frequency and so the experimental signals are, in a static scenario (stationary mobile, fixed sources),

$$H_i(\omega) = \int \int \boldsymbol{E}(\theta, \phi, \omega) \cdot \boldsymbol{h}_i(\theta, \phi)\, \mathrm{d}\theta\, \mathrm{d}\phi \tag{6.7.35}$$

and the correlation is

$$R^{(B)}_{ij} = \frac{1}{2B} \int_{-B}^{B} H_i(\omega)\, H^*_j(\omega)\, \mathrm{d}\omega, \tag{6.7.36}$$

and this can be equivalently expressed in terms of the impulse responses, $h(\tau)$ (note that $h(\tau)$ is different to the receiving pattern $\boldsymbol{h}(\theta, \phi)$), i.e.

$$R^{(B)}_{ij} = \int_0^{\infty} h_i(\tau) h^*_j(\tau)\, \mathrm{d}\tau. \tag{6.7.37}$$

6.7.3.4 Correlation over time and bandwidth

For combining the finite bandwidth effects and the moving terminal, the angular correlation is

$$\Gamma'_{\phi\phi}(\theta_1, \theta_2; \phi_1, \phi_2) = \int_0^{2\pi} \int_0^{\pi} \frac{1}{T} \int_0^{T} \frac{1}{2B} \int_{-B}^{B} E_{\phi}(\theta_1, \phi_1, \omega, t)\, E^*_{\phi}(\theta_2, \phi_2, \omega, t)\, \mathrm{d}\omega\, \mathrm{d}t\, \mathrm{d}\theta\, \mathrm{d}\phi, \tag{6.7.38}$$

and an experimental measure for this, which necessarily includes antenna effects, can be expressed

$$R_{ij} = \frac{1}{2B}\int_0^\infty \int_{-B}^{B} H_i(\omega, t)\; H_j^*(\omega, t)\; \mathrm{d}\omega\; \mathrm{d}t \tag{6.7.39a}$$

$$= \frac{1}{2B}\int_0^\infty \int_0^\infty h_i(\tau, t)\; h_j^*(\tau, t)\; \mathrm{d}\tau\; \mathrm{d}t, \tag{6.7.39b}$$

where $H_i(\omega, t)$ is the time varying radio channel transfer function for the ith antenna, and similarly $h_i(\tau, t)$ is the associated time varying impulse response.

6.7.4 Gaussian scenario model

A statistically appealing form for the angular distribution of same-polarised waves is the Gaussian distribution about some direction $(\theta_\mathrm{m}, \phi_\mathrm{m})$ which may represent the nominal line-of-sight direction or at least the nominal centre of a bundle of rays. The distribution is of course restricted to the support of the coordinate system, i.e.

$$S(\theta, \phi) \propto \exp\left[-\left(\frac{(\theta-\theta_m)^2}{2\sigma_\theta^2} + \frac{(\phi-\phi_m)^2}{2\sigma_\phi^2}\right)\right], \quad 0 \le \phi < 2\pi, \;\; 0 \le \theta \le \pi, \tag{6.7.40}$$

and the angular spread is given by $\sigma_\theta, \sigma_\phi$. For a single incident signal, $(\sigma_\theta, \sigma_\phi) \to (0, 0)$, and

$$S^{(\mathrm{s})}(\theta, \phi) = \delta(\theta - \theta_m)\delta(\phi - \phi_m). \tag{6.7.41}$$

For uniformly distributed incident signals, $(\sigma_\theta, \sigma_\phi) \to (\infty, \infty)$, a field-of-view is defined by angular upper and lower limits,

$$S(\theta, \phi) = S^{(\mathrm{u})}, \quad \theta_\mathrm{l} \le \theta \le \theta_\mathrm{u}, \quad \phi_\mathrm{l} \le \phi \le \phi_\mathrm{u};$$
$$= 0, \quad \text{elsewhere;} \tag{6.7.42}$$

where

$$S^{(\mathrm{u})} = \frac{1}{(\theta_\mathrm{u} - \theta_\mathrm{l})(\phi_\mathrm{u} - \phi_\mathrm{l})}. \tag{6.7.43}$$

The uniform field model is a handy simplification to the Gaussian distribution for scenarios.

6.7.5 Laplacian scenario model

Although the Gaussian-shaped model for the scenario appeals from a modelling viewpoint, measurements at 1800 MHz in both rural (mostly with line-of-sight) and urban (mostly with no line-of-sight) locations are reported by Pedersen *et al*. (1987) which support the Laplacian distribution being a better fit than the Gaussian distribution, although the Gaussian was a better fit for the occasional instantaneous (unaveraged) measurement. The Laplacian is the double-sided exponential,

$$S(\theta, \phi) = \frac{c}{\sqrt{2}\sigma}\mathrm{e}^{-(\sqrt{2}/\sigma)|\phi|}, \quad -\pi \le \phi < \pi, \tag{6.7.44}$$

where $c = (1 - \exp(-\sqrt{2}\pi/\sigma))^{-1}$ is the normalising constant with σ controlling the spread. For a rural environment with a base station height of 47 m, Pedersen *et al.* (1987) report a spread of 1.8°; and for the urban environments, spreads (and respective base-station heights) of 5° (32 m) and 10° (20 m, which was also the rooftop height). As shown in Chapter 4 there is no physical basis for the Laplacian distribution, and other distributions like

$$S(\phi) = \frac{c}{\sin^2 \phi + c} \tag{6.7.45}$$

(c is a constant) give a better fit to the apparent distribution of power. Only super resolution techniques can resolve the true distribution of power.

6.7.6 Vehicular mobile

Much of the scenario development in mobile communications has been with vehicular terminals for landmobile communications in mind. The antenna, if fixed to the vehicle, can then be referenced to the environment, at least in the elevation angle sense. This assumption breaks down somewhat where there are steeply canted or hilly roads, of course. The following is a brief review of models proposed for vehicular mobiles, which are naturally outdoor models.

Clarke's (1966) popular model is the two-dimensional ring source, whose pdf is

$$S^{(\mathrm{C})}(\theta, \phi) = S^{(\mathrm{C})}(\theta) = \delta\left(\theta - \frac{\pi}{2}\right), \tag{6.7.46}$$

where the mobile is located at the centre of the ring. The model is very simple and effective. The Doppler spectrum is a reasonable fit to experimental results. For simple simulations, a minimum of about ten sources on the ring with similar (or equal) amplitudes and uniformly distributed phases, is sufficient for most simulation purposes.

Aulin (1972) added a single ray from an elevated direction, $(\theta_0\phi_0)$, to Clarke's model, arguing that such a ray could represent a line-of-sight or dominant path to a mobile, and this was claimed to improve the fit to experimental Doppler spectra for vehicular terminals. The scenario, with the line-of-sight component as a fixed direction, is

$$S^{(\mathrm{A})}(\theta, \phi) \propto \left\{\delta\left(\theta - \frac{\pi}{2}\right) + K\delta(\theta - \theta_0)\delta(\phi - \phi_0)\right\}, \tag{6.7.47}$$

where K is a proportionality constant. Aulin also proposed a 3D averaged scenario which is uniform in azimuth and has an almost-uniform distribution over an angular support of $2\beta_\mathrm{a}$, centred on the horizon. Expressed using the elevation angle from the horizon, β, it is

$$\begin{aligned} S^{(\mathrm{A})}(\beta, \phi) &\propto \frac{\cos\beta}{\sin\beta_\mathrm{a}}, \quad |\beta| \leq |\beta_\mathrm{a}| \leq \frac{\pi}{2}; \\ &= 0, \quad \text{elsewhere.} \end{aligned} \tag{6.7.48}$$

This was chosen for its advantage of a closed form of integration for finding the Doppler spectrum. Parsons (1992) gives a similar form, uniform in azimuth but having an elevation distribution which is a cosine form, again centred on the horizon,

$$\begin{aligned} S^{(\mathrm{P})}(\beta, \phi) &\propto \cos\left(\frac{\pi}{2}\frac{\beta}{\beta_{\mathrm{a}}}\right), \quad |\beta| \leq |\beta_{\mathrm{a}}| \leq \frac{\pi}{2}; \\ &= 0, \quad \text{elsewhere.} \end{aligned} \tag{6.7.49}$$

Neither of these appear to have experimental justification from direction-of-arrival measurements and no specific support size for the elevation is proposed. Vaughan (1986) used scant published experimental results for direction-of-arrival, to propose uniformity in azimuth (cf. Clarke) and uniformity in elevation from the horizon up to 30° above the horizon. This corresponds to $\theta_{\mathrm{l}} = 90°, \theta_{\mathrm{u}} = 60°; \phi_{\mathrm{l}} = 0°, \phi_{\mathrm{u}} = 360°$, in the above uniform case for the Gaussian expression. This averaged scenario has no incident waves from below the horizon, which is clearly a possibility for ground-bounce type rays and an elevated antenna, although roof-mounted monopole type antennas normally have low gain to these regions. Such a model contrasts with the Aulin and Parsons models, which both have half of the incident power arriving from below the horizon. Chao *et al.* (1991), also using experimental direction-of-arrival evidence, concluded a similar uniform scenario, but extending to 60° in elevation (catering for high-rise urban environment), i.e. corresponding to $\theta_{\mathrm{l}} = 90°, \theta_{\mathrm{u}} = 60°; \phi_{\mathrm{l}} = 0°, \phi_{\mathrm{u}} = 360°$.

In summary, the azimuthal distribution of a time-averaged scenario is considered omnidirectional. The averaged elevation distribution can be argued to be Gaussian-like from a statistical viewpoint, truncated of course by the support of the coordinate system. This can be simplified to a uniform distribution. The extent of the distribution in any experiment will depend on the type of environment. For example, averaging in a high-rise city may produce higher elevation angles than averaging in suburban areas as noted above. Also the azimuthal distribution may exhibit higher powers from the 'forward' and 'backwards' directions, corresponding to the waveguide-like propagation of power along urban canyons. The averaged scenario attempts to cover many different environments. With the vehicular mobile above, the antenna elevation reference can be considered fixed with respect to the physical environment. This means that the scenarios discussed here represent physical elevation angles. The movement of the mobile antenna can be included to produce the scenario for terminals whose elevation reference changes.

6.7.7 Personal terminals

With personal terminals, it seems reasonable to maintain the assumption of uniformity in azimuth for the averaged scenario of incident waves. The natural movement of a personal terminal (specifically, a human with a portable phone) is considerable in the elevation direction as well. This movement of the terminal results in the receiving pattern convolving with the instantaneous scenario to produce the received signal. This convolution can be expressed in simplified form, where $\alpha(t)$ is the time-varying

elevation angle of the antenna, $\boldsymbol{h}(\theta)$ is the elevation pattern of the antenna and $\boldsymbol{E}(\theta)$ is the static incoming wave distribution in geographical coordinates. The time varying transfer function is then

$$h(t) = \int_0^{\pi} \boldsymbol{E}\,(\theta - \alpha(t)) \cdot \boldsymbol{h}(\theta)\,\mathrm{d}\theta, \tag{6.7.50}$$

in which the frequency dependencies of the antenna and transfer function are understood. The scenario that is referenced to the receiver coordinates thus has a larger elevation angle extent than the fixed coordinate scenario of the environment. Also, personal terminals are used indoors, and it is likely that the elevation angles of the incoming waves extend further than for the outdoor vehicular scenarios. Some experimental results are indicated in Chapter 4. Without experimental evidence, a convenient scenario for personal receivers is the entire sphere, here also made uncorrelated in polarisation and angle, which is

$$S^{(h)}(\theta, \phi) = S^{(h)} = \frac{1}{4\pi}. \tag{6.7.51}$$

6.7.8 Base stations

There are several coverage requirements for base stations depending on the type of cell or coverage area.

Broadcast-type coverage for a wide area field-of-view typically has the base station antenna well elevated above the scatterers. On reasonably flat terrain, the fields illuminating such a base station will clearly be from below the horizon relative to the base station height. Matching the base station antenna patterns to the scenario (by 'tilting' the pattern downwards), is motivated not only by a reduction of the path loss (from the increased base station antenna gain) but also by a decrease in the delay spread and less interference with distant stations using the same or adjacent frequencies. The azimuthal coverage is tailored for the application, in some layouts being omnidirectional and in others, sectorised. For the elevation distribution, an ideal situation is given in Figure 6.7.4, where an elevated base station is illuminated by a

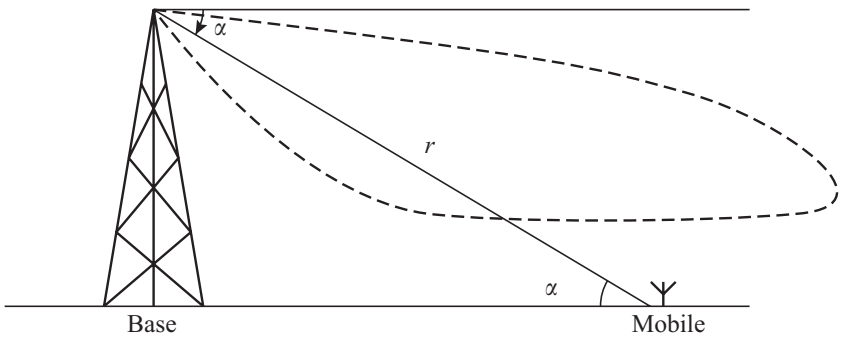


Figure 6.7.4 An elevated base station with a shaped pattern for broadcast-type coverage over flat terrain

uniform distribution of sources on flat terrain. The base-station antenna is at height h and the direct wave range is r; α denotes the elevation angle from the mobile to the base station and is also used as an elevation angle coordinate at the base, referring to the angle from the horizontal down to the direction to the mobile.

If the mobiles on the terrain radiate isotropically over the elevation angles of interest and their radiated power is also constant (say P_m), then the elevation distribution of power incident at the base station is

$$P(\alpha) \propto \frac{P_m}{r^n} \propto \sin^n \alpha, \quad 0 \leq \alpha \leq \frac{\pi}{2}. \tag{6.7.52}$$

The decay exponent lies in the range $2 < n < 4$ (see Chapter 4), where n is two for the time-averaged incident field distribution for the elevation direction where only the direct wave propagation is considered. The 'matched' base-station power pattern is $\text{cosec}^n \alpha$, which then provides constant power density to the flat terrain independent of range. Even though this is an idealised situation, it gives a useful design guideline for large cell sizes. The scenarios are again considered uncorrelated in angle and in polarisation.

If local scattering about the mobile is considered, a distributed scenario over the local region occupied by the mobile comprises the illumination. This is illustrated in Figure 6.7.5, where the scenario comprises a narrow solid angle containing a distribution which, in a time-average sense, is uncorrelated in angle and between polarisations.

For the azimuthal power distribution, the Gaussian distribution above again appeals for modelling the time-averaged scenario, with the mean direction not being part of the averaging for illumination by a single mobile. The cosine distribution was used by Lee (1973), with

$$S(\phi) \propto \cos^n(\phi - \phi_0), \quad \phi_\text{m} \leq \phi \leq \frac{\pi}{2} + \phi_\text{m}, \tag{6.7.53}$$

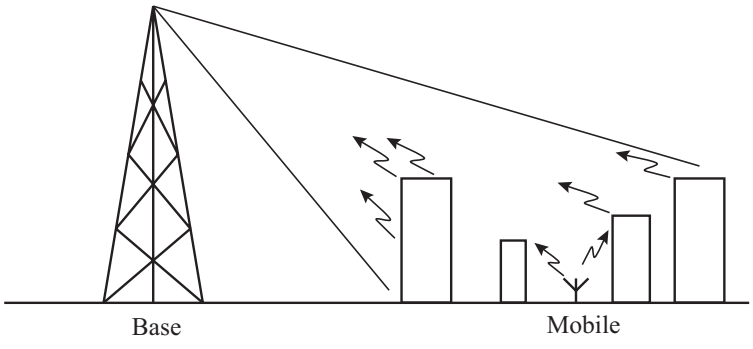


Figure 6.7.5 The local scatterers around a mobile terminal provide a scenario seen by a large-cell base station which can occupy a small part of the field-of-view

where ϕ_m is the mean direction, and n determining the beamwidth. This is a close approximation to the Gaussian in the main lobe region. The uniform scenario has been used by Salz and Winters (1994) for this situation.

For micro-cell-type base stations, the antennas are often wall-mounted, or, for indoor situations, sometimes ceiling-mounted. For a wall-mounted base station, the incident wave distribution automatically excludes the 'rear' hemisphere, in principle, which is occupied by the wall. However, such an excluded region for the scenario does not necessarily translate into a 'don't care' region for matched antenna patterns. For indoor applications, for example, the base-station radiation in the direction of the wall (or ceiling) on which the antenna is mounted can be important since it may cause interference with adjacent cells. This is particularly the case when different floors of a building have frequency reuse and employ ceiling-mounted antennas. Again, the scenarios are assumed uniform over the field of coverage and uncorrelated with angle and between polarisations. A difference between micro-cell scenarios and large-cell or broadcast-type scenarios, is that the micro-cell base-station field-of-view can be fully occupied, instead of partially occupied, by the scenario from a single mobile terminal. Indoors is again a good example, where the multipath is ubiquitous.

6.7.9 Mobile satellite

6.7.9.1 Geostationary

This is normally a line-of-sight service, and the situation may be viewed as a terrestrial cell with an elevated base station with a line-of-sight. For the vehicular terminal, the elevation angle to the satellite is essentially fixed, although steep hills and canted roads can create some variation as mentioned above. For some boats the pitch and yaw statistics for this variation are known, but this type of information does not seem to be readily available for land vehicles.

For urban/suburban situations, the multipath components can be accounted for by including the diffuse scenarios of the previous sections. The polarisation is circular, and if the diffuse components comprise mainly single reflections off the ground, then there is a good chance that much of the diffuse power is cross-polarised. If the antenna has good polarisation purity in the lower elevation angles, then the multipath incident on the antenna will be reduced by the polarisation mismatch.

The mobile can again be assumed to have uniformly distributed azimuthal bearing. If the elevation angle to the satellite is α then the polar zenith angle is $\theta_0 = \pi/2 - \alpha$, and the scenario can be written

$$S(\theta, \phi) = k_{\text{Rice}}\delta(\theta - \theta_0) + (1 - k_{\text{Rice}})S^{(U)}, \tag{6.7.54}$$

where k_{Rice} is the Rice factor for total power in the scenario. An instantaneous scenario and its time- or distance-averaged form are depicted in Figure 6.7.6.

The upper zenith angle limit θ_u can be taken as 30° or less. Allowing for vehicle pitch and yaw, and to cover for any significant latitude changes (such as for long-haul trucks operating in the latitudinal direction), the limits for the direct path need to be extended appropriately.

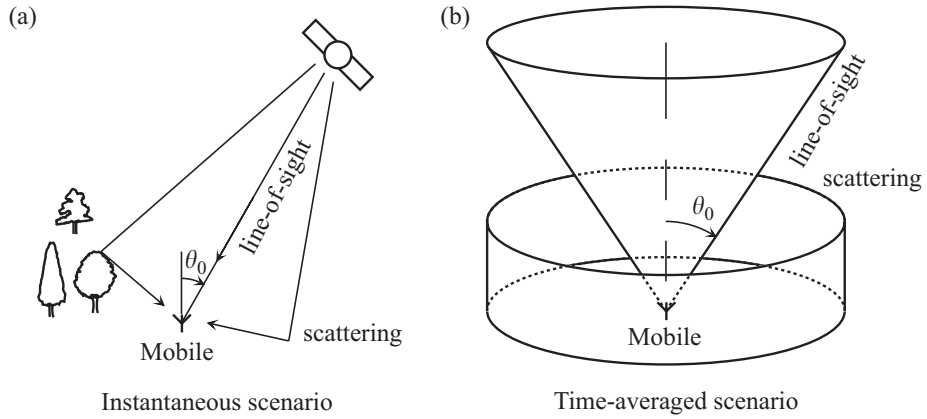


Figure 6.7.6 *(a) An instantaneous scenario for a geosynchronous satellite link to a vehicular mobile. A direct line-of-sight is combined with scattered contributions. (b) The time-averaged form, where the mobile takes on all azimuthal bearings, but has a fixed elevation angle reference (no pitch and yaw of the vehicle)*

6.7.9.2 Non-geostationary

Recent development of low-earth-orbit (LEO) satellite constellations has enabled global coverage for personal communications. The global positioning system (GPS) was the first such constellation that provided a navigation global service. For deriving the time-averaged source (defined from line-of-sight to the satellite) scenario, both the constellation and the terrestrial terminal motion must be considered. The mobile terminal can be taken as being equilikely in its azimuthal bearing. The satellites are assumed to have a uniform angular distribution referenced to the orbit centre, i.e. the satellites are taken as being equilikely on the circle of their orbit. When including the motion of the terrestrial terminal, the satellite locations, with respect to the terminal coordinates, are taken to be uniformly distributed on a sphere containing the orbit(s). Assuming polarisation matching, the space loss with respect to elevation angle and the satellite gain pattern become the primary parameters for the scenario (Vaughan and Scott, 1993).

A terrestrial terminal and a satellite are depicted in Figure 6.7.7.

The mean orbital radius is denoted r_s, and the mean radius to the terminal (earth radius) is r_e and their ratio is

$$r_n = \frac{r_e}{r_s}. \tag{6.7.55}$$

The distance d between the terminal and satellite is a function of the zenith angle referenced to the terminal, θ. The space-loss factor is then

$$[d(\theta)]^{-2} = \left[r_s^2 + r_e^2 - 2r_s r_e \cos\beta(\theta)\right]^{-1}, \tag{6.7.56}$$

where

$$\cos\beta(\theta) = r_n \sin^2\theta + \cos\theta\sqrt{1 - r_n^2 \sin^2\theta}, \quad |\theta| < \frac{\pi}{2}. \tag{6.7.57}$$

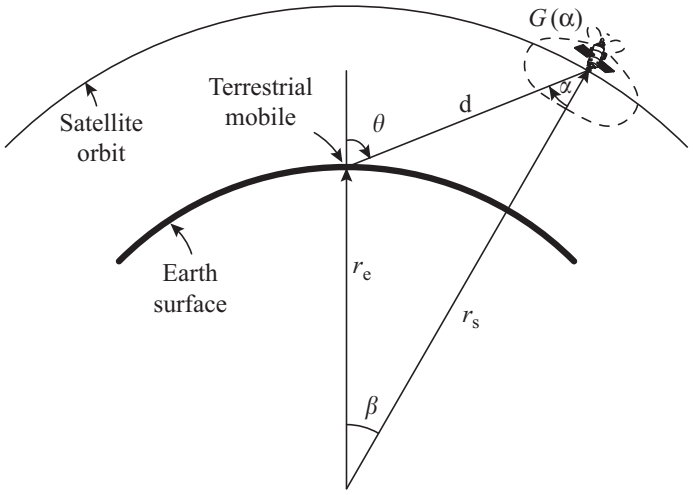


Figure 6.7.7 An instantaneous scenario for a satellite link to a mobile

The satellite power pattern, $G_S(\alpha)$, is taken to be rotationally symmetric about its zenith angle, α, to the centre of orbit ($\alpha = 0$) and so

$$\alpha(\theta) = \theta - \beta(\theta). \tag{6.7.58}$$

The incident power density at the terminal is therefore proportional to the space loss factor and the satellite gain in the elevation plane of the terminal,

$$S^{(0S)}(\theta, \phi) = S^{(S)}(\theta) = [d(\theta)]^{-2} \cdot G_S(\alpha(\theta)). \tag{6.7.59}$$

So for an omni-satellite beam, the incident power density is simply proportional to the space-loss factor.

The angular density of the satellites with respect to the angle β, which is assumed constant and denoted $K(\beta) = K$, can be transformed to be a function of θ, using

$$\frac{\partial \beta}{\partial \theta} = -\frac{2r_n \sin\theta\cos\theta - \sin\theta\sqrt{1 - r_n^2\sin^2\theta} - r_n^2\cos^2\theta\sin\theta/\sqrt{1 - r_n^2\sin^2\theta}}{\sqrt{1 - \left(r_n\sin^2\theta + \cos\theta\sqrt{1 - r_n^2\sin^2\theta}\right)^2}}. \tag{6.7.60}$$

The averaged satellite density, as seen by the mobile terminal, can then be written

$$S^{(SD)}(\theta) = K\left|\frac{\partial \beta}{\partial \theta}\right|. \tag{6.7.61}$$

In a multipath situation, this density (representing independent sources) would be multiplied into the incident power pdf. But a satellite constellation provides services which may require only one 'visible' satellite while other services may require several satellites being simultaneously visible. The space-loss factor and patterns to compensate for it are covered in Section 8.3.

6.8 A diffuse model for outdoor environments

The Clarke scenario is an example of a quasi-physical situation for modelling the spatial channel. This section describes a model that incorporates some directional information from some of the physical scattering processes in the channel. In this sense it sits in between the deterministic approach of ray-tracing and purely statistical models which contain no angular insight. This is an extension of the description given in Chapter 4.

The model is for an elevated base station and mobile above a planar ground plane which contains uniformly dense scatterers. For the base-station transmit situation, power arrives at the mobile only from the scatterers on the ground. An omnidirectional antenna is assumed at the mobile so that the effective scatterers relate directly to the propagation model scatterers. The base station and mobile antenna above a ground plane are depicted in Figure 6.8.1. An ellipsoid has its foci at the base-station antenna and mobile antenna. The ellipsoid is the locus of a given delay time from any effective scatterer. The scatterers are confined to the ground plane, so the locus of the scatterers in the model is on the interception of the ellipsoid and the ground plane. This locus is an ellipse. Consequently the scatterers seen by the base station and the mobile, at given delay time, occupy an ellipse. The position of the base and mobile relative to this ellipse is complicated. This model produces a directional scenario in azimuth for the mobile. The directionality depends on the delay time and is addressed further below.

6.8.1 Parameters of model

The parameters of Figure 6.8.1, are summarised in Table 6.1.

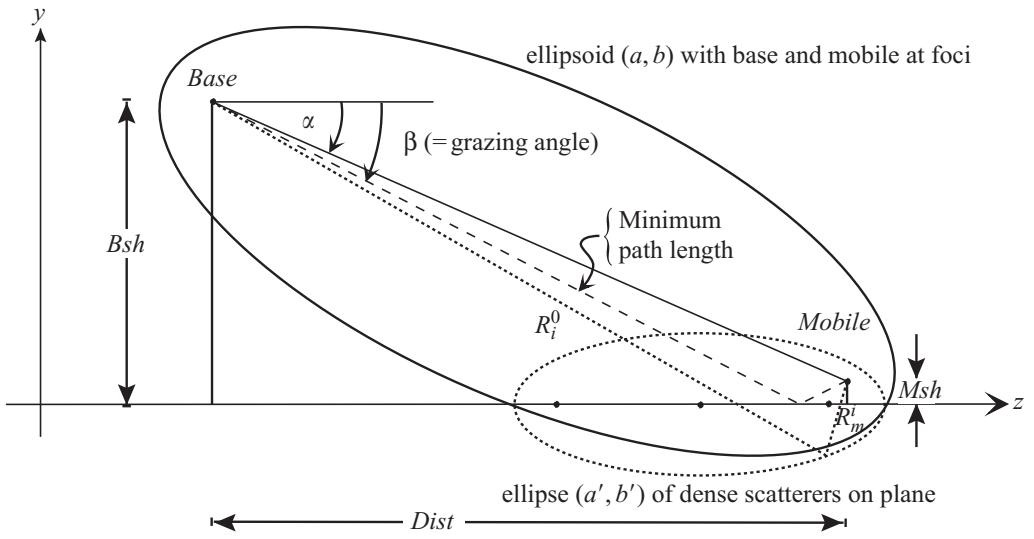


Figure 6.8.1 The base station and mobile with ellipsoid and ellipse of scatterers on the ground plane for a given excess delay time. At a later delay time the ellipsoid and ellipse are larger

Table 6.1 Summary of parameters in model

Parameter	Description
Bsh	base station antenna height
Msh	mobile antenna height
Dist	horizontal spacing between base and mobile
α	angle between base antenna and mobile antenna relative to horizontal (tilt angle of ellipsoid)
β	grazing angle of reflection
a	major axis parameter of ellipsoid
$R_{\mathrm{i}}^{\mathrm{o}}$	distance from base antenna to scatterer
$R_{\mathrm{m}}^{\mathrm{i}}$	distance from scatterer to mobile
b	minor axis parameter of ellipsoid
a'	major axis of ellipse of ground level scatterers
b'	minor axis of ellipse of ground level scatterers
y_{c}	height from ground plane of centre of ellipsoid

The elliptical relations are as follows. The grazing angle is given by

$$\beta = \tan^{-1}\left(\frac{Bsh + Msh}{Dist}\right) \tag{6.8.1}$$

and this is for the case of a single reflection from the ground plane between the mobile and base. The corresponding path length of the grazing ray is

$$R_{\min} = \frac{Bsh + Msh}{\sin\beta} \tag{6.8.2}$$

which has a delay of $\tau_{\min} = R_{\min}/c$, where c is the speed of propagation. In this diffuse scattering model, there is no line of sight, so this ray represents the earliest time of arrival. So for this situation, the ellipse of scatterers is vanishingly small and the excess delay is zero. At increasing delay times, the ellipse increases in size until the ellipse becomes essentially circular.

The ellipsoid primary parameters are (a, b). We have

$$a = R/2, \tag{6.8.3}$$

where $R = R_{\min} + R_{\mathrm{ex}}$ is the path length between the base and mobile antenna via the single scatterer, and R_{ex} is the distance corresponding to the excess delay. In fact $R = R_{\mathrm{i}}^{0} + R_{\mathrm{m}}^{\mathrm{i}}$ as depicted in Figure 6.8.1. We also have

$$b = \frac{\sqrt{R^2 - \left(Dist^2 + (Bsh - Msh)^2\right)}}{2}. \tag{6.8.4}$$

The tilt angle of the ellipsoid is

$$\alpha = \tan^{-1}\left(\frac{Bsh - Msh}{Dist}\right). \tag{6.8.5}$$

The parameters of the ellipse of scatterers are:

$$a' = \sqrt{\frac{1 - g_{ab}^2/c_{ab} - y_{ab}^2}{c_{ab}}}; \tag{6.8.6}$$

and

$$b' = b\sqrt{1 - \frac{g_{ab}^2}{c_{ab}} - y_{ab}^2}; \tag{6.8.7}$$

in which

$$g_{ab} = \left(\frac{1}{a^2} - \frac{1}{b^2}\right) y_c \sin\alpha \cos\alpha, \tag{6.8.8}$$

$$c_{ab} = \left(\frac{\sin\alpha}{a}\right)^2 + \left(\frac{\cos\alpha}{b}\right)^2, \tag{6.8.9}$$

$$y_{ab}^2 = c_{ab} y_c^2, \tag{6.8.10}$$

and

$$y_c = Msh + \sqrt{a^2 - b^2}\sin\alpha \tag{6.8.11}$$

is the height of the centre of the ellipse above the ground plane.

6.8.2 Directionality in the simple ellipse model

The basic elliptic model is given in Figure 6.8.2, with a and b the usual elliptical parameters. The left and right foci have 'complementary' angle definitions.

The scatterers are taken as uniformly distributed on the plane, which means a uniform distribution on the ellipse. The length of the arc is therefore proportional to the number of dense scatterers, or incident power at the foci, in a given angle. The transmitter can be considered at the left focus transmitting to the mobile at the right focus. The following development is a somewhat different approach to that in Nørklit *et al*. (1998).

For the origin at the centre of the ellipse (as drawn in Figure 6.8.2), the basic ellipse has a polar equation

$$r(\theta) = \frac{ab}{\sqrt{a^2\sin^2\theta + b^2\cos^2\theta}} \tag{6.8.12}$$

and the angles for the left and right foci are related to the centre angle θ, for $-\pi \leq \theta \leq \pi$, as

$$\theta_L(\theta) = \tan^{-1}\left(\frac{r\sin\theta}{r\cos\theta + c}\right), \quad \theta_R(\theta) = \tan^{-1}\left(\frac{r\sin\theta}{r\cos\theta - c}\right). \tag{6.8.13}$$

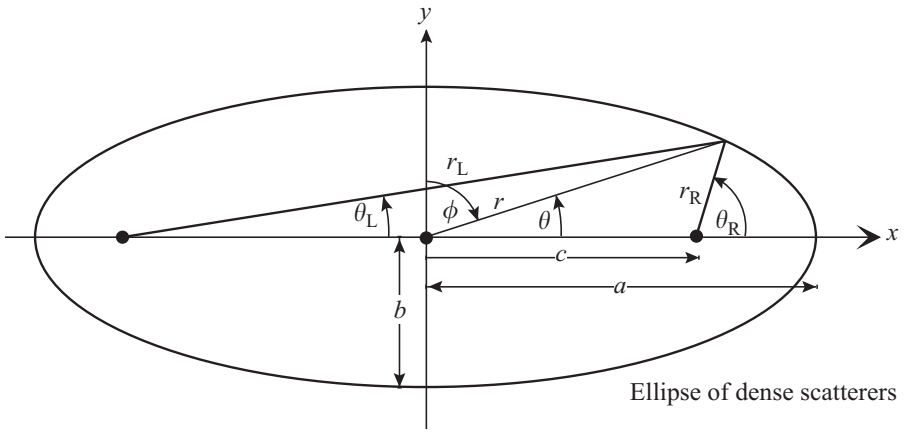


Figure 6.8.2 The basic elliptical model with transmitter and receiver at the foci of an ellipse which contains the scatterers for a given delay time

The elemental arc length is found from the usual formula $s^2 = (\partial x/\partial\theta)^2 + (\partial y/\partial\theta)^2$ to be

$$s(\theta) = a\sqrt{1 - \epsilon^2 \sin^2(\pi/2 - \theta)} \tag{6.8.14}$$

where

$$\epsilon = \frac{\sqrt{a^2 - b^2}}{a} \tag{6.8.15}$$

is the eccentricity of the ellipse. Setting $\phi = \pi/2 - \theta$ (see Figure 6.8.2), the arc length for a given angular range $[\phi_1, \phi_2]$ is given by the incomplete elliptical integral as $s(\phi_1, \phi_2) = E\left(\epsilon^2, \phi_2\right) - E\left(\epsilon^2, \phi_1\right)$. The pdf for the arc length follows as

$$p_s(\theta) = \frac{a\sqrt{1 - \epsilon^2 \sin(\pi/2 - \theta)}}{\int_0^{2\pi} a\sqrt{1 - \epsilon^2 \sin(\pi/2 - \theta)}\,\mathrm{d}\theta} = \frac{\sqrt{1 - \epsilon^2 \sin(\pi/2 - \theta)}}{4E(\epsilon^2)}. \tag{6.8.16}$$

The result for the origin at the left and right foci can be found the same way. The required polar equations are, respectively,

$$r_L(\theta_L) = \frac{a(1 - \epsilon^2)}{1 - \epsilon\cos\theta_L}, \quad r_R(\theta_R) = \frac{a(1 - \epsilon^2)}{1 + \epsilon\cos\theta_R}. \tag{6.8.17}$$

Examples are given in Figure 6.8.3. The plot represents the pdfs of the power of the dense scatterers in an angle at the left and right foci. This would be the angular power pdf if there was no space loss, i.e. $\alpha = 0$. The plots for the left and right foci are the same, but have a displacement of π rad from each other. They show the changing pdf with changing excess delay time which is represented by the different ellipses. The double peak arises from the long arc length seen from the foci at the far end of the ellipse. For very large excess delays, the ellipse approaches a circle and the angular

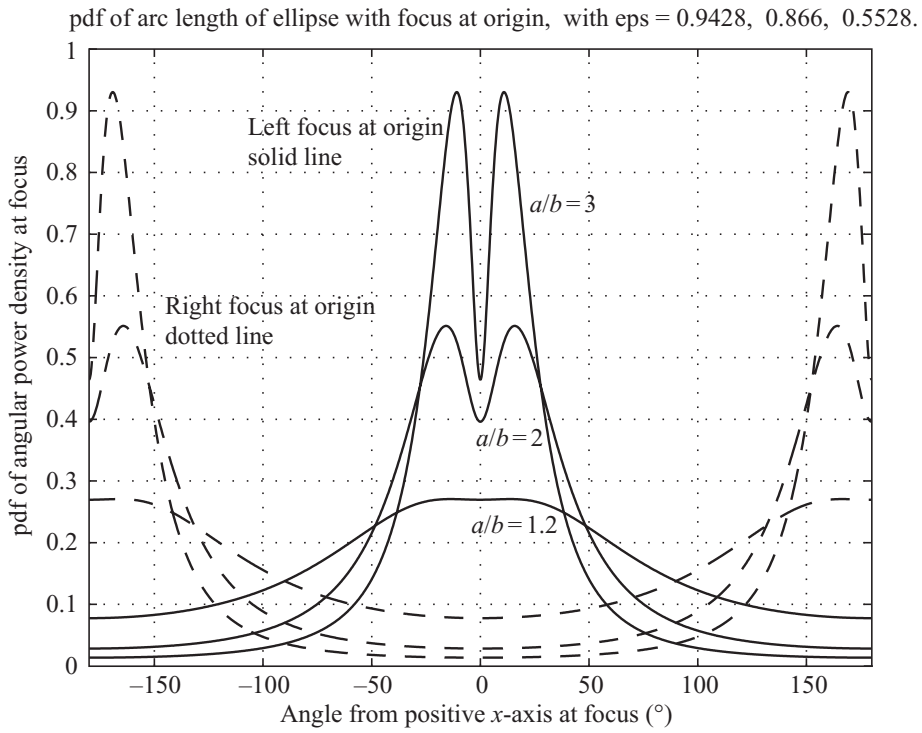


Figure 6.8.3 *The probability density function of elliptical arc length (considered as a random variable) as a function of angle at the left and right foci, for three different delay times defined through the ellipse geometry. The Doppler power profile can be found by including the azimuthal antenna pattern (if directional) and a space loss factor*

pdf approaches a uniform distribution. For the ellipse approaching a line, the pdfs approach delta-like functions at $\theta_L = 0$ and $\theta_R = \pi$.

The power delay profile, found by integrating over all ellipses (all delay times) has some interesting, tractable cases (Andersen and Eggers, 1992). A delay time, normalised to the delay corresponding to the interfocal distance, is written

$$\tilde{\tau} = \frac{\tau}{\left(\frac{2\sqrt{a^2-b^2}}{c}\right)}, \tag{6.8.18}$$

where c is again the speed of propagation. When $\tilde{\tau} = 1$ there is a direct ray between the focal points, and this case is not considered here, i.e. we deal only with $\tilde{\tau} > 1$. For the free space case, $\alpha = 2$, the power delay profile reduces to

$$P(\tilde{\tau}) = \frac{4\pi}{\sqrt{\left(\tilde{\tau}^2 - 1\right)\left(\left((2\tilde{\tau})^2 - 1\right)^2 - 1\right)^2}} \tag{6.8.19}$$

and as noted above, this produces an infinite delay spread. This does not imply a non-physical result. It simply reflects the fact that the distance to the relevant scatterers

extends to infinity, and in practice, a finite time aperture or noise enforces a truncation of the delay profile, and also ensures a finite delay spread. When $\tilde{\tau}$ is large,

$$P(\tilde{\tau}) = 2\sqrt{2}\pi\tilde{\tau}^{-3}, \quad \tilde{\tau} \gg 1,$$

and finally, again for large $\tilde{\tau}$ only, but general α,

$$P(\tilde{\tau}) \approx \tilde{\tau}^{(1-2\alpha)}, \quad \tilde{\tau} \gg 1.$$

6.9 References

AMOROSO, F.: 'Use of DS/SS signalling to mitigate Rayleigh fading in a dense scatter environment', *IEEE Personal Communications*, April 1996, **3** (2), pp. 52–61.

ANDERSEN, J. BACH, LAURITZEN, S.L., and THOMMESEN, C.: 'Distributions of phase derivatives in mobile communications', *Proceedings of the IEE*, August 1990, **137**, Pt.H, (4), pp. 197–201.

ANDERSEN, J. and EGGERS, P.C.F.: 'A heuristic model of power delay profiles in landmobile communications', *Proceedings URSI International Symposium on Electromagnetic Theory*, Sydney, Australia, August 1992, pp. 55–57.

ANDERSEN, J.BACH and PEDERSEN, K.I.: 'Angle-of-arrival statistics for low resolution antennas', *Proceedings of the 4th European Personal Mobile Communications Conference, EPMCC*, Vienna, February 2001.

AULIN, T.: 'A modified model for the fading signal at a mobile radio channel', *IEEE Transactions*, 1972, **VT-21** (1), pp. 27–38.

BELLO, P.A.: 'Characterization of randomly time-variant linear channels', *IEEE Transactions*, December 1963, **CS-11**, pp. 360–393.

CHAO, R.-U., FUJIMOTO, K., and HIRAZAWA, K.: 'Three-dimensional performance of an LMS adaptive array with inverted-F antennas', *IEEE Transactions*, August 1991, **VT-40** (3), pp. 575–583.

COLLIN, R.E. and ZUCKER, F.: *Antenna Theory*, Part I, McGraw-Hill, New York, 1969.

EGGERS, P.C.F., THOMMESEN, C., and ANDERSEN, J. BACH: 'On spatial power gradients', *IEEE Transactions*, January 2000, **VT-49** (1), pp. 235–238.

EGGERS, P.C.F., TOFTGÅRD, J., and OPREA, A.M.: 'Antenna systems for base station diversity in urban small and micro cells', *Journal of Selected Areas in Communications*, September 1993, **11** (7), pp. 1046–1057.

KALKAN, M. and CLARKE, R.H.: 'Prediction of the space-frequency correlation function for base station diversity reception', *IEEE Transactions*, February 1997, **VT-46** (1), pp. 176–184.

KOZONO, S., TSURUHATA, H., and SAKAMOTO, M.: 'Base station polarisation diversity reception for mobile radio', *IEEE Transactions*, 1984, **VT-33** (4), pp. 301–306.

LEE, W.C.-Y.: 'Effects of correlation between two mobile radio base station antennas', *IEEE Transactions*, November 1973, **COM-21**, pp. 1214–1224.

LEONARD-TAYLOR, M.: 'Delay, spread simulations of an indoor multipath environment,' Industrial Development Limited, Report 759, July 1997.

LIBERTI, J.C. and RAPPAPORT, T.S.: 'A geometrically based model for line-of-sight multipath radio channels', *Proceedings of 46th IEEE Vehicular Technology Conference*, Atlanta, Georgia, USA, May 1996, pp. 844–848.

NØRKLIT, O. and ANDERSEN, J. BACH: 'Diffuse channel model and experimental results for array antennas in mobile communications', *IEEE Transactions*, June 1998, **AP-46** (6), pp. 834–840.

PAPOULIS, A.: *Signal Analysis*, McGraw-Hill, 1977.

PARSONS, J.D.: *The Mobile Radio Channel*, Pentech, London, 1992.

PEDERSEN, K.I., MOGENSEN, P.E., and FLEURY, B.H.: 'Power azimuth spectrum in outdoor environments', *Electronic Letters*, 28 August 1987, **33**, pp. 1583–1584.

PIERCE, J.R. and STEIN, S.: 'Multiple diversity with non-independent fading', *Proceedings of the IRE*, January 1960, **48**, pp. 89–104.

RICE, S.O.: 'Mathematical analysis of random noise', *Bell System Technical Journal*, 1944, **23**, pp. 282–332.

RICE, S.O.: 'Statistical properties of a sine wave plus random noise', *Bell System Technical Journal*, 1948, **27**, pp. 109–157.

SALZ, J. and WINTERS, J.H.: 'Effect of fading correlation on adaptive arrays in digital mobile radio', *IEEE Transactions*, November 1994, **43** (4), pp. 1049–1057.

TURIN, G.L.: 'Communication through noisy, random-multipath channels', MIT Lincoln Laboratory Technical Report, No.116, May 1956.

VAUGHAN, R.G.: 'Signals in mobile communications – a review', *IEEE Transactions*, 1986, **VT-35** (4), pp. 135–145.

VAUGHAN, R.G.: 'Polarization diversity in mobile communications', *IEEE Transactions*, August 1990, **VT-39** (3), pp. 177–186.

VAUGHAN, R.G.: 'Spaced directive antennas for mobile communcations by the Fourier method', *IEEE Transactions*, July 2000, **AP-48** (7), pp. 1025–1032.

VAUGHAN, R.G.: 'Finite sample estimates for mobile channels', *Proceedings of the IEEE Vehicular Technology Society Conference* (VTC-2000), Boston, September 2000, **2**, pp. 797–804.

VAUGHAN, R.G. and ANDERSON, J., BACH: 'Antenna diversity in mobile communications, *IEEE Transactions*, November 1987, **VT-36** (4), pp. 149–172.

VAUGHAN, R.G. and SCOTT, N.L.: 'Antennas for FPLMTS', *Proceedings of The Fourth International Symposium on Personal, Indoor and Mobile Radio Communications* (PIMRC'93), Yokohama, Japan, September 1993, pp. 562–566.

Chapter 7

Aspects of simulation and measurement

7.1 Short-term fading simulation

In generating signals similar to those encountered in mobile communications, it is practical to compute signal samples from the scenario model by moving the receiving antenna amongst discrete sources. In simulations, calculations of any metric over a finite sample size will provide only an estimate of the ensemble average value. In this section, basic considerations for computer simulations of the short-term fading channel are discussed. Two aspects of simulation are: the generation of discrete channel samples, that may be mutually independent and have the correct ensemble-averaged statistics; and the generation of time (i.e. space-) series which are samples of continuous channel behaviour, and these may comprise closely spaced, correlated samples, which should also have the correct ensemble-averaged statistics.

7.1.1 Model for narrowband signals

Two-dimensional scenarios are the most used. In order to produce accurate multipath statistical distributions, such as Rayleigh-distributed envelopes, a minimum of about ten equiamplitude and random-phase sources should be received at any one time. The sources need not have the same amplitude for simulation purposes, the main point being that the real and imaginary parts approach Gaussian distributions. More sources than this may be required to achieve uncorrelated samples in long record lengths (see Section 7.1.4). In this section, the uniform scenario is of main interest.

The uniform, or Clarke, scenario is simulated using a ring of discrete sources with the receiver at the nominal centre of the ring. The sources can be equal amplitude, equal excess delay, and random phase. This models a uniform incident power distribution in azimuth. Two techniques can be used in simulations for the uniform scenario. The directions to the sources can be fixed, so that the moving receiver 'sees' each effective source at the same bearing as it moves; or the effective sources are spatially fixed so that as the receiver moves, the angles to the sources change. The former technique is more computationally efficient since the distances to the individual sources

do not need to be calculated. In the latter technique, the distances to the sources can be made large so that it gives essentially the same situation as fixed bearing angle sources. In other words, in such a simulation, the mobile trajectory is kept small relative to the distance to the sources. For example, for a 100 wavelength trajectory, the distance to the source positions should be several thousand wavelengths in order to restrict the angle changes to the sources to milliradians. Under such conditions, there is no need to include the free-space path-loss term in the summations, since this will not affect the short-term signal behaviour. However, incorporating some path loss factor does allow some long-term effects to be included, and also affects some statistical parameters because the path loss acts by selectively scaling the directional contributions. An example for the delay spread is given further below. The important point here is that for the modelling of the short-term nature of narrowband signals, the distances between the mobile and the sources are not important in the model. The set of directions to the sources is the critical parameter.

The uniform scenario is treated here as instantaneous, i.e. the scenario is not changing with time. However, the uniform scenario normally represents a time-averaged quantity. In practice, an instantaneous scenario (a snapshot of incoming waves at a given time or position) may have few sources with a non-uniform distribution. Taking the instantaneous scenario as the same as the time-averaged model imposes limitations on the validity of the simulated signals. Nevertheless, many statistical quantities will be correctly represented if the instantaneous scenario is somehow changed a sufficient number of times to represent the ensemble average during the simulation. The basic modelling step is shown in Figure 7.1.1. The directive scattering patterns, here shown for the mobile receiving case, are drawn to indicate that the scattering is independent of the mobile position.

This approach implies that the mobile itself is not a significant contributor to the scattering, and this is a limitation of the model. In the case of vehicular communications, there is some evidence that a car in an urban environment acts to perturb the scenario. This was quantified by measuring the correlation of

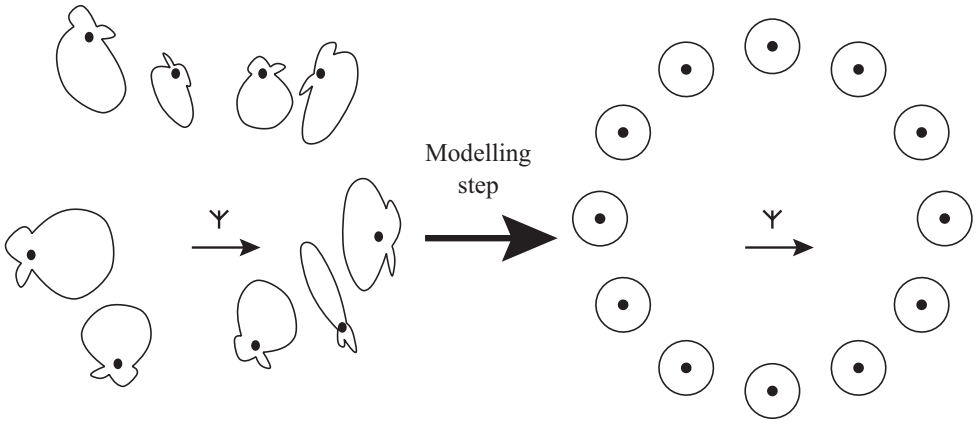


Figure 7.1.1 A receiver moving through directive scatterers (here depicted as constant with receiver position) is modelled for simulation by a ring of sources at a large radius

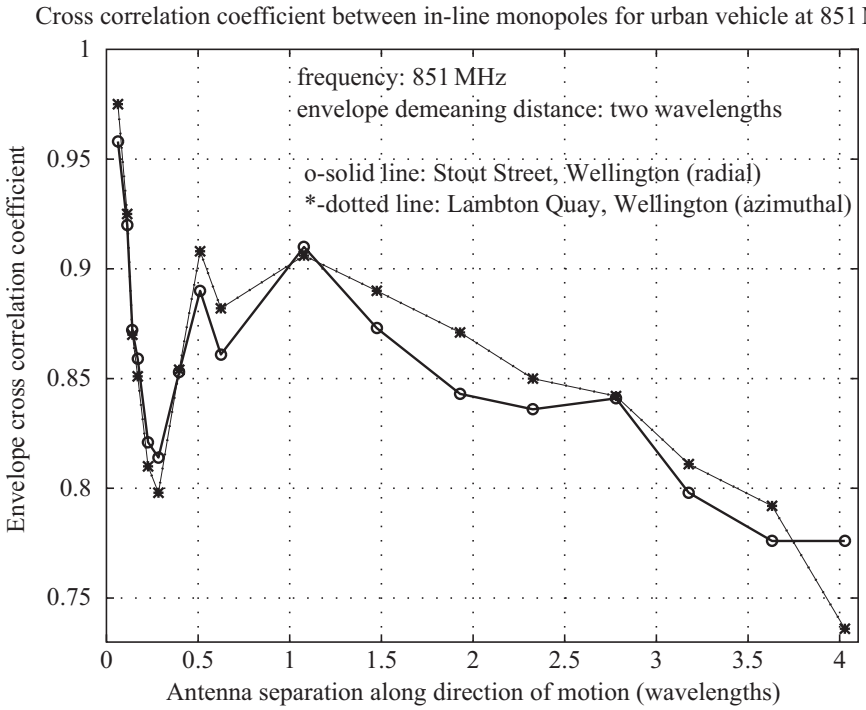


Figure 7.1.2 *The peak value of the cross-correlation coefficient of in-line monopoles mounted on a large groundplane on a vehicle in an urban environment at 851 MHz. The decreasing correlation at larger distances shows that the mobile itself may influence the scenario of sources seen by the antennas, and this is a limitation of the basic modelling step*

narrowband envelopes received by well-separated, in-line monopoles on a large roof-mounted groundplane which makes the patterns 'seen' by each element as similar as possible. The results are shown in Figure 7.1.2 (Vaughan *et al.* 1992). The correlation coefficient for the Rayleigh-fading envelopes is shown for two measured scenarios, one in a basically radial trajectory with respect to the base station, and the other along a basically azimuthal trajectory. For very close spacing, there is very high mutual coupling as the monopoles almost merge, and the correlation coefficient approaches unity. At moderately close spacing, the mutual coupling acts to decouple the antennas (see Section 8.4), and at the larger spacings where the mutual coupling dies away, the correlation coefficient decreases steadily. From the modelling approach above, the correlation coefficient should remain at unity, with the traces of the two monopoles, with similar as possible patterns, cutting an identical locus through the standing wave environment. The steadily decreasing nature of the experimental curve shows that the vehicle itself is changing the standing wave field, or the antenna and the vehicle form one antenna, which changes phase and amplitude pattern depending on the position of the antenna on the groundplane.

The Clarke (1966) modelling step is to postulate the time average of the instantaneous scenarios as a ring of sources at a large radius and then use this as an instantaneous, static scenario.

For a 2D scalar case, the sources lie in the xz plane at angle θ from the z-axis and the receiver moves along the z-axis. The 2D receiving pattern is denoted

$$\begin{aligned} g(\theta) &= |g(\theta)|\mathrm{e}^{\mathrm{j}\xi}, \quad \text{complex amplitude pattern,} \\ G(\theta) &= |g(\theta)|^2, \quad \text{power pattern (gain).} \end{aligned} \tag{7.1.1}$$

This pattern can be factored to include a multiplicative directional function which can be used to include a directionality in the scenario. So $G(\theta)$ can be considered as the product of the antenna directivity and any directivity in the scenario.

The signal at the receiver is the summation over the N sources and has the complex envelope

$$r(z)\mathrm{e}^{\mathrm{j}\psi(z)} = \sum_{i=1}^{N} g(\theta)a_i \mathrm{e}^{\mathrm{j}\alpha_i}\mathrm{e}^{\mathrm{j}k_\mathrm{C}\cos\theta_i z}\mathrm{e}^{\mathrm{j}\omega\tau_i}, \tag{7.1.2}$$

where ω is the baseband frequency. Recall that the relationship between the actual radio frequency, ω_RF, the baseband frequency and the radio carrier frequency, ω_c, is

$$\omega_\mathrm{RF} = \omega + \omega_\mathrm{c}. \tag{7.1.3}$$

In this section, ω_D is the Doppler angular frequency and this is related to the spatial Doppler frequency, $u = k_\mathrm{C}\cos\theta$, by the speed $V = z/t$, by

$$\omega_\mathrm{D} = Vu = (V/c)\omega_\mathrm{c}\cos\theta, \tag{7.1.4}$$

where c is the speed of propagation and the maximum value for the Doppler frequency is

$$\omega_\mathrm{Dm} = Vu_\mathrm{max} = Vk_\mathrm{C} = (V/c)\omega_\mathrm{c}. \tag{7.1.5}$$

The terminology is revisited below. The important point is that the baseband frequency ω is a separate variable from the Doppler frequency ω_D.

The excess delays, τ_i, are set to zero here in order to concentrate on the narrowband signal behaviour. The delay is reintroduced later when wideband signal behaviour is addressed (see Section 7.1.5). The received signal is now expressed

$$r(z)\mathrm{e}^{\mathrm{j}\psi(z)} = x(z) + \mathrm{j}y(z) = \sum_{i=1}^{N} g(\theta)a_i \mathrm{e}^{\mathrm{j}\alpha_i}\mathrm{e}^{\mathrm{j}u_i z}, \tag{7.1.6}$$

where x and y are the in-phase and quadrature components, and there is no space loss, since only the incident angular power is currently the parameter of interest. The a_i can be all the same for Rayleigh-signal behaviour, and the α_i are uniformly distributed.

If the mobile velocity is along some direction $\theta = \gamma$, then the equations modify by replacing θ_i with $(\theta_i - \gamma)$.

The spatial variable, z, can be interchanged with time, t, using the velocity of the mobile, as noted above. The situation is shown in Figure 7.1.3, where N sources, here equispaced in angle and with equal amplitude and random phases, α_i, simulate the uniform scenario. For short-term signal behaviour, the u_i can be fixed which means

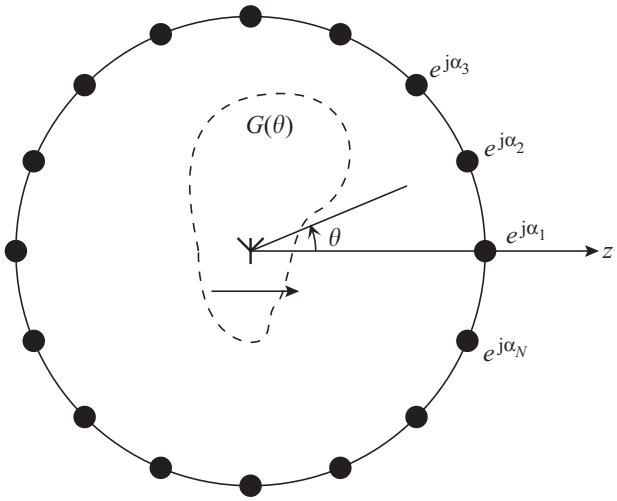


Figure 7.1.3 The uniform scenario is simulated using N equispaced and equi-amplitude discrete sources. G(θ) is the directive power gain

that the direction to the scatterers is independent of the mobile position, i.e. the u_i do not need to be calculated for each sample of the mobile position. This can be interpreted as the sources being a long way from the mobile, relative to the length of its trajectory. Recall from above that the number of effective sources being summed, i.e. the number of sources in the main beam of the pattern, should always be more than about ten in order to maintain accurate Gaussian statistics.

7.1.2 A simple Matlab simulation

It is straightforward to simulate a channel based on the equations above like

$$V_i = \sum_{k=1}^{N} A_k e^{j2\pi \cos\theta_k z} \tag{7.1.7}$$

where N is the number of random sources of complex magnitude A_k and direction θ_k, and θ_k is uniformly distributed between 0 and π. V_i is the ith realisation of the random environment. In the program below, z is a constant, but it could also be a variable for a fixed (but random) set of sources.

```
%A simulation example
N=10; %number of sources;
M=1000; %number of samples;
z=1;
for i=1:M;
 A=randn(1,N)+j*randn(1,N); %N complex Gaussian sources;
 theta=pi*rand(1,N); %uniform random directions to sources between 0 and pi;
 V(i)=A*exp(-j*cos(theta)*2*pi*z)';%summation over all sources;
end
Vmean=sqrt(mean(abs(V).^2)); %rms value;
PdB=20*log10(abs(V)/Vmean); %normalised power values in dB;
```

```
[n1,x1]=hist(PdB,50); %creating a histogram;
n1m=n1/M;
a=cumsum(n1m); %cumulative summation
b=log10(a); %on a log scale;
figure(1)
plot(x1,b,'*'),grid %plotting simulated values on a log cumulative plot;
hold on
x2=-40:10; %dB values used for exact Rayleigh plot;
r2=exp(x2/10*log(10)); %corresponding envelope values;
b2=log10(1-exp(-r2)); %cumulative Rayleigh plot on a log scale;
plot(x2,b2,'g','LineWidth',1.5);
axis([-40 10 -4 0]);
xlabel('P dB');
ylabel('log Power<P dB');
```

The program calculates a distribution of samples which are plotted on a log scale. An example is shown in Figure 7.1.4 for $M = 1000$ samples and $N = 10$ sources.

It should be noted that to have a good fit to the distribution at small probabilities ϵ, it is necessary to have the number of samples larger than $10/\epsilon$ (in Figure 7.1.4 $M = 1000$ and the fit is good down to 10^{-2}).

For this specific case of only finding a Rayleigh distributed variable it was not necessary to define the angles θ. We could also have chosen the random variables outside the loop and stepping through values of z, as in the following modification,

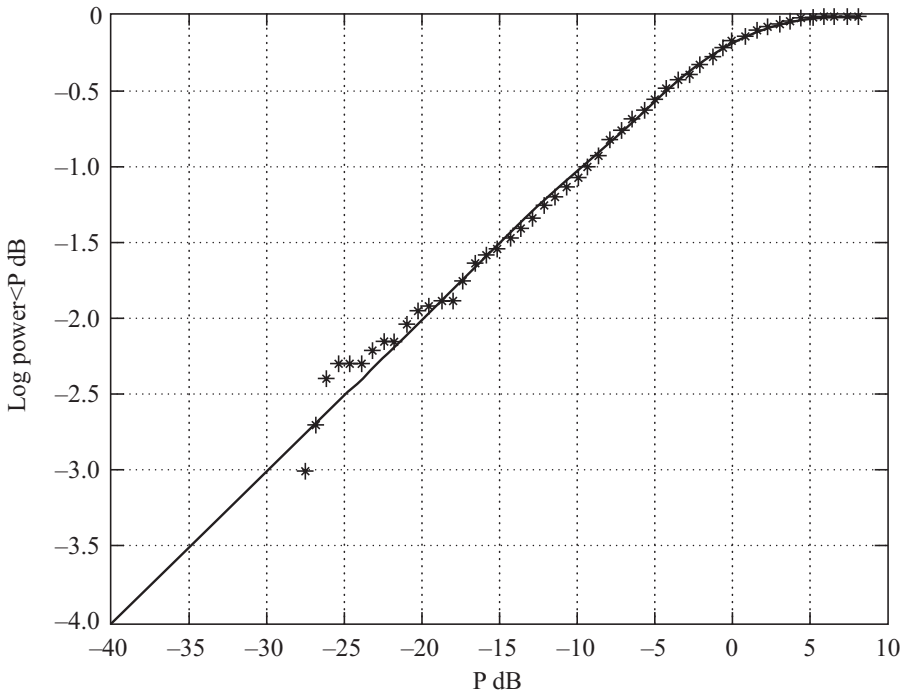


Figure 7.1.4 A distribution of random samples from a simple program compared with the exact Rayleigh distribution. The number of sources is ten, and the number of realisations is 1000, which explains the lack of samples for log(probability) less than −3

where the distribution of angles matters. The choice of random angles avoids the cyclostationarity to be described in the following;

$$V_i = \sum_{k=1}^{N} A_k e^{j2\pi \cos\theta_k z_i} \tag{7.1.8}$$

i.e.

```
A=randn(1,N)+j*randn(1,N); % N complex Gaussian sources;
theta=pi*rand(1,N); %uniform random directions to sources between 0 and pi;
for i=1:M;
 V(i)=A*exp(-j*cos(theta)*2*pi*z(i))';%summation over all sources and varying z;
end
```

An example is shown in Figure 7.1.5 with the fading power versus distance measured in wavelengths. With this approach, however, the statistics of the time series is defined as a *k*th snapshot of the sources, which is not likely to be uniformly distributed. So for a simulation producing such a time series, the appropriate scenario distribution (for example, uniform) is required.

7.1.3 Discretisation options for the uniform scenario

Using discrete sources in a computer simulation brings in the need for discretised distribution functions. It is often convenient to deal with constant amplitude samples, which are spaced variably in order to represent the variable value of the analogue

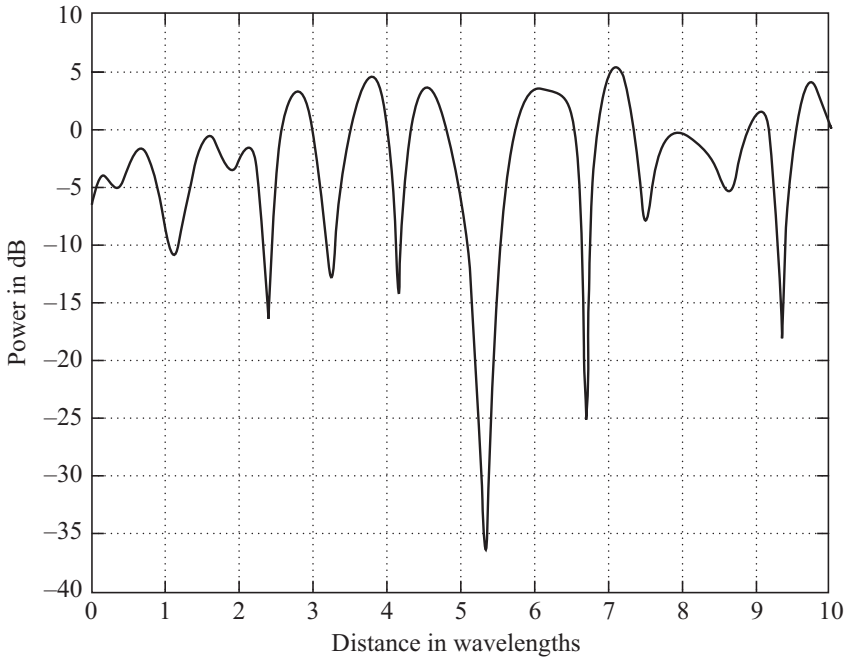


Figure 7.1.5 Example of narrowband fading using the matlab program

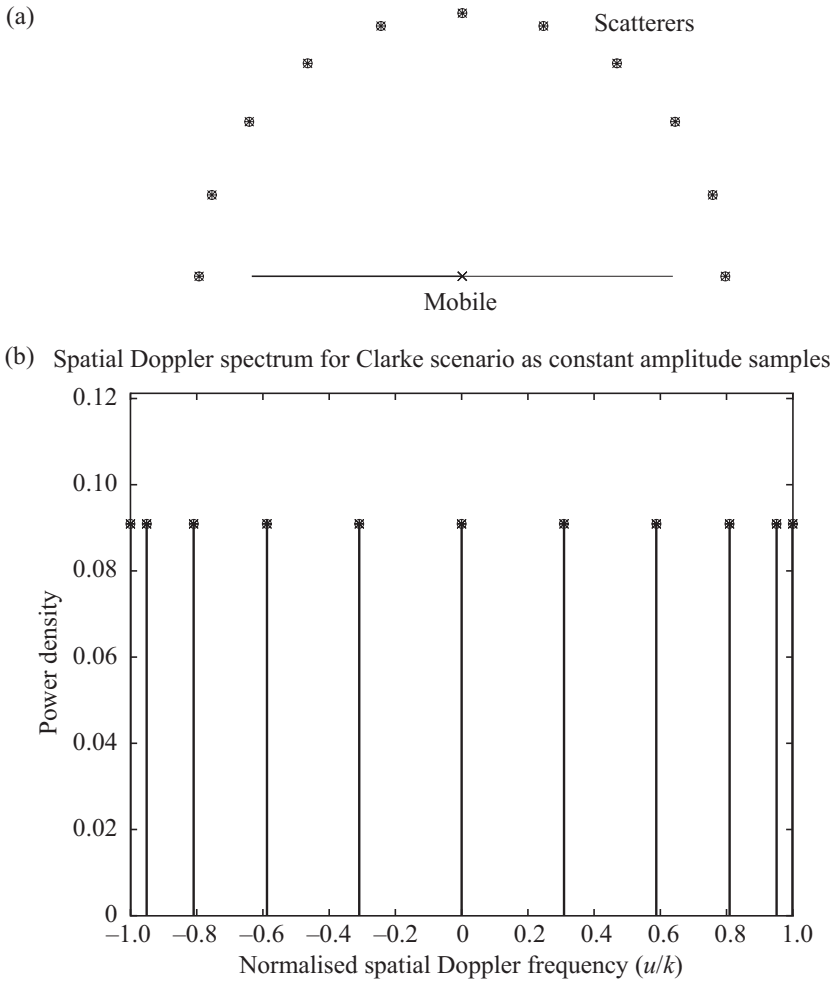


Figure 7.1.6 *(a) An example of a uniform scenario with 11 evenly spaced (in θ) sources, with (b) the Doppler spectrum as constant amplitude contributions with variable spacing*

function. This is in contrast to conventional sampling where regularly spaced samples take the value of the analogue distribution function (i.e. the sample values correspond to the sources) at the sample point.

An example is the Clarke scenario and its Doppler spectrum. For simulations, the sources are discrete and so the spectrum is also discrete. Figure 7.1.6 shows the case of $N = 11$ sources, spaced evenly on the semicircle $0 \leq \theta \leq \pi$ (the Doppler shifts are the same for corresponding sources at $-\pi \leq \theta \leq 0$), and the Doppler spectrum.

For simulations using an antenna pattern which is not symmetric, i.e. the same in the semicircles ($0 \leq \theta \leq \pi$) and ($-\pi \leq \theta \leq 0$), the scenario must of course occupy the full circle to cater for the whole pattern.

Finding the spacing of the discrete sources in the Doppler spectrum, as in the figure, requires the cdf of the Doppler spectrum. In terms of an algorithm for finding the spacing, it is convenient to begin by placing sources at the end points of the Doppler distribution ($u = \pm k_C$). The nth sample positions for the N sources are then

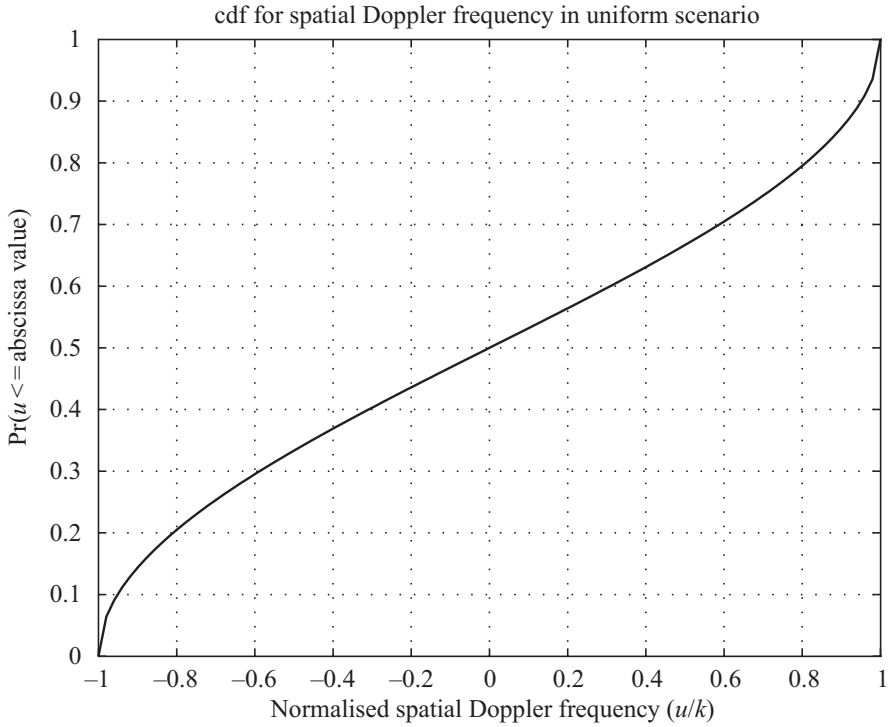


Figure 7.1.7 The cdf for the Doppler frequency for an omni antenna in a uniform scenario

found from

$$P(u \leq u_n) = \frac{n}{N}. \tag{7.1.9}$$

The cdf for the uniform scenario is

$$\begin{aligned} P(u \leq u_n) = \int_{-k_C}^{u_n} p_u(u)\,\mathrm{d}u &= \int_{-k_C}^{u_n} \frac{1}{\pi\sqrt{k_C^2 - u^2}}\mathrm{d}u \\ &= \frac{1}{2} + \frac{1}{\pi}\sin^{-1}\left(\frac{u_n}{k_C}\right) \end{aligned} \tag{7.1.10}$$

which is plotted in Figure 7.1.7, and equating the two results,

$$u_n = \frac{k_C}{\pi}\sin^{-1}\left(\frac{n}{N} - \frac{1}{2}\right). \tag{7.1.11}$$

7.1.4 Signal repetition distance (cyclostationarity from simulations)

The received signal comprises a summation from the discrete sources, i.e. $r(z) \sim \sum \mathrm{e}^{\mathrm{j}u_i z}$. (Recall that the distance to the scatterers is large relative to the trajectory of the mobile and so the directions to the sources are effectively constant as the receiver moves.) As the receiver moves, the different spatial Doppler frequencies will form 'spatial beat frequencies', which have 'beat distances'. This means that the

received signal will be repetitive in space, i.e. it is cyclostationary. (Note that this is unlikely to occur in a real-world situation.) The repetition distance can be large, however, for many scatterers spaced at appropriate angles. The important point is that a given data record may contain repeated, or almost-repeated, signal sequences, and therefore not contain its expected number (the whole record length) of 'independent'[1] data samples.

The beat Doppler frequency is given by the largest common divisor for the component frequencies, and this frequency gives the beat distance. For integer-related Doppler frequencies, e.g. for sources equally spaced in u, then the beat frequency is the common frequency difference between adjacent frequencies. For example, five sources at $\boldsymbol{u} = k_{\mathrm{C}}\left[-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\right]$, i.e. at $\theta = [180°, 120°, 90°, 60°, 0°]$, gives a true beat distance of $1/u_{\mathrm{B}} = 2$ wavelengths. For N sources, spaced evenly in u in this way, we get a beat distance of $(N-1)/2$ wavelengths. An autocorrelation function of the received signal would show a repetition (autocorrelation going to unity) at this spacing. So for statistical analysis of the signals from such simulations, the record lengths are limited by the number of sources.

By taking a set of quasi-random directions (in the sense that choices with short correlation distances should be avoided) the beat distances can be made large. However, in allowing uneven directions for the discrete sources, care must be taken to keep control of the incident power density, which should be effectively uniform; cf. Figure 7.1.6(b). The regular angular spacing case, which is the discretised uniform power density, is now treated.

The constant angular (θ) spacing serves to make the spacings within the set of Doppler frequencies (u) uneven and this creates a much longer beat distance than the case of linear, constant-u spacing. In fact, the true beat distances can become very long, but 'pseudo-beats' occur, when the autocorrelation coefficient function reaches a large peak which is not quite one. This is the distance at which many, but not all, of the received sources' signals essentially return to their original phase values.

To calculate the beat distance for a general set of Doppler frequencies, $\boldsymbol{u} = [u_1, u_2, \ldots, u_N]$, the greatest common divisor (gcd) is used, denoted

$$u_{\mathrm{B}} = \gcd\left[u_1, u_2, \ldots, u_N\right]. \tag{7.1.12}$$

This can be calculated iteratively using an algorithm such as (seeking $u_{\mathrm{B}} = u_{\mathrm{B}}^{(N)}$)

$$\begin{aligned} u_{\mathrm{B}}^{(1)} &= \gcd\left[u_1, u_2\right], \\ \text{then for } i &= 2 \text{ to } N: \\ u_{\mathrm{B}}^{(i)} &= \gcd\left[u_{\mathrm{B}}^{(i-1)}, u_i\right] \\ \text{then } u_{\mathrm{B}} &= u_{\mathrm{B}}^{(N)}. \end{aligned} \tag{7.1.13}$$

[1] In a simulation using a given set of effective sources, all the samples of a data record are generated from the same sources. This means that the samples in the record are difficult to justify as truly independent in a statistical context, although a zero correlation between all samples implies independence for Gaussian signals. Using an independent set of sources (generated from a different seed in the computer) will give a record that is statistically independent of the original record.

In treating the $\boldsymbol{u}$ as integers for this calculation, it is necessary to scale up the original floating point $\boldsymbol{u}$ and then use rounding. It is this process (rather than the gcd operation) that limits the accuracy of the beat frequency calculation by effectively limiting the accuracy of phase values of the frequency components. For example, a scale factor α is used which gives $\alpha\boldsymbol{u}$ in degrees m^{-1}. The $\alpha\boldsymbol{u}$ now is rounded leaving the phase accuracy for each frequency component as say $\pm 0.5°$. Then for $\boldsymbol{u}$ in rad m^{-1}, the accuracy-limiting process can be summarised by

$$\frac{u_{\mathrm{B}}}{\alpha} = \gcd\left[\text{round}\,(\alpha u_1),\ \text{round}\,(\alpha u_2)\right]. \tag{7.1.14}$$

For example, $\alpha = 180/\pi$ gives $f_1 = \text{round}(\alpha u_1)$ in degrees m^{-1}. For different accuracies (values of α), the beat distances change. These changes can be abrupt since only one of several deviations (from the several sources) need violate the phase alignment condition which gives the period of cyclostationarity. A measure for the phase alignment which has less abrupt behaviour, is a standard deviation of the phase,

$$\sigma_\phi(\Delta z) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\phi_i(z) - \phi_i(z+\Delta z))^2}, \tag{7.1.15}$$

where ϕ_i is the phase of the signal from the ith source. When this standard deviation approaches zero, the pseudo beat distance approaches a true beat distance of Δz.

The pseudo beat distances are evident in the correlation coefficient function which has well-defined peaks spaced quasi-randomly. Figure 7.1.8 gives the numerical results for the maximum correlation coefficient (i.e. the largest of the several peaks which occur) against the number of evenly spaced, random phase, equal amplitude sources. The results are from a simulation with the sources at a radius of $(10)^6$ wavelengths (the u_i change slightly with the mobile position), a record length of 8000 wavelengths and a maximum autocorrelation displacement distance of 500 wavelengths, centred on the centre of the semicircle of sources. The plotted correlation

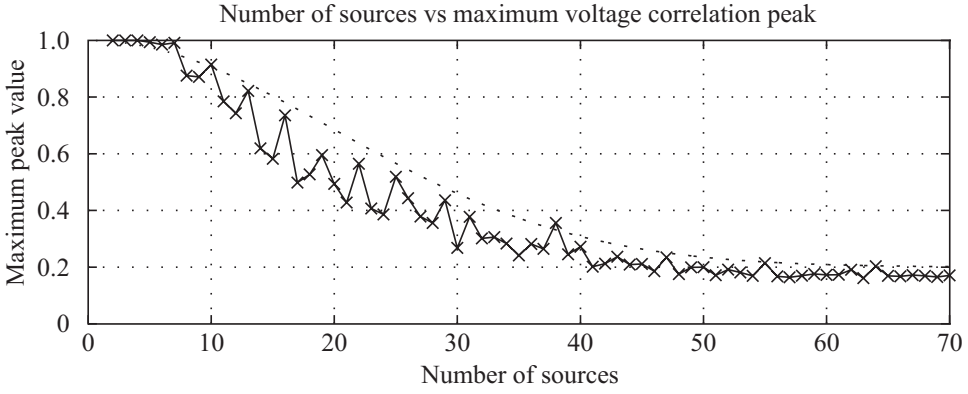


Figure 7.1.8 The maximum correlation coefficient peak within a range of 500λ for the signal from an omni antenna in a uniform scenario simulated using many sources spaced evenly on a semicircle. The envelope correlation coefficient is approximately the square root of this

maximum is the largest only from the range of displacements. (This is not the same as the first correlation (non-maximum) peak, i.e. that with the smallest pseudo beat distance.) The correlation function used here is the magnitude square of the complex voltage correlation coefficient. The results in the plot are insensitive to the random phase seed. The power signal (envelope squared) correlation coefficient has a slightly higher sensitivity to the random phase seed used for the phases of the sources, and has slightly larger variations.

The plot can be used for estimating the minimum number of sources required for simulating a uniform, or Clarke, scenario such that the data record contains 'uncorrelated' (the actual correlation value given by the value of the correlation coefficient peak) data which can be considered as uncorrelated in the sense that it is non-repetitive over the length of the data record. For example, for data records which maintain their autocorrelation coefficient between well-displaced samples below 0.3 (envelope correlation coefficient $\rho_e <\sim 0.1$), the plot indicates that about 50 sources are required for a simulation. A simple, empirical relationship for approximately relating the largest values of the maximum correlation, $|\rho|^2_{\max}$, and the number of sources, N, is

$$|\rho|^2_{\max} = 0.15 + (1 - 0.15)\mathrm{e}^{-N^2/(2(20)^2)}, \quad \text{i.e.} \quad N = 20\sqrt{2\ln\left(\frac{1-0.15}{|\rho|^2_{\max} - 0.15}\right)} \tag{7.1.16}$$

and this can be used to find N for a given $|\rho|^2_{\max}$ or vice versa.

7.1.5 Delay spread example with path-loss effect

An example of simulation using a spatially fixed set of sources is to study the effect of path loss on the delay spread as the receiver moves all the way through a Clarke scenario. The simulation has a single ring of N dense, random phase sources, with the receiver starting at the centre and moving to a long way outside the ring. The excess delay of each of the sources is the same, so at the centre of the ring, the delay spread is zero and there is flat fading, i.e. the signal features narrowband behaviour. The received voltage from an omnidirectional antenna is

$$s(z) = \sum_{i=1}^{N} s_i(z) = \sum_{i=1}^{N} \mathrm{e}^{\mathrm{j}\phi_i} \times d_i^{-2\gamma}\, \mathrm{e}^{\mathrm{j}u_i z} \tag{7.1.17}$$

or the position-dependent impulse response can be written

$$h(\tau, z) = \sum_{i=1}^{N} s_i(z)\delta\left(\tau - \tau_i(z)\right) \tag{7.1.18}$$

where d_i is the distance between the ith source and the receiver position, z, γ is the power path-loss factor, and the delay to the ith source is $\tau_i(z) = d_i(z)/c$ and c is the speed of propagation.

In Figure 7.1.9, the delay spread is given as a function of the radial position of the receiver. The receiver passes between the discrete sources when it reaches the radius

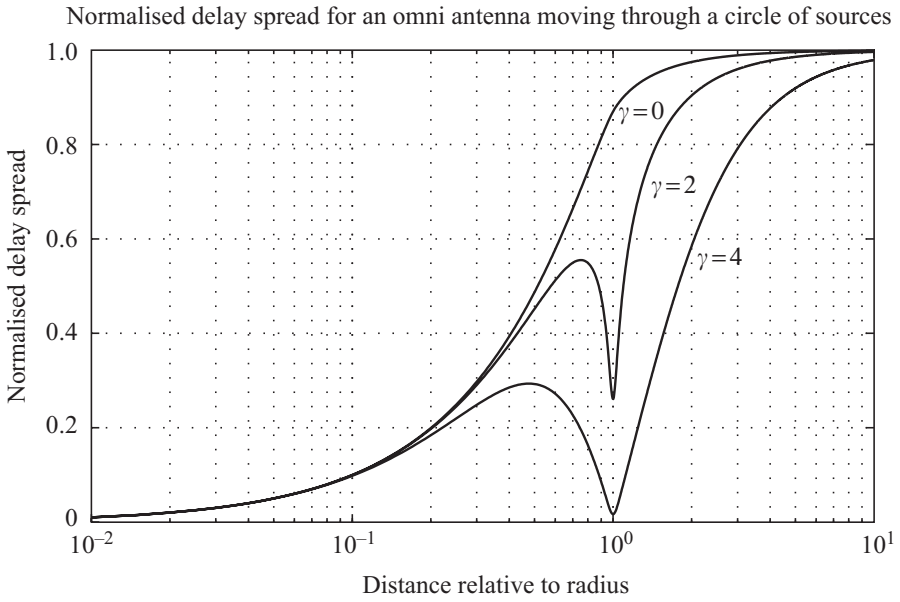


Figure 7.1.9 *The normalised delay spread for an omni antenna as a function of its radial position in a uniform scenario, with the power path-loss factor as a parameter (produced with assistance from M. Leonard-Taylor, 1997)*

of the ring of sources so that d_i is always non-zero. Independent of the path-loss factor, the delay spread is zero at the centre of the ring. As the receiver moves well away from the ring of sources it receives essentially plane waves, i.e. the Doppler spread goes to zero. Here, the delay spread is

$$s^2(z \to \infty) = \left(\frac{\int \tau^2 |h(\tau, \infty)|^2 \, \mathrm{d}\tau}{\int |h(\tau, \infty)|^2 \, \mathrm{d}\tau} - \left(\frac{\int \tau |h(\tau, \infty)|^2 \mathrm{d}\tau}{\int |h(\tau, \infty)|^2 \, \mathrm{d}\tau} \right)^2 \right)^{1/2}$$
$$= \frac{R}{\sqrt{2}c}, \tag{7.1.19}$$

where R is the radius of the ring of sources in metres. For an increasing path-loss factor the changing radial position causes the receiver to see an increasingly directional incident angular power distribution, and an increasingly deep dip in the delay spread appears at the radial position of the sources. This is because the sources which are close to the receiver have small delay times (because they are now close to the receiver), and these few sources become a more dominant influence on the received signal.

The results in this type of example are difficult to find analytically, but are easy to simulate. For specific source models or more complicated scenarios, such results can be found only from simulation (or measurement). Other examples are given in Chapter 5.

7.2 Reference parameters for simulations

In undertaking simulations, it is important to be able to benchmark the results, which use finite sample sizes, with the expected values. In this section, basic results are reviewed for the short term signal behaviour.

7.2.1 Basic pdf for incident power

The pdf for the continuous angular incident power is recalled as (Clarke, 1968)

$$p_\theta(\theta) = \frac{1}{2\pi}, \quad 0 \le \theta < 2\pi, \tag{7.2.1}$$

or, in the u-notation,

$$p_u(u) = \frac{1}{\pi\sqrt{k_C^2 - u^2}}, \qquad -k_C \le u \le k_C, \tag{7.2.2}$$

and the simulated quantity is a summation of the discrete sources, here evenly spaced in the θ_i directions,

$$p_\theta(\theta) = \sum_i p_\theta^{(i)}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \delta\left(\theta - \left(\frac{i-1}{N}\right)2\pi\right). \tag{7.2.3}$$

As long as N is large enough (or equivalently, for a given N, that the antenna pattern is not too directive) to include a sufficient number (say ten) of sources in the pattern lobes, then the continuous and discrete pdfs are effectively the same[2] This can be called dense multipath and the simulated signals will be representative of those from a continuous, uncorrelated scenario.

7.2.2 Basic received signal

The in-phase component is

$$\begin{aligned} x(z) &= \mathrm{Re}\sum_{i=1}^{N} |g(\theta_i)| e^{j\xi_i} e^{j\alpha_i} e^{ju_i z} \\ &= \sum_{i=1}^{N} |g(\theta_i)| \left(b_i \cos\alpha_i + c_i \sin\alpha_i\right), \end{aligned} \tag{7.2.4}$$

where

$$\begin{aligned} b_i &= \cos\xi_i \cos(u_i z) - \sin\xi_i \sin(u_i z) \\ c_i &= -\sin\xi_i \cos(u_i z) - \cos\xi_i \sin(u_i z) \end{aligned} \tag{7.2.5}$$

[2] For continuous distributions, $\int p_\theta(\theta)\, d\theta = 1$ and for discrete distributions, $\sum_i p_\theta^{(i)}(\theta_i) = 1$.

are functions of the distance (or time) and include the phase of the receiving pattern, ξ_i, in the direction $\theta = \theta_i$. (Recall that $u_i = k_C \cos\theta_i$.) The quadrature component, $y(z)$, is found the same way.

7.2.3 Averaging

To find the averaged value of x and y, the averaging is firstly over distance (or time) for each direction, which is

$$\langle b_i \rangle_z = \langle c_i \rangle_z = 0 \tag{7.2.6}$$

from the result that the received signal has random, uniformly distributed phase over distance z (or time), i.e.

$$\langle \cos(u_i z) \rangle_z - \langle \sin(u_i z) \rangle_z = 0. \tag{7.2.7}$$

In a simulation, the distance z must be averaged for long enough to invoke an estimate with the required accuracy.

The averaging of x and y is secondly over the ensemble of directions, θ, i.e. over i, which for the uncorrelated scenario (random phases of the N sources) results in

$$\langle \cos\alpha_i \rangle_i = \langle \sin\alpha_i \rangle_i = 0. \tag{7.2.8}$$

Again, in a simulation or measurement, there should be sufficient samples (here, sources) in order to invoke a sufficiently accurate estimate of this result.

Applying the averaging in both distance or time (z) and direction to the signal (index i for the direction $u_i = k_C \cos\theta_i$) takes the form

$$\begin{aligned} E\{x\} &= \sum_{i=1}^{N} \langle |g(\theta_i)| b_i \rangle_i \langle \cos\alpha_i \rangle_z + \langle |g(\theta_i)| c_i \rangle_i \langle \sin\alpha_i \rangle_z \\ &= 0 \end{aligned} \tag{7.2.9}$$

and so the zero mean result stems from either the distance averaging or the directional averaging. The same result applies to the quadrature component.

The condition for random phase of the uncorrelated sources at directions indexed by *i,l* can be written

$$\langle \cos\alpha_i \sin\alpha_l \rangle_{i,l} = \langle \cos\alpha_i \sin\alpha_l \rangle_{i,l} = 0, \quad i, l \in N, \tag{7.2.10}$$

and this result is required for finding the second-order moments: $E\{x^2\}$, $E\{y^2\}$; and

$$E\{xy\} = 0. \tag{7.2.11}$$

The latter result confirms that the received signal is a complex Gaussian form, independent of the gain or shape of the receiving pattern. However, the other second-order

moments, being energy terms, depend on the pattern (and any scenario directionality), $G(\theta)$, namely

$$\begin{aligned}\langle x^2\rangle_z &= \left\langle \left(\mathrm{Re}\sum_{i=1}^{N}|g(\theta_i)|\mathrm{e}^{\mathrm{j}\alpha_i}\mathrm{e}^{\mathrm{j}u_i z}\right)\left(\mathrm{Re}\sum_{l=1}^{N}|g(\theta_l)|\mathrm{e}^{\mathrm{j}\alpha_l}\mathrm{e}^{\mathrm{j}u_l z}\right)\right\rangle_z \\ &= \frac{1}{N}\sum_{i=1}^{N}|g(\theta_i)|^2\left(\left\langle b_i^2\right\rangle_z\cos^2\alpha_i + \left\langle c_i^2\right\rangle_z\sin^2\alpha_i\right) \\ &= \frac{1}{2N}\sum_{i=1}^{N}|g(\theta_i)|^2 \\ &= \frac{1}{2}\langle G(\theta_i)\rangle_i . \end{aligned} \tag{7.2.12}$$

In this step, the values

$$\langle b_i^2\rangle_z = \langle c_i^2\rangle_z = 1/2 \tag{7.2.13}$$

have been used. The result also applies to give $\langle y^2\rangle = \langle x^2\rangle$ and so the average power received is

$$P_{\mathrm{rec}} = \langle r^2\rangle_z = \langle x^2\rangle_z + \langle y^2\rangle_z = \langle G(\theta_i)\rangle_i , \tag{7.2.14}$$

i.e. the signal energy averaged over distance is given by the mean (with respect to angle θ_i) power gain to the sources. The same result can be found from the formal definition of the mean gain using the discretised or continuous pdf from above and this is the same as the received power,

$$P_{\mathrm{rec}} = \langle G(\theta_i)\rangle_i = \sum_i p^{(i)}(\theta_i)\,G(\theta_i) = \int_0^{2\pi} G(\theta)\,p_\theta(\theta)\,\mathrm{d}\theta . \tag{7.2.15}$$

Averaging with distance, which does not change the angular incident power pdf or the antenna gain, gives the mean received power, so the mean gain is equivalent to the received power. The isotropic antenna ($G(\theta) = 1$) gives the obvious result that $\langle G(\theta)\rangle_i = 1$ and the received power is unity.

7.2.4 Statistical quantities for signals from a uniform scenario

Using the above approach with N sources, then the in-phase and quadrature components of the received signal can be expressed, here for a uniform scenario, and isotropic antenna ($G(\theta) = 1$),

$$E\{x\} = E\{y\} = 0, \tag{7.2.16}$$

$$E\{x^2\} = E\{y^2\} = \sigma^2 = N, \tag{7.2.17}$$

$$E\{xy\} = 0. \tag{7.2.18}$$

The envelope is

$$r = \sqrt{x^2 + y^2} \tag{7.2.19}$$

with

$$E\{r\} = \sqrt{\frac{\pi}{2}}\sigma = \sqrt{\frac{\pi N}{2}}, \tag{7.2.20}$$

$$E\{r^2\} = E\{x^2\} + E\{y^2\} = 2\sigma^2 = 2N, \tag{7.2.21}$$

$$\sigma_r^2 = E\{r^2\} - (E\{r\})^2 = \frac{4-\pi}{4}2\sigma^2 = \frac{4-\pi}{2}N, \tag{7.2.22}$$

$$E\{r^4\} = 2\left(E\{r^2\}\right)^2 = 8\sigma^4 = 8N^2, \tag{7.2.23}$$

$$\sigma_{r^2}^2 = E\{r^4\} - \left(E\{r^2\}\right)^2 = 4\sigma^4 = 4N^2. \tag{7.2.24}$$

The median of the envelope is

$$m_r = 0.832\sqrt{2\sigma^2} = 1.176\sqrt{N}. \tag{7.2.25}$$

These represent the true (ensemble) values and the sample values will estimate these with a statistical accuracy according to the sample size.

For a terminal moving at velocity V, the derivatives with respect to distance (or time) are of interest below, and are

$$E\{\dot{x}\} = E\{\dot{y}\} = 0, \tag{7.2.26}$$

$$E\{\dot{x}^2\} = E\{\dot{y}^2\} = \tfrac{1}{2}\left(k_C V\right)^2 N. \tag{7.2.27}$$

7.2.5 Two signals

For two signals, here denoted as functions of time rather than distance (which follows historical convention), the terminology is

$$\begin{aligned} s_1(t) &= x_1(t) + \mathrm{j}y_1(t) = r_1(t)\mathrm{e}^{\mathrm{j}\psi_1(t)} \\ s_2(t) &= x_2(t) + \mathrm{j}y_2(t) = r_2(t)\mathrm{e}^{\mathrm{j}\psi_2(t)}, \end{aligned} \tag{7.2.28}$$

which are complex Gaussians where $E\{x_i y_i\} = 0$. The correlation coefficient of the signals is the fundamental parameter.

For uncorrelated signals, the interpretation is

$$\begin{aligned} E\{x_1 x_2\} &= E\{y_1 y_2\} = 0, \\ E\{x_1 y_2\} &= -E\{x_2 y_1\} = 0. \end{aligned} \tag{7.2.29}$$

For correlated signals, it is convenient to consider them the same signal with a displacement relationship between the signals,

$$\begin{aligned} s_1(t) &= s(t) \\ s_2(t) &= s(t + \Delta t), \end{aligned} \tag{7.2.30}$$

with a suitable correlation function featuring the displacement, Δt. The displacement can be time (or distance) for space diversity studies, or frequency for frequency diversity studies, or a combination of both time and frequency. In the following parts of this section, only one variable is used much of the time, but both distance and frequency can be used as separate parameters for the correlation displacement.

The complex correlation is defined here as

$$\begin{aligned} E\{s_1^* s_2\} &= \tfrac{1}{2}\left((E\{x_1 x_2\} + E\{y_1 y_2\}) + \mathrm{j}\,(E\{x_1 y_2\} - E\{x_2 y_1\})\right) \\ &= R(\Delta t) \quad \Leftrightarrow \quad S(\omega_\mathrm{D}) = S_\mathrm{R}(\omega_\mathrm{D}) + \mathrm{j}S_\mathrm{I}(\omega_\mathrm{D}), \end{aligned} \tag{7.2.31}$$

where the Doppler spectrum $S(\omega_\mathrm{D})$, of the signal $s(t)$, follows the Fourier convention denoted

$$\int R(\Delta t)\mathrm{e}^{-\mathrm{j}\omega_\mathrm{D}\Delta t}\mathrm{d}\Delta t = S(\omega_\mathrm{D}) \quad \Leftrightarrow \quad R(\Delta t) = \frac{1}{2\pi}\int S(\omega_\mathrm{D})\mathrm{e}^{\mathrm{j}\omega_\mathrm{D}\Delta t}\,\mathrm{d}\omega_\mathrm{D}. \tag{7.2.32}$$

A reclarification of terminology is prudent here. In the spatial (u) notation, the 2D Fourier relationship is written

$$R(\Delta z, \omega) \quad \Leftrightarrow \quad P(u, \Delta\tau). \tag{7.2.33}$$

Here, ω refers to the baseband frequency $\omega = \omega_\mathrm{RF} - \omega_\mathrm{c}$ and the $u = k_\mathrm{C}\cos\theta$ is the spatial Doppler frequency in rad m^{-1}; z is the distance variable and τ is the excess time delay (of the effective sources).

In the time notation used in this section, the conversion to the Doppler frequency uses $t = z/V$ and we denote the Doppler frequency, ω, in rad s^{-1}, as noted above, by

$$uV = \omega_\mathrm{D} = V\frac{\omega_\mathrm{c}}{c}\cos\theta. \tag{7.2.34}$$

In terms of the time variable, the Fourier pair is written

$$R(\Delta t, \omega) \quad \Leftrightarrow \quad P(\omega_\mathrm{D}, \Delta\tau), \tag{7.2.35}$$

and for just the time variable, here also expressed using the space variable, the transform is

$$R(\Delta t) \quad \Leftrightarrow \quad S(\omega_\mathrm{D}) \quad \text{or} \quad R(\Delta z) \quad \Leftrightarrow \quad P(u). \tag{7.2.36}$$

From the relations

$$x = r\cos\psi, \qquad y = r\sin\psi, \tag{7.2.37}$$

it is traditional to write the correlation between the complex numbers s_1 and s_2 as

$$R(\Delta t) = R_\mathrm{c}(\Delta t) + \mathrm{j}R_\mathrm{cs}(\Delta t) \tag{7.2.38}$$

where

$$\begin{aligned} R_{\mathrm{c}}(\Delta t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \left(S_{\mathrm{R}}(\omega_{\mathrm{D}})\cos\omega_{\mathrm{D}}\Delta t - S_{\mathrm{I}}(\omega_{\mathrm{D}})\sin\omega_{\mathrm{D}}\Delta t\right)\,\mathrm{d}\omega_{\mathrm{D}}, \\ R_{\mathrm{cs}}(\Delta t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \left(S_{\mathrm{I}}(\omega_{\mathrm{D}})\cos\omega_{\mathrm{D}}\Delta t + S_{\mathrm{R}}(\omega_{\mathrm{D}})\sin\omega_{\mathrm{D}}\Delta t\right)\,\mathrm{d}\omega_{\mathrm{D}}. \end{aligned} \tag{7.2.39}$$

and the integral limits reduce to $\pm\omega_{\mathrm{Dm}} = \pm k_{\mathrm{C}} V$. Since s_1 and s_2 are here defined from the same signal, their (cross) correlation is always an autocorrelation which has the property of being even in its real part and odd in its imaginary part. So the spectrum is always purely real, $S(\omega_{\mathrm{D}}) = S_{\mathrm{R}}(\omega_{\mathrm{D}})$ and is even. This means that

$$\begin{aligned} R_{\mathrm{c}}(\Delta t) &= E\{x_1 x_2\} = E\{y_1 y_2\} = \frac{1}{\pi}\int_0^{\infty} S(\omega_{\mathrm{D}})\cos\omega_{\mathrm{D}}\Delta t\,\mathrm{d}\omega_{\mathrm{D}}, \\ R_{\mathrm{cs}}(\Delta t) &= E\{x_1 y_2\} = -E\{x_2 y_1\} = \frac{1}{\pi}\int_0^{\infty} S(\omega_{\mathrm{D}})\sin\omega_{\mathrm{D}}\Delta t\,\mathrm{d}\omega_{\mathrm{D}}. \end{aligned} \tag{7.2.40}$$

The above definition makes the correlation function R a complex number in general. In some contexts, R is defined traditionally as a real number directly from

$$R^2(\Delta t) = R_{\mathrm{c}}^2(\Delta t) + R_{\mathrm{cs}}^2(\Delta t) \tag{7.2.41}$$

and no magnitude notation, expected to be present for the complex R, is necessary. Similarly, the magnitude notation is omitted in some texts for the corresponding correlation coefficient and its square.

In a measurement, the correlation can be estimated by a calculation over a finite data record. Knowing the pattern and a scenario allows the true value to be found either directly from the summations of the discrete random sources or else from the integrals using the spatial or Doppler spectra of the scenario model. The Doppler spectrum of the scenario is found by transforming from the pdf of the spatial Doppler frequency, $u = k_{\mathrm{C}}\cos\theta$, of the sources, to the pdf of the Doppler frequency ω_{D}, i.e.

$$p_{\omega}(\omega_{\mathrm{D}}) = p_u(u)\left|\frac{\partial u}{\partial\omega_{\mathrm{D}}}\right|. \tag{7.2.42}$$

The Doppler spectrum of the signal received by the antenna is then

$$S(\omega_{\mathrm{D}}) = p_{\omega}(\omega_{\mathrm{D}})\; G_{\omega}(\omega_{\mathrm{D}}) = p_{\theta}(\theta)\; G(\theta)\left|\frac{\partial\theta}{\partial\omega_{\mathrm{D}}}\right| \tag{7.2.43}$$

where $G_{\omega}(\omega_{\mathrm{D}})$ is the power pattern in terms[3] of the Doppler frequency. Taking a 2D isotropic antenna as a reference, the scaling is denoted

$$S(\omega_{\mathrm{D}}) = p_{\omega}(\omega_{\mathrm{D}})\; b, \quad \big(G(\theta) = 1\big); \tag{7.2.44}$$

[3] $G_{\omega}(\omega_{\mathrm{D}})$ can be called the Doppler spectrum of the antenna pattern.

so the scale factor b represents the power available to an isotropic antenna. For N sources, $b = 2N$ where the factor of two results because the in-phase and quadrature parts of the arriving waves (from the effective sources) are both defined as having unit energy.

For a uniform scenario and $G(\theta) = 1$, then

$$S(\omega_D) = \frac{b}{\sqrt{\omega_{\mathrm{Dm}}^2 - \omega_{\mathrm{D}}^2}} \tag{7.2.45}$$

where ω_{Dm} is recalled as the maximum angular Doppler frequency. So for an omni antenna with a gain different to one, the spectrum is just scaled by the gain. If we write the radio frequency as the carrier plus Doppler frequency, $\omega_{\mathrm{RF}} = \omega_{\mathrm{c}} + \omega_{\mathrm{D}}$, then the shifted spectrum is written

$$S(\omega_{\mathrm{RF}}) = \frac{b}{\sqrt{\omega_{\mathrm{Dm}}^2 - \left(\omega_{\mathrm{RF}} - \omega_{\mathrm{c}}\right)^2}}. \tag{7.2.46}$$

In the u notation, the spatial Doppler frequency spectrum is

$$S(u) = \frac{G(\theta)}{\pi\sqrt{k_{\mathrm{C}}^2 - u^2}} = \frac{G\left(\cos^{-1}\left(u/k_{\mathrm{C}}\right)\right)}{\pi\sqrt{k_{\mathrm{C}}^2 - u^2}}. \tag{7.2.47}$$

If the polarisations are separated and a 3D scenario is used, then the formulation is

$$S(\omega_{\mathrm{D}}) = b\left(p_{\hat{\theta}}(\theta, \phi)G_{\hat{\theta}}(\theta, \phi) + p_{\hat{\phi}}(\theta, \phi)G_{\hat{\phi}}(\theta, \phi)\right)\left|\frac{\partial\theta}{\partial\omega_{\mathrm{D}}}\right| \tag{7.2.48}$$

where $p_{\hat{\theta}}(\theta, \phi)$ is the 3D pdf of the vertically polarised power, and $p_{\hat{\phi}}(\theta, \phi)$ is the horizontally polarised power pdf, and the average received power, b, now includes both polarisations.

The joint pdf for the two signals can now be found in terms of the Gaussian components and transformation then gives the pdfs of the envelope and phase quantities.

7.2.6 Moments

The moments of the Doppler frequency, ω_{D}, can now be found from the Doppler spectrum of the sources, $p_\omega(\omega_{\mathrm{D}})$, and the Doppler spectrum of the antenna pattern, $G_\omega(\omega_{\mathrm{D}})$, as:

$$\begin{aligned} b_n = E\{\omega_{\mathrm{D}}^n\} = \int_{-\omega_{\mathrm{m}}}^{\omega_{\mathrm{m}}} \omega^n S(\omega_{\mathrm{D}})\,\mathrm{d}\omega_{\mathrm{D}} &= \int_{-\omega_{\mathrm{m}}}^{\omega_{\mathrm{m}}} \omega_{\mathrm{D}}^n p_\omega(\omega_{\mathrm{D}})G_\omega(\omega_{\mathrm{D}})\,\mathrm{d}\omega_{\mathrm{D}} \\ &= b\int_{-\omega_{\mathrm{m}}}^{\omega_{\mathrm{m}}} \omega_{\mathrm{D}}^n p_\omega(\omega_{\mathrm{D}})\,\mathrm{d}\omega_{\mathrm{D}} \quad \text{(omni)}; \end{aligned} \tag{7.2.49}$$

or, in terms of the spatial Doppler frequency $u = k_C \cos\theta = \omega_D V$ (V is the velocity of the mobile),

$$b_n = E\{\omega_D^n\} = V^n \; E\{u^n\} = V^n \int_{-k_C}^{k_C} u^n \, p_u(u) \; G\left(\cos^{-1}\frac{u}{k}\right) \; \mathrm{d}u; \tag{7.2.50}$$

or, directly from the scenario form,

$$b_n = b\big(\omega_m\big)^n \int_{-\pi}^{\pi} p_\theta(\theta) G(\theta) \; \cos^n\theta \; \mathrm{d}\theta; \tag{7.2.51}$$

or, from the moment theorem,

$$b_n = \begin{cases} (-1)^{n/2} \; R_c^{(n)}(0) & n \text{ even} \\ (-1)^{(n+3)/2} \; R_{cs}^{(n)}(0) & n \text{ odd} \end{cases}. \tag{7.2.52}$$

In the last equation,

$$R^{(n)}(0) = \frac{\mathrm{d}^n}{\mathrm{d}t^n} R(\Delta t)|_{\Delta t = 0}. \tag{7.2.53}$$

So given the pattern and source distribution, all the moments can be found. For N sources and $G(\theta) = 1$, $b = b_0$. For other patterns, the moments, including b_0, can be expressed in terms of b.

The centralised moments include the Doppler spread, which is

$$s_D = \sqrt{\frac{b_2}{b_0} - \left(\frac{b_1}{b_0}\right)^2}. \tag{7.2.54}$$

For wideband considerations, the moment definition requires the delay time variable as well, e.g.

$$b_{n,m} = V^n \int_0^{\infty} \int_{-k_C}^{k_C} u^n \tau^m \, p(u, \tau) \; G\left(\cos^{-1}\frac{u}{k}\right) \; \mathrm{d}u \; \mathrm{d}\tau. \tag{7.2.55}$$

7.2.7 Joint pdfs of two signals

The following sections draw from Rice (1944–45, 1948), Davenport and Root (1958), Middleton (1960), Jakes (1974), and Lee (1982). The derivation details are omitted with the emphasis on basic methods and results.

The joint pdfs are kept compact using matrix notation. Define

$$\boldsymbol{X}^T = [x_1, y_1, x_2, y_2] \tag{7.2.56}$$

and the covariance matrix is the matrix of correlations (the covariance is the same as the correlations since $\boldsymbol{X}$ contains zero mean variables) which is

$$\begin{aligned}\Lambda &= E\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\} \\ &= \begin{bmatrix} \sigma_1^2 & 0 & R_{\mathrm{c}} & R_{\mathrm{cs}} \\ 0 & \sigma_1^2 & -R_{\mathrm{cs}} & R_{\mathrm{c}} \\ R_{\mathrm{c}} & -R_{\mathrm{cs}} & \sigma_2^2 & 0 \\ R_{\mathrm{cs}} & R_{\mathrm{c}} & 0 & \sigma_2^2 \end{bmatrix}.\end{aligned} \tag{7.2.57}$$

The correlations may be functions of both baseband frequency and time displacements, but only time will be used here (narrowband case) for brevity except when discussing explicitly the frequency involvement.

The joint pdf is then written compactly as

$$p(\boldsymbol{X}) = \frac{1}{(2\pi)^2 |\Lambda|^{\frac{1}{2}}} \exp\left[-(1/2|\Lambda|^{\frac{1}{2}})\ \boldsymbol{X}^{\mathrm{T}} \Lambda^{-1} \boldsymbol{X}\right] \tag{7.2.58}$$

or

$$p(x_1, x_2, y_1, y_2) = \frac{1}{(2\pi)^2\ |\Lambda(\Delta t)|^{1/2}} \exp\left(-\frac{1}{2|\Lambda(\Delta t)|^{1/2}} \begin{bmatrix} \sigma^2\left(x_1^2 + x_2^2 + y_1^2 + y_2^2\right) \\ -2R_{\mathrm{c}}(\Delta t)\left(x_1 x_2 + y_1 y_2\right) \\ -2R_{\mathrm{cs}}(\Delta t)\left(x_1 y_2 - x_2 y_1\right) \end{bmatrix}\right), \tag{7.2.59}$$

where

$$|\Lambda(\Delta t)|^{\frac{1}{2}} = (\sigma_1\sigma_2)^2 - \left(R_{\mathrm{c}}^2(\Delta t) + R_{\mathrm{cs}}^2(\Delta t)\right) = (\sigma_1\sigma_2)^2 - R^2(\Delta t). \tag{7.2.60}$$

7.2.8 Joint envelope and joint phase pdfs

To get the envelope and phase quantities, the Jacobian is

$$\frac{\partial(x_1, y_1, x_2, y_2)}{\partial(r_1, \psi_1, r_2, \psi_2)} = r_1 r_2 \tag{7.2.61}$$

and

$$p(r_1, \psi_1, r_2, \psi_2) = \frac{1}{(2\pi)^2 |\Lambda(\Delta t)|^{1/2}} \exp\left(-\frac{1}{2|\Lambda(\Delta t)|^{1/2}} \begin{bmatrix} \sigma^2\left(r_1^2 + r_2^2\right) \\ -2R_{\mathrm{c}}(\Delta t)\ r_1 r_2 \cos(\psi_2 - \psi_1) \\ -2R_{\mathrm{cs}}(\Delta t) r_1 r_2 \sin(\psi_2 - \psi_1) \end{bmatrix}\right) \tag{7.2.62}$$

for the range $r_1, r_2 > 0$ and $0 \leq \psi_1, \psi_2 < 2\pi$.

This can now be reduced to the joint envelope distribution,

$$p_r(r_1, r_2) = \frac{r_1 r_2}{\sigma_1^2 \sigma_2^2 (1 - |\rho|^2)} I_0\left(\frac{|\rho|}{(1 - |\rho|^2)} \frac{r_1 r_2}{\sigma_1 \sigma_2}\right) \times \exp\left[-\frac{1}{2(1 - |\rho|^2)}\left(\left(\frac{r_1}{\sigma_1}\right)^2 + \left(\frac{r_2}{\sigma_2}\right)^2\right)\right] \tag{7.2.63}$$

in which there is the correlation coefficient of the zero mean complex Gaussians,

$$|\rho(\Delta t)|^2 = \frac{R(\Delta t)}{R(0)} = \frac{R_c^2(\Delta t) + R_{cs}^2(\Delta t)}{R_c^2(0) + R_{cs}^2(0)} = \frac{R_c^2(\Delta t) + R_{cs}^2(\Delta t)}{\sigma_1^2 \sigma_2^2} = 4\frac{R_c^2(\Delta t) + R_{cs}^2(\Delta t)}{\sigma_{r_1^2} \sigma_{r_2^2}}. \tag{7.2.64}$$

The power signal and its mean value are respectively denoted

$$\gamma = r^2, \quad E\{r^2\} = \Gamma = 2\sigma^2; \tag{7.2.65}$$

and we can interpret (γ/Γ) as the ratio of the 'instantaneous' carrier-to-noise ratio (γ/N_0) to the mean carrier-to-noise ratio (Γ/N_0). The Jacobian for converting to power quantities is $1/(4r_1 r_2)$ and the joint pdf of the powers follows as

$$p_\gamma(\gamma_1, \gamma_2) = \frac{1}{\Gamma_1 \Gamma_2 (1 - |\rho|^2)} I_0\left(\frac{2|\rho|}{(1 - |\rho|^2)} \sqrt{\frac{\gamma_1 \gamma_2}{\Gamma_1 \Gamma_2}}\right) \times \exp\left[-\frac{1}{(1 - |\rho|^2)}\left(\frac{\gamma_1}{\Gamma_1} + \frac{\gamma_2}{\Gamma_2}\right)\right]. \tag{7.2.66}$$

In a similar way, the joint phase can be found as (Davenport and Root, 1958)

$$p_\psi(\psi_1, \psi_2) = \frac{|\Lambda|^{1/2}}{(2\pi)^2 \sigma^4}\left(\frac{(1 - \beta^2)^{1/2} + \beta(\pi - \cos^{-1}\beta)}{(1 - \beta^2)^{3/2}}\right) \tag{7.2.67}$$

in which

$$\beta = \frac{R_c(\Delta t)}{\sigma^2} \cos(\psi_2 - \psi_1) + \frac{R_{cs}(\Delta t)}{\sigma^2} \sin(\psi_2 - \psi_1). \tag{7.2.68}$$

For uncorrelated signals with equal powers, $|\Lambda|^{1/2} = \sigma^4$ and $p(r_1, r_2) = p(r_1)p(r_2)$, $p(\psi_1, \psi_2) = p(\psi_1)p(\psi_2)$.

The envelope and phase of a single signal are independent, i.e. $p(r, \psi) = p(r)p(\psi)$, which means that samples of the envelope and phase are independent. However, the envelopes and phases of multiple signals are not, in general, independent for multiple signals, i.e. $p(r_1, r_2, \psi_1, \psi_2) \neq p(r_1, r_2)p(\psi_1, \psi_2)$. This means that samples of the envelopes (r_1, r_2) and phases (ψ_1, ψ_2) are not independent. Only when the multiple signals are independent do we get $p(r_1, r_2, \psi_1, \psi_2) = p(r_1)p(r_2)p(\psi_1)p(\psi_2)$.

7.2.9 Envelope correlation coefficients

In electrical engineering convention the relationship between the correlation coefficient, ρ, which is the normalised covariance, and the correlation function, $R_x = E\{x(t)x(t+\Delta t)\}$, is

$$\rho_x(\Delta t) = \frac{R_x(\Delta t) - \left(E\{x(t)\}\right)^2}{R_x(\Delta t = 0) - \left(E\{x(t)\}\right)^2}. \tag{7.2.69}$$

The correlations can be found directly from the pdfs. For the Gaussian components, for example,

$$R_{\mathrm{c}}(\Delta t) = E\{x_1 x_2\} = E\{y_1 y_2\} = \iint x_1 x_2 p(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2. \tag{7.2.70}$$

For an uniform scenario with exponential delay distribution and an omni antenna, this gives the correlation coefficient as (here the frequency displacement is reintroduced)

$$\rho(\Delta t, \Delta\omega) = \frac{R(\Delta t, \Delta\omega)}{R(0,0)} = \frac{J_0(k_{\mathrm{C}} V \Delta t)}{\sqrt{1 + (\Delta\omega)^2 s^2}} \quad \text{or} \quad \rho(\Delta z, \Delta\omega) = \frac{J_0(k_{\mathrm{C}} \Delta z)}{\sqrt{1 + (\Delta\omega)^2 s^2}}, \tag{7.2.71}$$

where s is recalled as the delay spread and the single variable (x and y) quantities from above confirm that

$$R_x(0,0) = \sigma^2 = N. \tag{7.2.72}$$

The envelope correlation is similarly found from $p(r_1, r_2)$ and is (again omitting the frequency displacement for brevity)

$$\begin{aligned} R_r(\Delta t) &= \frac{\pi}{2} b_0\, {}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; 1; \rho^2(\Delta t)\right) \\ &= b_0\left(2E(\rho^2(\Delta t)) - (1 - \rho^2(\Delta t))\, E(\rho^2(\Delta t))\right) \\ &\approx \frac{\pi}{2} b_0 \left(1 + \frac{1}{4}\rho^2(\Delta t)\right). \end{aligned} \tag{7.2.73}$$

Here K and E are the complete elliptic integral of the first and second kind, respectively, and the approximation of the hypergeometric function is good to a couple of percent. The notation is $E(m) = E(k^2)$, where m is the modulus and k is the parameter. The envelope correlation coefficient follows as (see also Section 2.2)

$$\begin{aligned} \rho_r(\Delta t) &= \frac{R_r(\Delta t) - \left(E\{r(t)\}\right)^2}{R_r(0) - \left(E\{r(t)\}\right)^2} \\ &= \frac{2E\left(\rho^2(\Delta t)\right) - \left(1 - \rho^2(\Delta t)\right) K\left(\rho^2(\Delta t)\right) - \pi/2}{2 - \pi/2}. \end{aligned} \tag{7.2.74}$$

Recall that a good approximation is to take the envelope correlation as the same as the square of the Gaussian component correlation, $\rho_r \approx \rho^2 = \rho_{\mathrm{c}}^2 + \rho_{\mathrm{cs}}^2$.

7.2.10 Phase correlation coefficient

The phase correlation is (Middleton, 1960, p. 408; with a modification by Jakes, 1974, p. 53)

$$R_\psi(\Delta t, \Delta\omega) = \pi^2 \Big(1 + 2A\big(\rho_c(\Delta t, \Delta\omega)\big) + 2A^2\big(\rho_c(\Delta t, \Delta\omega)\big) - \tfrac{1}{24}\Omega\big(\rho(\Delta t, \Delta\omega)\big)\Big) \tag{7.2.75}$$

where the correlation coefficients are

$$\rho = \frac{R(\Delta t, \Delta\omega)}{\sigma^2}, \quad \rho_c = \frac{R_c(\Delta t, \Delta\omega)}{\sigma^2}, \tag{7.2.76}$$

and

$$A\big(\rho_c(\Delta t, \Delta\omega)\big) = \frac{1}{2\pi}\ \sin^{-1}\big(\rho_c(\Delta t, \Delta\omega)\big), \tag{7.2.77}$$

$$\Omega\big(\rho(\Delta t, \Delta\omega)\big) = \frac{6}{\pi^2}\sum_{m=1}^{\infty}\frac{\rho^{2m}(\Delta t, \Delta\omega)}{m^2}, \quad \big(\Omega(0) = 0, \Omega(1) = 1\big). \tag{7.2.78}$$

The signal phase is uniformly distributed, i.e. $p_\psi(\psi) = 1/2\pi$, so $\langle\psi_1\rangle = \langle\psi_2\rangle = \pi$ and $\langle\psi_1^2\rangle = \langle\psi_2^2\rangle = 4\pi^2/3$, and the correlation coefficient is

$$\begin{aligned}\rho_\psi(\psi_1, \psi_2) &= \frac{R_\psi(\Delta t, \Delta\omega) - E\{\psi_1\}E\{\psi_2\}}{R_\psi(0, 0) - E\{\psi_1\}E\{\psi_2\}} \\ &= \frac{3}{\pi^2}\big(R_\psi(\Delta t, \Delta\omega) - \pi^2\big) \\ &= 3A(\rho_c)\big(1 + 2A(\rho_c)\big) - \tfrac{1}{8}\Omega\ (\rho),\end{aligned} \tag{7.2.79}$$

and if also the delay profile is exponential, then

$$\rho^2(\Delta t, \Delta\omega) = \rho_c^2(\Delta t, \Delta\omega) = \frac{J_0^2(\omega_{Dm}\Delta t)}{1 + (\Delta\omega \cdot s)^2}, \tag{7.2.80}$$

where ω_{Dm} is recalled as the maximum Doppler frequency. For a zero time displacement, this phase correlation coefficient drops to 0.5, i.e. $p_\psi\big(0, \Delta\omega = B_c\big) = 0.5$ when $B_c = 1/(4\pi s)$. This is half of the frequency spacing for a similar drop in the envelope correlation coefficient. This means that the phase of the received signal is changing more quickly than the envelope.

The phase correlation coefficient function for this example is plotted in Figure 7.2.1 for two cases. The first case is for when the two signals are at the same frequency but are separated in space by Δz. The correlation displacement is given on the abscissa as

$$\omega_{Dm} \cdot \Delta t = \Delta z/\lambda \quad (\Delta z = V\Delta t). \tag{7.2.81}$$

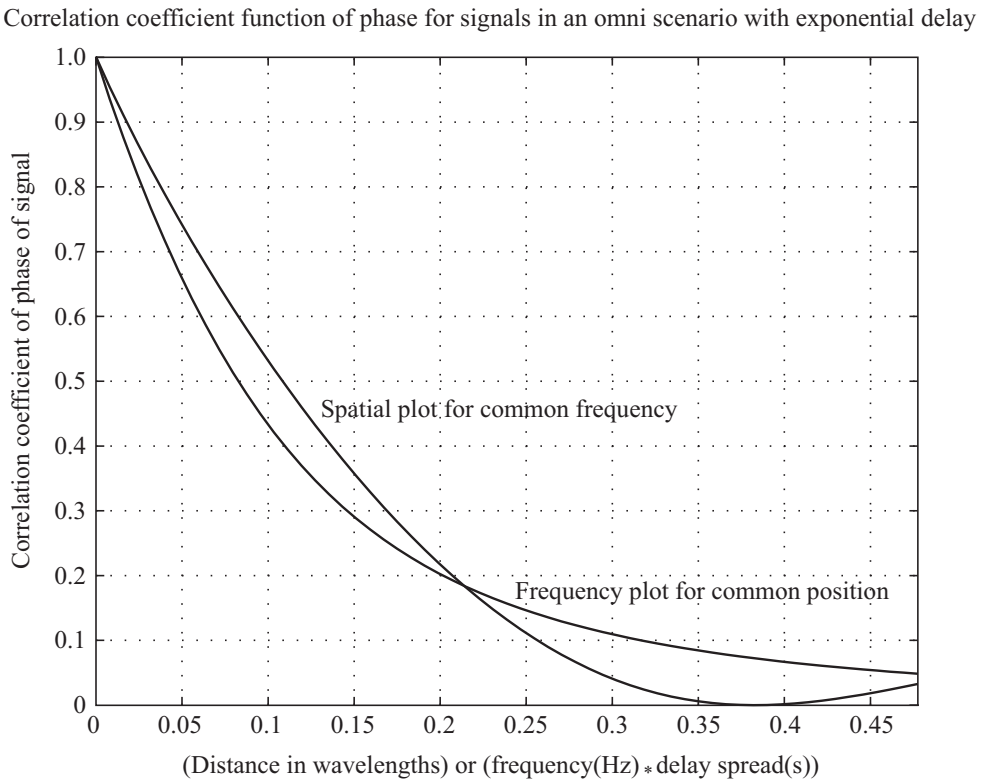


Figure 7.2.1 The correlation coefficient of the phases of signals in a uniform scenario with exponentially distributed delay time. The spatial plot is against distance in wavelengths and the frequency plot is against frequency times delay spread.

The second case is for frequency separation and a common spatial position, and the correlation displacement on the abscissa is

$$s\Delta\omega/2\pi = s \cdot \Delta f \tag{7.2.82}$$

where s is the delay spread and Δf is the frequency spacing in Hz. The curves are similar in form near the origin owing to the source distribution models used (uniform in angular power and exponential in delay time). They drop quickly from a sharp peak at the origin.

7.2.11 Signal dynamics for moving receiver

The time derivatives of the quadrature signals are introduced as variables for the moving receiver case. This allows the calculation of properties such as the phase derivative, or random FM, and also allows level crossing problems to be formulated. The time (or distance) derivatives are denoted by a dot. A single signal calculation uses the joint pdf $p\left(x, y, \dot{x}, \dot{y}\right)$, and for multiple signals the starting point is a joint pdf such as $p\left(x_1, y_1, \dot{x}_1, \dot{y}_1; x_2, y_2, \dot{x}_2, \dot{y}_2\right)$.

In order to calculate the distribution of the maximum of the envelope, we use the covariance matrix (Rice, 1945, p. 80) (again the matrix of correlations is the same as

the matrix of covariances since the means are zero)

$$\begin{bmatrix} x^2 & x\dot{y} & x\ddot{x} & xy & x\dot{x} & x\ddot{y} \\ \dot{y}x & \dot{y}\dot{y} & \dot{y}\ddot{x} & \dot{y}y & \dot{y}\dot{x} & \dot{y}\ddot{y} \\ \ddot{x}x & \ddot{x}\dot{y} & \ddot{x}^2 & \ddot{x}y & \ddot{x}\dot{x} & \ddot{x}\ddot{y} \\ yx & y\dot{y} & y\ddot{x} & y^2 & y\dot{x} & y\ddot{y} \\ \dot{x}x & \dot{x}\dot{y} & \dot{x}\ddot{x} & \dot{x}y & \dot{x}^2 & \dot{x}\ddot{y} \\ \ddot{y}x & \ddot{y}\dot{y} & \ddot{y}\ddot{x} & \ddot{y}y & \ddot{y}\dot{x} & \ddot{y}^2 \end{bmatrix} = \begin{bmatrix} b_0 & b_1 & -b_2 & 0 & 0 & 0 \\ b_1 & b_2 & -b_3 & 0 & 0 & 0 \\ -b_2 & -b_3 & b_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_0 & -b_1 & -b_2 \\ 0 & 0 & 0 & -b_1 & b_2 & b_3 \\ 0 & 0 & 0 & -b_2 & b_3 & b_4 \end{bmatrix} \tag{7.2.83}$$

To convert to envelope and phase quantities, the relations are

$$\begin{aligned} x &= r\cos\psi \\ y &= r\sin\psi \\ \dot{x} &= \dot{r}\cos\psi - r\dot{\psi}\sin\psi \\ \dot{y} &= \dot{r}\sin\psi + r\dot{\psi}\cos\psi \\ \ddot{x} &= \ddot{r}\cos\psi - 2\dot{r}\dot{\psi}\sin\psi - r\dot{\psi}^2\cos\psi - r\ddot{\psi}\sin\psi \\ \ddot{y} &= \ddot{r}\sin\psi + 2\dot{r}\dot{\psi}\cos\psi - r\dot{\psi}^2\sin\psi + r\ddot{\psi}\cos\psi \end{aligned} \tag{7.2.84}$$

with the Jacobian

$$\frac{\partial(x, \dot{x}, \ddot{x}, y, \dot{y}, \ddot{y})}{\partial(r, \dot{r}, \ddot{r}, \psi, \dot{\psi}, \ddot{\psi})} = r^3. \tag{7.2.85}$$

This allows the expression of $p\left(r, \dot{r}, \ddot{r}, \psi, \dot{\psi}, \ddot{\psi}\right)$. Rice notes that the ψ and $\ddot{\psi}$ terms integrate out easily but integrating out the $\dot{\psi}$ term requires the spectrum $S(\omega)$ to be symmetric (i.e. $G(\theta)$ is symmetric about the direction of motion, $\theta = 0$) which means that b_1 and b_3 (and all other odd moments) are zero. Rice used this to calculate $p\left(r, \dot{r}, \ddot{r}\right)$ from which the distribution of the maximum, which is a complicated expression, is found using (Rice, 1945, p. 82)

$$p(r) = -\int_{-\infty}^{0} \ddot{r}\, p(r, \dot{r} = 0, \ddot{r})\, \mathrm{d}\ddot{r}. \tag{7.2.86}$$

Rice's expression of $p\left(r, \dot{r}, \ddot{r}\right)$ is the basic pdf for addressing level crossing problems.

7.2.12 Level crossing problems

The number of level crossings in a positive direction of the envelope at a level $r = R$ is given by Rice, 1948, p. 125, as

$$N(R) = \int_{0}^{\infty} \dot{r}\, p(r = R, \dot{r})\, \mathrm{d}\dot{r} \tag{7.2.87}$$

and it turns out that the envelope and its derivative are statistically independent,

$$p(r, \dot{r}) = p(r) \cdot p(\dot{r}) = \frac{r}{b_0}\, \mathrm{e}^{-r^2/(2b_0)} \cdot \frac{1}{\sqrt{2\pi b_2}}\, \mathrm{e}^{-\dot{r}^2/(2b_2)} \tag{7.2.88}$$

(note the useful result that $\dot{r}$ is zero mean Gaussian) giving the level crossing rate as (Jakes, 1974, p. 34)

$$N(R) = \frac{\overline{R}\mathrm{e}^{-\overline{R}^2}}{\sqrt{\pi b_0}}\sqrt{b_2 - \frac{b_1^2}{b_0}} \tag{7.2.89}$$

where

$$\overline{R} = \frac{R}{\sqrt{E\{r^2\}}} = \frac{R}{\sqrt{2b_0}}. \tag{7.2.90}$$

For a given antenna pattern and scenario, the spectrum $S(\omega)$ is first calculated, and then the moments can be found. Alternatively, the moments are found directly from the scenario equation form, $p_\theta(\theta)$. Then substitution of the moments gives the one-way level crossing rate.

For a uniform scenario, and some simple, symmetric power pattern examples,

$$\begin{aligned} N(R) &= \overline{R}\mathrm{e}^{-\overline{R}^2}\ \omega_m \frac{1}{\sqrt{2\pi}}, \quad G(\theta) = 1, \\ N(R) &= \overline{R}\mathrm{e}^{-\overline{R}^2}\ \omega_m \frac{1}{2\sqrt{\pi}}, \quad G(\theta) = \sin^2\theta, \\ N(R) &= \overline{R}\mathrm{e}^{-\overline{R}^2}\,\omega_m\sqrt{\frac{3}{4\pi}}, \quad G(\theta) = \cos^2\theta. \end{aligned} \tag{7.2.91}$$

Finally, Pätzold *et al.* (1998) give the level crossing rate for the Rician case for $G(\theta) = 1$.

7.2.13 Level crossings of diversity combined signals

For multiple signals, the starting point is again the joint pdf of the component Gaussians, and from

$$\langle x_1 \dot{y}_2 \rangle_i = \langle \dot{x}_1 y_2 \rangle_i = 0, \tag{7.2.92}$$

the independence relation

$$p(x_1, y_1, \dot{x}_1, \dot{y}_1, x_2, y_2, \dot{x}_2, \dot{y}_2) = p(x_1, y_1, \dot{x}_1, \dot{y}_1) \cdot p(x_2, y_2, \dot{x}_2, \dot{y}_2) \tag{7.2.93}$$

results, and this can be used for the level crossing solutions.

Lee (1982) uses the equal gain signal combination

$$r = r_1 + r_2, \qquad \dot{r} = \dot{r}_1 + \dot{r}_2, \tag{7.2.94}$$

to calculate the level crossing rate of a two branch antenna and notes that for more branches than two, the method is too cumbersome and complicated. Lee recommends an approximation which uses the fact that the variables $(x_i, \dot{x}_i, \dot{r}_i)$ are zero mean Gaussians to get

$$p(x_1, y_1, x_2, y_2, \ldots, \dot{r}_1, \dot{r}_2, \ldots) \approx p(r)\,p(\dot{r}) \tag{7.2.95}$$

with a resulting level crossing rate of

$$N(R) = p(r = R) \int_0^\infty \dot{r}\, p(\dot{r})\, \mathrm{d}\dot{r}$$
$$= p(r = R) \frac{\sigma_{\dot{r}}}{2\pi}. \tag{7.2.96}$$

Lee (1982, p. 323) gives the standard deviation of the envelope derivatives for up to four branches:

$$\begin{aligned}
\sigma_{\dot{r}} &= \left(2\langle \dot{r}^2 \rangle - 2\langle \dot{r}_1 \dot{r}_2 \rangle\right)^{1/2}, \quad \text{two branches} \\
&= \left(3\langle \dot{r}^2 \rangle + 4\langle \dot{r}_1 \dot{r}_2 \rangle + 2\langle \dot{r}_1 \dot{r}_3 \rangle\right)^{1/2}, \quad \text{three branches} \\
&= \left(4\langle \dot{r}^2 \rangle + 6\langle \dot{r}_1 \dot{r}_2 \rangle + 4\langle \dot{r}_1 \dot{r}_3 \rangle + 2\langle \dot{r}_1 \dot{r}_4 \rangle\right)^{1/2}, \quad \text{four branches,}
\end{aligned} \tag{7.2.97}$$

where

$$\langle \dot{r}_m \dot{r}_n \rangle = -\frac{\mathrm{d}^2}{\mathrm{d}t^2} \langle r_m\ r_n \rangle\ |_{t=0} \tag{7.2.98}$$

which can be calculated from

$$\langle r_m\ r_n \rangle = R_r = 2\sigma^2 \left(2E\left(\rho_{rmn}\right) - \left(1 - \rho_{rmn}\right) K\left(\rho_{rmn}\right)\right) \tag{7.2.99}$$

in which $2\sigma^2 = R_r(0)$ and the approximation $\rho_{rmn} \approx \rho_{mn}^2$ has been made. Applying this to antenna diversity signals requires that the mutual coupling effects of the antenna elements are included in ρ_r.

Lee (1982) also gives a method to calculate the envelope correlation coefficient for equal-gain combined linear array antennas. Diversity combining is discussed in Section 8.4.

7.2.14 Fade duration

The fade duration has a direct effect on communications through the channel. For a given signal level $r = R$, the mean fade duration is

$$t_{\mathrm{F}}(r = R) = \frac{P\ (r \le R)}{N(r = R)}. \tag{7.2.100}$$

For a given antenna pattern and source distribution, together giving $G(\theta)$, the cdf, $P(r \le R)$, and level crossing rate, $N(r = R)$, are calculated as above and substituted. For a uniform scenario, the envelope cdf is, for the simple symmetric power pattern examples,

$$\begin{aligned}
&P(r \le R) = 1 - \mathrm{e}^{-R^2/(2\sigma^2)}, \qquad G(\theta) = 1, \\
&P(r \le R) = 1 - \mathrm{e}^{-R^2/(\sigma^2)}, \qquad G(\theta) = \sin^2\theta, \text{ or } G(\theta) = \cos^2\theta;
\end{aligned} \tag{7.2.101}$$

and

$$t_{\mathrm{F}}(R) = \frac{\mathrm{e}^{-\bar{R}^2} - 1}{\bar{R}} \frac{1}{\omega_{\mathrm{m}}} \sqrt{2\pi}, \qquad G(\theta) = 1,$$

$$t_{\mathrm{F}}(R) = \frac{\mathrm{e}^{-\bar{R}^2} - 1}{\bar{R}} \frac{1}{\omega_{\mathrm{m}}} \sqrt{4\pi}, \qquad G(\theta) = \sin^2\theta, \tag{7.2.102}$$

$$t_{\mathrm{F}}(R) = \frac{\mathrm{e}^{-\bar{R}^2} - 1}{\bar{R}} \frac{1}{\omega_{\mathrm{m}}} \sqrt{\frac{4\pi}{3}}, \qquad G(\theta) = \cos^2\theta.$$

An example for the omni antenna is given in Figure 7.2.2. In this example, the frequency is 900 MHz and the speed is $10\,\mathrm{km\,h^{-1}}$, so it models a pedestrian cell-phone situation. The 'half wavelength duration' (the time required to travel a half wavelength) of 60 ms is included for reference and occurs at an envelope level of about 1.5 dB below its rms value. The durations can be used for coverage calculations, where the signal-to-noise ratio is the primary parameter. In terms of digital communications, consider a 1200 bps signalling rate. The bit duration is $t_{\mathrm{B}} = 0.83$ ms and from the figure, this is the average fade duration at 35 dB below the signal rms value. If the digital channel is modelled as a simplified Gilbert–Elliot form, which is instantaneously 'good' (e.g. $\langle BER \rangle = 0$ if $t_{\mathrm{F}} > 0.5 t_{\mathrm{B}}$) or 'bad' (e.g. $\langle BER \rangle = 0.5$

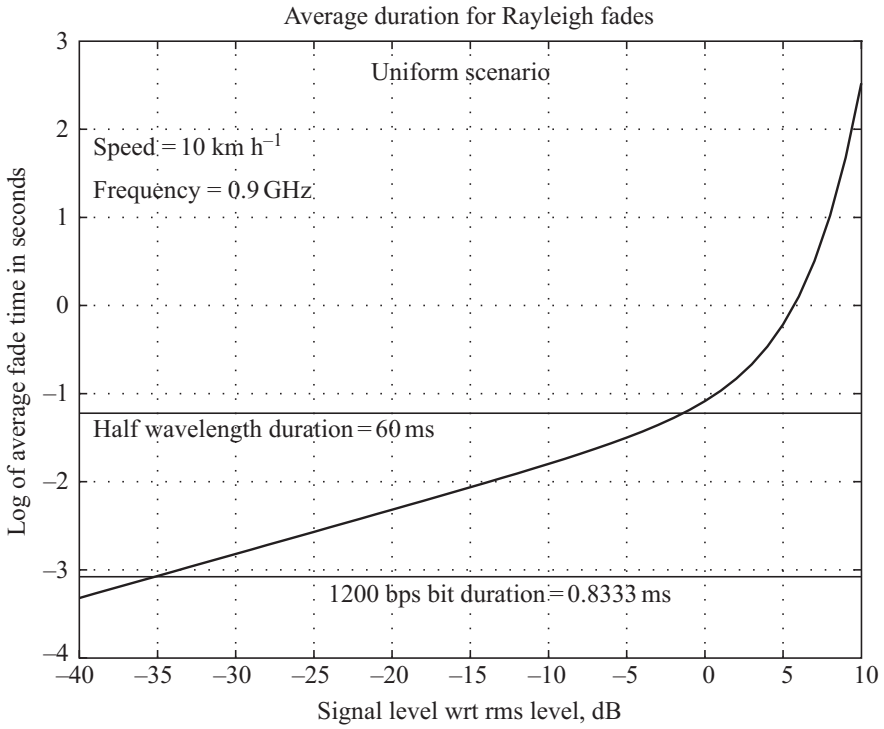


Figure 7.2.2 An example for the average fade duration from an omni antenna in a uniform scenario. The frequency is 1.8 GHz and the speed of the terminal is 10 km h^{-1}. The duration of a bit that has a signalling rate of 1200 bit s^{-1} is included

if $t_F < 0.5t_B$), then the $0.5t_B$ level, which in the example is at 40 dB below the rms signal value, is the primary parameter for performance. The signal level (i.e. SNR) can be set accordingly.

Normally, more sophisticated models are required for data throughput analysis, and these depend on system dependent parameters such as the coding, modulation and receiver structure.

For further results, Abdi *et al*. (2000) give the duration for Rice and Nakagami fading for the uniform case.

7.2.15 Random FM spectrum and rate

The pdf of the random FM, i.e. the phase derivative with respect to time (or distance), is given by Rice (1948, p. 128). By taking a symmetric Doppler spectrum (i.e. an antenna pattern symmetric about $\theta = 0$) so that the odd moments are zero, the pdf reduces to (Rice, 1948, p. 131)

$$\begin{aligned} p(\dot{\psi}) &= \iiint p(r, \dot{r}, \psi, \dot{\psi})\, \mathrm{d}r\ \mathrm{d}\dot{r}\ \mathrm{d}\psi \\ &= \frac{1}{2}\sqrt{\frac{b_0}{b_2}}\left(1 - \frac{b_1^2}{b_0 b_2}\right)\left(1 - 2\frac{b_1}{b_2}\dot{\psi} + \frac{b_0}{b_2}\dot{\psi}^2\right)^{-3/2} \\ &= \frac{1}{2}\sqrt{\frac{b_0}{b_2}}\left(1 + \frac{b_0}{b_2}\dot{\psi}^2\right)^{-3/2} \qquad \text{(symmetric case).} \end{aligned} \tag{7.2.103}$$

The cdf is the student-T form discussed earlier in the context of the group delay (recall that the random FM and the group delay are analogous quantities), and is

$$P\left(\dot{\psi} \le \dot{\Psi}\right) = \frac{1}{2}\left(1 + \frac{\dot{\Psi}}{\sqrt{(b_2/b_0) + \dot{\Psi}^2}}\right). \tag{7.2.104}$$

The Doppler spread is

$$\begin{aligned} s_D &= \sqrt{E\{\omega_D^2\} - (E\{\omega_D\})^2} = \sqrt{b_2/b_0 - (b_1/b_0)^2} \\ &= \sqrt{b_2/b_0} \qquad \text{(symmetric pattern)}, \end{aligned} \tag{7.2.105}$$

and so the $\sqrt{b_0/b_2}$ term is the Doppler spread for the symmetric case and this is recognised as the analogy of delay spread for the distribution of the group delay given in Section 6.2. For an omni antenna in a uniform scenario, for example, the substitution $b_0/b_2 = 2/\omega_{Dm}^2$ (Doppler spread squared) is used.

The power spectrum of the random FM (or, analogously, group delay) is found from the Fourier transform of the autocorrelation (Rice, 1948, pp. 122,124; Jakes, 1974; Lee, 1982) for the symmetric Doppler spectrum case ($b_1 = 0, R'_{cs} = 0,$

$R''_{\rm cs} = 0$, $R'''_{\rm cs} = 0$) as

$$R_{\dot\psi}(\Delta t) = \left\langle \dot\psi(t)\dot\psi(t+\Delta t)\right\rangle = -\frac{1}{2}\left(\left(\frac{R'_{\rm c}(\Delta t)}{R_{\rm c}(\Delta t)}\right)^2 - \frac{R''_{\rm c}(\Delta t)}{R_{\rm c}(\Delta t)}\right)\cdot \ln\left(1-\left(\frac{R_{\rm c}(\Delta t)}{R_{\rm c}(0)}\right)^2\right), \quad (7.2.106)$$

and the spectrum is

$$S_{\dot\psi}(\omega_\psi) = \int_{-\infty}^{\infty} R_{\dot\psi}(\Delta t)\mathrm{e}^{-\mathrm{j}\omega_\psi \Delta t}\,\mathrm{d}\Delta t = 4\int_0^{\infty} R_{\dot\psi}(\Delta t)\cos(\omega_\psi \Delta t)\,\mathrm{d}\Delta t. \quad (7.2.107)$$

This integration has been undertaken for the uniform scenario and omni antenna case (Gans, 1972; Jakes, 1974; Lee, 1982) and has a 'low-pass' form. The low frequency limit is

$$\left\{\omega_\psi \to 0\right\}\ S_{\dot\psi}(\omega_\psi) \to \frac{4\omega_{\rm Dm}}{\pi}\ln(K\omega_\psi), \quad (7.2.108)$$

where K is a constant. There is a 'knee' at $\omega_\psi \sim 2\omega_{\rm m}$, after which the spectrum has the asymptotic form (Gans, 1972)

$$\left\{\mathrm{Lim}\ \omega_\psi \to \infty\right\} S_{\dot\psi}(\omega_\psi) \sim \frac{2\pi}{\omega_\psi}\left(\frac{b_2}{b_0} - \frac{b_1^2}{b_0^2}\right). \quad (7.2.109)$$

For the uniform scenario, $b_0 = \sigma = N,\ b_1 = 0,\ b_2 = Nk_{\rm C}V^2/2$ and

$$\{\mathrm{Lim}\ \omega \to \infty\}\ S_{\dot\psi}(\omega) \sim \frac{2\pi}{\omega}\cdot\frac{1}{2}(k_{\rm C}V)^2 \quad \text{(uniform)}. \quad (7.2.110)$$

7.2.16 Level crossing rate of random FM

Rice (1948, p. 134) gives the level crossing rate of the random FM for the Rayleigh case from the form

$$N\left(\dot\psi = \dot\Psi\right) = \iiiint p\left(r, \dot r, \ddot r, \psi\ \dot\psi = \dot\Psi, \ddot\psi\right)\,\mathrm{d}r\,\mathrm{d}\dot r\,\mathrm{d}\ddot r\,\mathrm{d}\psi\,\mathrm{d}\ddot\psi = \frac{\sqrt{b_4/b_2 - b_2/b_0 + 4\dot\Psi^2}}{2\pi\left(1+(b_0/b_2)\dot\psi^2\right)} \quad \text{(symmetric pattern)}, \quad (7.2.111)$$

and the latter result is only valid for a symmetric Doppler spectrum. For the uniform scenario, the maximum rate is $N(\dot\Psi) = 0.322\beta$, which occurs at $\dot\Psi = 1.45\beta$, where β is the two-sided bandwidth in Hz of an ideal (brickwall) filter centred at the carrier, and which limits the input signal to be strictly narrowband. Rice's original exposition of course gives the formulations for the more generalised Rician envelope, in which a dominant source is present.

7.3 Direct generation of channel functions

If interest is confined to the generation of signals according to some specified distribution, then direct methods are preferable. The relation to a physical channel environment is further removed with this approach, so it is not well suited to antenna problems, for example. The advantage is that signal samples which belong to controllable distributions are generated efficiently. This is not always the case with sampled experimental signals and modelled physical scenario simulations.

The generation of Gaussian variables, including correlated Gaussians with a defined correlation coefficient, is first covered. These techniques are used in the generation of impulse responses and transfer functions, and use of the Kahunen–Loeve expansion for the latter is described.

7.3.1 Gaussian independent samples

For a given statistical distribution, samples can be generated by transforming readily available random variables using the inverse of the desired cdf. For Gaussian variables, the inverse is therefore required of the standard normal function (see Appendix C). This requires a numerical solution, perhaps followed by the use of a look-up table which can then be used for fast computation.

If the tails of the distribution are not critical, the central limit theorem can be 'approximated' by adding only a finite set of random samples. For example (Jeruchim *et al.*, 1992), adding 12 uniformly distributed random numbers occupying [0,1], with mean 1/2 and variance 1/12, to approximate a Gaussian x, which is zero mean and unit variance, is expressed as

$$x = \sum_{k=1}^{12} u(k) - 6. \tag{7.3.1}$$

The numerical solution for the inverse of the standard normal function, look-up table, or above approximation may not be adequate for some applications. Another convenient computation method, which is also accurate, is the Box–Muller (1958) technique, which generates two independent Gaussian variables (x, y) from two independent uniformly distributed variables (u_1, u_2). It applies the inverse cdf to the envelope, $r = \sqrt{x^2 + y^2}$, and phase, $\theta = \tan^{-1}(y/x)$. The relations between the envelope and a uniform variable, u_1; and the phase, and the other uniform variable, u_2; are:

$$u_1 = P(r \geq r_0) = \mathrm{e}^{-r_0^2/(2\sigma^2)}, \quad r \geq 0; \tag{7.3.2}$$

$$u_2 = P(\theta \geq \theta_0) = \frac{1}{2\pi}\theta_0, \quad 0 \leq \theta < 2\pi, \tag{7.3.3}$$

where σ^2 is the variance of the Gaussian components. For unit variance, the Gaussians follow as

$$x = \sqrt{-2\ln u_1}\cos 2\pi u_2; \quad y = \sqrt{-2\ln u_1}\sin 2\pi u_2. \tag{7.3.4}$$

So for generating a sequence of independent Gaussian samples, the Box–Muller algorithm which transforms $2M$ independent uniform random numbers into M complex Gaussian samples follows as:

- choose: μ, σ^2;
- for $k = 1$ to M: generate u_{2k-1} and u_{2k} and then

$$x_{2k-1} = \mu + \sigma\sqrt{-2\ln u_{2k-1}}\sin 2\pi u_{2k},$$
$$x_{2k} = \mu + \sigma\sqrt{-2\ln u_{2k-1}}\cos 2\pi u_{2k}, \tag{7.3.5}$$

  and the x_{2k} variable can be taken as the y_{2k-1}.

7.3.2 Gaussian correlated samples

For generating correlated Gaussian samples, the process of decorrelation is reversed (Yong, 1990). The starting point is a pair of Gaussian variables (x_{A1}, x_{A2}) with standard deviations $(\sigma_{A1}, \sigma_{A2})$, means (μ_{A1}, μ_{A2}), and which are correlated. The correlation coefficient (normalised covariance between x_{A1} and x_{A2} with zero lag) is

$$\rho_{A1A2}(0) = \rho = \frac{E\{x_{A1}x_{A2}\} - \mu_{A1}\mu_{A2}}{\sigma_{A1}\sigma_{A2} - \mu_{A1}\mu_{A2}}, \tag{7.3.6}$$

where $|\rho| < 1$.

From here, a pair of uncorrelated Gaussians can be derived by axis rotation. Specifically, a 45° counterclockwise rotation results in uncorrelated Gaussians with variances $1 \pm \rho$. The transform to zero mean, uncorrelated Gaussian variables (x_{B1}, x_{B2}) with unit variances, is

$$\begin{bmatrix} x_{B1} \\ x_{B2} \end{bmatrix} = \begin{bmatrix} (1-\rho)^{-1/2} & 0 \\ 0 & (1+\rho)^{-1/2} \end{bmatrix} \begin{bmatrix} 2^{-1/2} & -2^{-1/2} \\ 2^{-1/2} & 2^{-1/2} \end{bmatrix} \begin{bmatrix} \sigma_{A1}^{-1} & 0 \\ 0 & \sigma_{A2}^{-1} \end{bmatrix} \begin{bmatrix} (x_{A1} - \mu_{A1}) \\ (x_{A2} - \mu_{A2}) \end{bmatrix}. \tag{7.3.7}$$

The steps are, from right to left:

- setting the means to zero;
- setting the variances to unity;
- rotating the axes by 45° (this produces variances of $1 \pm \rho$ and zero correlation);
- resetting the variances to unity.

The reverse process follows as

$$\begin{bmatrix} x_{A1} \\ x_{A2} \end{bmatrix} = \boldsymbol{C} \begin{bmatrix} x_{B1} \\ x_{B2} \end{bmatrix} + \begin{bmatrix} \mu_{A1} \\ \mu_{A2} \end{bmatrix}, \tag{7.3.8}$$

where

$$\begin{aligned} \boldsymbol{C} &= \begin{bmatrix} \sigma_{A1} & 0 \\ 0 & \sigma_{A2} \end{bmatrix} \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \\ -2^{-1/2} & 2^{-1/2} \end{bmatrix} \begin{bmatrix} \sqrt{1-\rho} & 0 \\ 0 & \sqrt{1+\rho} \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{A1}\sqrt{(1-\rho)/2} & \sigma_{A1}\sqrt{(1+\rho)/2} \\ -\sigma_{A2}\sqrt{(1-\rho)/2} & \sigma_{A2}\sqrt{(1+\rho)/2} \end{bmatrix}. \end{aligned} \tag{7.3.9}$$

Similarly to the forward process, the steps in $\boldsymbol{C}$, from right to left, are:

- setting the variances to unity;
- rotation by $-45°$;
- setting the variances to σ_{A1}, σ_{A2};

and so this reverse procedure allows the generation of pairs of Gaussians with a pre-defined correlation coefficient.

Note that this process generates two Gaussian sequences whose corresponding samples are correlated with each other in the ensemble sense, but neither of the sequences $\{x_{A1}\}$, $\{x_{A2}\}$ has an autocorrelation function defined within the process – this depends on the autocorrelation of the input sequence, i.e. the consecutive samples of both $\{x_{B1}\}$ and $\{x_{B2}\}$.

An alternative method (Yong, 1990) requires the input of three, instead of two, independent Gaussian variables, but is readily extended to many dimensions. Denote the input Gaussians, which are independent, zero mean and unity variance, as x_{B1}, x_{B2}, x_{B3}. Then the correlated Gaussians, x_{A1}, x_{A2}, with correlation coefficient, ρ, are:

$$\begin{aligned} x_{A1} &= \sigma_{A1}\left(\sqrt{|\rho|}\, x_{B1} + \sqrt{1-|\rho|^2}\, x_{B2}\right) + \mu_{A1}; \\ x_{A2} &= \sigma_{A2}\left(\sqrt{|\rho|}\, x_{B1} \pm \sqrt{1-|\rho|^2}\, x_{B3}\right) + \mu_{A2}; \end{aligned} \tag{7.3.10}$$

where the sign choice follows the sign of ρ. This process is constrained to have a covariance matrix form of

$$\begin{aligned} \sigma_{ij} &= \sigma_i^2 \quad \text{for } i = j \\ &= \lambda_i \lambda_j \quad \text{for } i \neq j, \end{aligned} \tag{7.3.11}$$

which is efficient but very restricting. The process for generating N correlated Gaussian samples this way is:

$$\begin{bmatrix} x_{A1} \\ x_{A2} \\ . \\ . \\ . \\ x_{AN} \end{bmatrix} = \begin{bmatrix} \lambda_1 & \sqrt{1-\lambda_1^2} & 0 & 0 & . & . & 0 \\ \lambda_2 & 0 & \sqrt{1-\lambda_2^2} & 0 & 0 & . & 0 \\ \lambda_3 & 0 & 0 & \sqrt{1-\lambda_3^2} & 0 & . & 0 \\ . & . & . & . & . & . & 0 \\ \lambda_N & 0 & 0 & 0 & 0 & . & \sqrt{1-\lambda_N^2} \end{bmatrix} \cdot \begin{bmatrix} x_{B1} \\ x_{B2} \\ . \\ . \\ . \\ x_{BN} \\ x_{BN+1} \end{bmatrix} + \begin{bmatrix} \mu_{A1} \\ \mu_{A2} \\ . \\ . \\ . \\ \mu_{BN} \end{bmatrix} \tag{7.3.12}$$

and the correlation coefficients between x_{Ai} and x_{Aj} are

$$\rho_{ij} = \lambda_i \lambda_j, \quad i \neq j. \tag{7.3.13}$$

As an example, if we take a correlation structure

$$[\rho_{11}, \rho_{12}, \rho_{13}, \rho_{14}, \rho_{15}] = \left[1, \tfrac{1}{2}, 0, -\tfrac{1}{4}, 0\right] \tag{7.3.14}$$

the parameters are

$$[\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5] = \left[\sqrt{\tfrac{1}{2}}, \sqrt{\tfrac{1}{2}}, 0, -\sqrt{\tfrac{1}{8}}, 0\right] \tag{7.3.15}$$

and this then also yields $\rho_{23} = 0,\ \rho_{24} = -1/4,\ \rho_{25} = 0,\ \rho_{34} = 0$, etc., to complete the correlation matrix, which is clearly not Toeplitz. This example demonstrates the limitations; for example the process is not wide-sense stationary (e.g. $\rho_{12} \neq \rho_{23}$), although this may not matter for some applications. Toeplitz correlation matrices, for a wide-sense stationary sequence, are also possible. The example is $\rho_{ij} = 1/v^2$ (all $i \neq j$) and this is given by $\lambda_i = 1/v$, all i.

7.3.3 General correlation matrix case

The general case is when the correlation matrix is fully defined and arbitrary. The following method does not require the one extra independent variable (relative to the number of variables to be generated) as an input but requires more computation.

For the Gaussian sequences $\boldsymbol{x_A} = [x_{A1}, x_{A2}, \ldots]^{\mathrm{T}}$ which have an arbitrary (but positive definite) correlation structure given by the covariance matrix

$$\sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & . \\ \sigma_{21} & \sigma_{22} & . \\ . & . & . \end{bmatrix}, \tag{7.3.16}$$

where $\sigma_{ij} = E\{x_{A1} x_{A2}\}$, it is necessary to decompose the matrix as $\sigma = \tau \tau^{\boldsymbol{T}}$. The correlation coefficients are related to the covariances through

$$\rho_{ij} = \frac{\sigma_{ij}^2 + \mu_i \mu_j}{\sigma_i \sigma_j + \mu_i \mu_j}. \tag{7.3.17}$$

The generation of the required Gaussians is then

$$\boldsymbol{x_A} = \tau \boldsymbol{x_B} + \boldsymbol{\mu}, \tag{7.3.18}$$

where $\boldsymbol{x_B}$ is the set of independent, zero mean input Gaussian sequences and $\boldsymbol{\mu}$ is the vector of means. Standard (Choleski) decomposition to a lower triangle form for the

covariance matrix is

$$\begin{aligned}
\tau_{ij} &= 0 && \text{for } 1 \le i < j \le n \text{ (upper triangle)},\\
\tau_{11} &= \sqrt{\sigma_{11}}, &&\\
\tau_{i1} &= \frac{\sigma_{i1}}{\sqrt{\sigma_{11}}}, && \text{for } i = 2 \text{ to } n \text{ (first column)},\\
\tau_{jj} &= \sqrt{\sigma_{jj} - \sum_{r=1}^{j-1} \tau_{jr}^2}, && \text{for } j = 2 \text{ to } n \text{ (rest of diagonal)},\\
\tau_{ij} &= \frac{1}{\tau_{jj}} \sqrt{\sigma_{ij} - \sum_{r=1}^{j-1} \tau_{ir}\tau_{jr}}, && \text{for } j < i \text{ and } i = 2 \text{ to } n-1 \text{ (rest of lower triangle)}.
\end{aligned} \tag{7.3.19}$$

The algorithm is summarised:

- input: the number of Gaussians, n, required samples in each of the sequences, N, and the set of means, $\boldsymbol{\mu}$;
- input the correlation structure, i.e. covariance matrix σ;
- calculate τ using (for example) Choleski decomposition;
- generate the sequences of length n uncorrelated Gaussians $\boldsymbol{x_B}$ using (for example) the Box–Muller technique;
- transform to the correlated Gaussians $\boldsymbol{x_A} = \tau \boldsymbol{x_B} + \boldsymbol{\mu}$.

An example is to generate five sequences $\{x_{A1}\}$, $\{x_{A2}\}$, $\{x_{A3}\}$, $\{x_{A4}\}$, $\{x_{A5}\}$ with zero means, unity variances and a Toeplitz correlation structure given by:

$$\rho_{1i} = J_0\left(2\pi(i-1)^*0.25\right), \quad i = 1, 2, 3, 4 \tag{7.3.20}$$

(corresponds to samples every quarter wavelength in a uniform scenario) which results in, for $t_n = 1$ to N,

$$\begin{aligned}
x_{A1}(t_n) &= 1 \cdot x_{B1}(t_n),\\
x_{A2}(t_n) &= 0.472 \cdot x_{B1}(t_n) + 0.882 \cdot x_{B2}(t_n),\\
x_{A3}(t_n) &= -0.304 \cdot x_{B1}(t_n) + 0.698 \cdot x_{B2}(t_n) + 0.648 \cdot x_{B3}(t_n),\\
x_{A4}(t_n) &= -0.266 \cdot x_{B1}(t_n) - 0.203 \cdot x_{B2}(t_n) + 0.822 \cdot x_{B3}(t_n) + 0.461 \cdot x_{B4}(t_n),\\
x_{A5}(t_n) &= 0.220 \cdot x_{B1}(t_n) - 0.420 \cdot x_{B2}(t_n) + 0.086 \cdot x_{B3}(t_n) + 0.814 \cdot x_{B4}(t_n)\\
&\quad + 0.326 \cdot x_{B5}(t_n).
\end{aligned} \tag{7.3.21}$$

Note that the correlated samples are indexed by the Gaussian sequence number, i.e. t_n is constant (n indexes the sample in each sequence). The sets of samples are wide sense stationary sequence from the Toeplitz nature of the correlation matrix. The correlation within the sequence of any one of the Gaussians, i.e. indexed by the t_n, is not addressed by this algorithm and with the assumption of independent samples in the input sequences, there is no correlation between output samples for non-zero lag. This can be stated

$$E\left\{x_{Ai}(t_n)\, x_{Ai}(t_m)\right\} = 0, \quad \text{all } i, \text{ all } m \neq n. \tag{7.3.22}$$

(For zero lag ($m = n$), the autocorrelation coefficient is unity of course.)

The correlation function for a set of samples is given by $\rho(\Delta t = \Delta z/V) = J_0(k\Delta z)$ and this can be used to produce the complex Gaussian $z(t_n) = x_{An}(1) + \mathrm{j}x_{An}(2)$, which will feature the correlation function $\rho_{|z|^2} \approx J_0^2(k\Delta z)$ (see Appendix C). For M independent complex Gaussians, the formula is simply

$$z_m(t_n) = x_{An}(2m-1) + \mathrm{j}x_{An}(2m), \quad m = 1, 2, \ldots, M. \tag{7.3.23}$$

7.3.4 Moving average technique for generating correlated signals

Another way to view the generation of correlated samples is the filtering of uncorrelated samples, also known as the moving average process. This requires more input samples and requires more computation to generate a given number of correlated samples than the above method, but is simpler to follow. A sequence of independent Gaussian samples, $\{x\}$, is generated and then filtered using a FIR structure with weights w_n, i.e. the weights are convolved with the input sequence,

$$y_m = \sum_{n=0}^{N} w_n \, x_{m-n}, \tag{7.3.24}$$

where $\{y\}$ is a moving average process of order N and it is seen to filter out the variations between the independent input samples. The gain of the FIR filter, which determines the variance of $\{y\}$, is $\sum_{n=0}^{N} w_n^2$. (Note that this is unity in the above example.) If the input samples are stationary, then so are the output samples (see Appendix B). The requirement is to find the filter coefficients from the chosen correlation function.

The correlation function of the output samples is given by

$$\begin{aligned} R_y(n\Delta t) &= \left(\frac{1}{\sum_{l=0}^{N} w_l^2}\right) \frac{N}{N-i} \sum_{m=n}^{N} w_m \, w_{m-n} \quad 0 \le |n| \le |N|, \\ &= 0 \quad \text{elsewhere;} \end{aligned} \tag{7.3.25}$$

where Δt is the interval between the samples. The power spectrum is given by the discrete Fourier transform of the correlation function, $P_y(\omega_n) \quad \Leftrightarrow \quad R_y(n\Delta t)$, which is

$$P_y(\omega_n) = \frac{1}{2\pi}\left(\frac{1}{\sum_{l=0}^{N} w_l^2}\right)\left|\sum_{m=0}^{N} w_m \, \mathrm{e}^{\mathrm{j}\omega_n x_n}\right|^2. \tag{7.3.26}$$

Taking the square root of the power spectrum and the inverse Fourier transform, gives the filter, here in continuous form,

$$w(t) = \int_{-\infty}^{\infty} \sqrt{P_y(\omega)} \, \mathrm{e}^{-\mathrm{j}\omega t} \, \mathrm{d}\omega, \tag{7.3.27}$$

and the samples of this are the filter coefficients.

An example is a correlation function (and its spectrum) which has Gaussian form,

$$R_y(n\Delta t) = \mathrm{e}^{-(n\Delta t/\tau_0)^2}, \tag{7.3.28}$$

$$\Rightarrow P_y(\omega_n) = \frac{1}{2\pi}\sum_{m=0}^{N} \mathrm{e}^{-(m\Delta t/\tau_0)^2}\mathrm{e}^{\mathrm{j}(m\Delta t)\omega_n}, \tag{7.3.29}$$

where τ_0 is the correlation time defined as the time lag for the correlation coefficient to fall to 1/e. For an infinite series for the correlation function, the continuous spectrum is

$$P_y(\omega) = \frac{\tau_0}{2\sqrt{\pi}}\,\mathrm{e}^{-(\tau_0^2\omega)/4}, \tag{7.3.30}$$

and so the filter is

$$w(t) = \frac{w_0}{\sqrt{\tau}_0}\,\mathrm{e}^{-2(t/\tau_0)^2} \tag{7.3.31}$$

or

$$w_n = \frac{w_0}{\sqrt{\tau}_0}\,\mathrm{e}^{-2(n\Delta t/\tau_0)^2}, \tag{7.3.32}$$

where w_0 is a normalising constant. Maintaining unity gain for the filter preserves the variance between the input and output sequences.

This method has been used (cf. Ogilvy, 1991) also for generating random surfaces for studying rough surface scattering (Section 3.2).

7.3.5 Generation of transfer functions in the time (space) domain

Hashemi (1979) generated snapshots at time t_n (or position, z_n) of the time domain channel transfer function, i.e. impulse response, $h(\tau)$, which is written

$$h(\tau, t_n) = \sum_i a_i(t_n)\,\mathrm{e}^{\mathrm{j}\theta_i(t_n)}\,\delta\left(t - \tau_i(t_n)\right). \tag{7.3.33}$$

Each time sample, t_n (or position sample for the moving receiver, spaced by $\Delta z = V/\Delta t$), the sets of amplitudes, a_i, phases, θ_i, and delay times, τ_i, follows defined distributions, and their evolution with time (or position) follows defined correlation functions.

The amplitude terms are drawn from a distribution such as Rayleigh, log-normal or Suzuki, Nakagami-m, etc. Their choice of statistical evolution is typically chosen from spatial considerations; for example, the classical Clarke scenario results in the envelope correlation function $\rho_{|h|^2}(\Delta z) \approx J_0^2(k\Delta z)$. The phases are uniformly distributed and their evolution can be expressed $\{\theta_i\}_n = \{\theta_i\}_{n-1} - u_i\,\Delta z$, also from classical spatial considerations.

For the arrival times, the modified Poisson process is used traditionally (Turin, 1972) because it caters for clusters of wave arrival times which had been observed experimentally by Turin (1972) and Suzuki (1977). Wu and Aghvami (1996) use an exponential distribution for τ in small cells which have a line of sight, and for large

cells where there may be no line of sight, they suggest using the more generalised Weibull distribution. The number of paths can also be assigned a distribution, e.g. Wu and Aghvami (1996) suggest the beta distribution. The transfer function of the changing channel can be found by Fourier transformation, for example $h(\tau, z) \Leftrightarrow H(\omega, z)$, i.e.

$$H(\omega, z) = \int_{-\infty}^{\infty} h(\tau, z)\, \mathrm{e}^{-\mathrm{j}\omega\tau}\, \mathrm{d}\tau, \tag{7.3.34}$$

as discussed in the previous sections, and the lower integral limit caters for time shifts which may result in non-causal impulse responses.

7.3.6 Generation of transfer functions in the frequency domain

If random transfer functions are required, then for complicated (comprising many paths) transfer functions, it is more efficient to generate directly in the frequency domain, rather than time domain generation followed by Fourier transformation. A fixed time delay, t_0, is incorporated using a linear phase shift in frequency and similarly a frequency offset between transmitter and receiver is included in the formulation, but excluded from the transfer function generation as explained below. The transfer functions of the wideband channel can be generated using the Kahunen–Loeve orthogonal expansion of a non-periodic process (Loeve,1955; Davenport and Root, 1958). The following draws from Clark (1992).

The equivalent baseband signal channel includes the filtering action of the antennas, front ends, etc. An equivalent transmit filter, $C_{\mathrm{Tx}}(\omega)$, and a receive filter, $C_{\mathrm{Rx}}(\omega)$, are now included in the model.

The inclusion of a frequency offset between the transmitter and receiver, $\omega_{\mathrm{Off}} = \omega_{\mathrm{Tx}} - \omega_{\mathrm{Rx}}$, produces a 'phase roll' at the receiver given by $\mathrm{e}^{\mathrm{j}\omega_{\mathrm{Off}}t}$. In the baseband signal channel, this appears explicitly in the impulse response and the transfer function:

$$\begin{aligned} h(t) &= \left(c_{\mathrm{Tx}}(t - t_n)\, \mathrm{e}^{\mathrm{j}\omega_{\mathrm{Off}}t}\right) * h_{\mathrm{P}}(t) * c_{\mathrm{Rx}}(t) \quad \Leftrightarrow \\ H(\omega) &= C(\omega) \cdot H_{\mathrm{P}}(\omega)\, \mathrm{e}^{-\mathrm{j}(\omega+\omega_{\mathrm{Off}})t_0}, \end{aligned} \tag{7.3.35}$$

where

$$C(\omega) = C_{\mathrm{Tx}}(\omega + \omega_{\mathrm{Off}}) \cdot C_{\mathrm{Rx}}(\omega) \tag{7.3.36}$$

is the composite baseband filter response and t_0 is an arbitrary time shift referred to the transmit filter; t_0 is included for generality; for example in multi-user simulations, it is of interest to have an ensemble of channel snapshots as being from different co-channel users (or interferers) which have different symbol times. For single user simulations, the t_0 can be set to zero. Note that the time argument of the rolling phase function cannot be interchanged with distance as in many of the other situations in the text. (The phase roll time is not a propagation parameter and is independent of the position of the mobile terminal.) $H_{\mathrm{P}}(\omega)$ denotes the baseband propagation channel.

The base eqivalent filters are limit in an ideal sense to a bandwidth $-\Omega/2 \leq \omega \leq \Omega/2$. This means that

$$C_{\mathrm{Tx}}(\omega) = 0, \quad |\omega| \geq \frac{\Omega}{2}; \tag{7.3.37}$$

$$C(\omega) = 0, \quad |\omega + \omega_{\mathrm{Off}}| \geq \frac{\Omega}{2}. \tag{7.3.38}$$

This completes the context for the channel definitions, and the focus is now placed on generating the baseband signal channel $H(\omega) = C(\omega)H_{\mathrm{P}}(\omega)$ as a function of time (i.e. position), which is denoted $H(\omega, z)$.

From now on, the phase roll function, while maintained in the formulation for reference, is set to zero ($\omega_{\mathrm{Off}} = 0$) so that the remaining time arguments can be interchanged with distance $z = Vt$ in the usual way. The fading, time-varying channel, $H(\omega, z)$, is taken as a zero-mean, complex Gaussian, wide-sense stationary random process, and has autocorrelation functions which are assumed separable, denoted:

$$\begin{aligned} R_H = E\{H^*(\omega_1, z_1)\ H(\omega_2, z_2)\} &= C^*(\omega_1)\ C(\omega_2)\ R_{HP}(\omega_1, \omega_2; z_1, z_2) \\ &= C^*(\omega_1)\ C(\omega_2)\ R_{HP}(\Delta\omega, \Delta z) \\ &= C^*(\omega_1)\ C(\omega_2)\ R_{HP\omega}(\Delta\omega)\ R_{HPz}(\Delta z) \\ &= R_{H\omega}(\omega_1, \omega_2)\ R_{Hz}(z_1, z_2). \end{aligned} \tag{7.3.39}$$

The KL expansion will give orthogonal, position- (or time-) varying coefficients. The expansion is expressed (Davenport and Root, 1958) as

$$H_{\mathrm{P}}(\omega, z) = \sum_{i=1}^{I} \sqrt{\lambda_i}\ c_i(z)\ \phi_i(\omega), \tag{7.3.40}$$

in which the equality holds in the limit of $I \to \infty$ for the continuous case, the λ_i are the set of I real eigenvalues, the $c_i(z)$ are zero-mean complex Gaussian random processes with unity variance and are uncorrelated in the sense

$$E\left\{c_i^* c_j\right\} = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{7.3.41}$$

and $\phi_i(\omega)$ are a set of I orthonormal eigenfunctions such that

$$\int_{-\Omega/2}^{\Omega/2} \phi_i^*(\omega)\ \phi_j(\omega)\ \mathrm{d}\omega = \delta_{ij}. \tag{7.3.42}$$

The $c_i(z)$ evolve following a specified distance (or time) correlation function. Note that the c_i are not related to $c_{\mathrm{Tx}}(t)$, $c_{\mathrm{Rx}}(t)$ or $C(\omega)$ above. The evolution of $c_{\mathrm{i}}(t)$ can be governed from the usual $\rho_{|c|^2}(\Delta z) \approx J_0^2(k\Delta z)$, for example, and be calculated from the preceding sections.

The eigensystem is solved from the integral equations

$$\int_{-\Omega/2}^{\Omega/2} R_{H\omega}(\omega_1, \omega_2)\, \phi_i(\omega_2)\, \mathrm{d}\omega_2 = \lambda_i \phi_i(\omega_1), \quad |\omega_1| \leq \frac{\Omega}{2}, \quad 0 \leq i \leq I, \tag{7.3.43}$$

once the correlation functions are specified.

The matrix formulation for the expansion is as follows. For generating M snapshot samples of the transfer function,

$$\boldsymbol{H}(\omega) = \begin{bmatrix} H_1(\omega) \\ H_2(\omega) \\ . \\ . \\ H_M(\omega) \end{bmatrix} = \boldsymbol{c} \Lambda^{1/2}\, \phi(\omega) \tag{7.3.44}$$

where

$$\boldsymbol{c} = \begin{bmatrix} c_{1,1} & c_{1,2} & \ldots & c_{1,I} \\ c_{2,1} & c_{2,2} & \ldots & c_{2,I} \\ . & . & . & . \\ . & . & . & . \\ c_{M,1} & c_{M,2} & \ldots & c_{M,I} \end{bmatrix}, \tag{7.3.45}$$

$$\Lambda = \mathrm{diag}\,[\lambda_1, \lambda_2, ..., \lambda_I]\,, \tag{7.3.46}$$

$$\phi(\omega) = [\phi_1(\omega)\ \phi_2(\omega)\ \ldots\ \phi_J(\omega)]^T\,, \tag{7.3.47}$$

and the orthonormal property is written

$$\int_{-\Omega/2}^{\Omega/2} \phi^*(\omega)\, \boldsymbol{\phi^T}(\omega)\, \mathrm{d}\omega = \overline{\boldsymbol{I}}, \tag{7.3.48}$$

where $\overline{\boldsymbol{I}}$ is an $I \times I$ identity matrix. The baseband signal channel transfer function is then

$$\boldsymbol{H}(\omega) = \begin{bmatrix} Q_1(\omega) \\ Q_2(\omega) \\ . \\ Q_M(\omega) \end{bmatrix} = \boldsymbol{c}\ \Lambda^{1/2}\, \phi(\omega)\, \mathrm{e}^{-\mathrm{j}\omega t_0}, \quad |\omega + \omega_{\mathrm{Off}}| \leq \frac{\Omega}{2}, \tag{7.3.49}$$

and $\omega_{\mathrm{Off}} = 0$ as explained above.

To generate random sets of $\boldsymbol{H}(\omega)$:

- generate $M \cdot I$ independent Gaussians with zero mean and unity variance for $\boldsymbol{c}$;
- if desired, generate uniform variables from the range $[-T_S/2, T_S/2]$ (T_S is symbol time) for the t_0;
- input $C(\omega)$ and $P_{hP}(\tau)$, generate $\Lambda^{1/2}\phi(\omega)$ (this is done just once at the start of ensemble generation and requires solving the eigensystem);
- calculate $\boldsymbol{H}(\omega)$.

7.3.7 Continuous delay profile

For solving the KL expansion for a continuous delay profile, the integral equation can be expressed as the system of I linear equations,

$$\boldsymbol{RD}\ \phi_i = \lambda_i\ \phi_i \tag{7.3.50}$$

where I is odd,

$$[\boldsymbol{R}]_{j,j'} = R_H(\omega_j, \omega_{j'}) = R_H(\omega_j - \omega_{j'}). \tag{7.3.51}$$

Simpson's integration yields (Clark, 1992)

$$[\boldsymbol{D}]_{j,j'} = \begin{cases} \frac{1}{3(I-1)}, & j = j' = 1 \text{ or } j = j' = I \\ \frac{4}{3(I-1)}, & j = j' = 2, 4, ..., I-1 \\ \frac{2}{3(I-1)}, & j = j' = 3, 5, ..., I-2 \\ 0 & j \neq j' \end{cases} \tag{7.3.52}$$

and

$$[\phi_i]_j = \phi_i\left(-\frac{\Omega}{2} + \left(\frac{(j-1)}{I-1}\right)\cdot\frac{\Omega}{2}\right). \tag{7.3.53}$$

The eigensystem can now be solved by using standard methods.

7.3.8 Discrete delay profile

For solving the KL expansion for a discretised delay profile with J component bins, we have

$$P_{hP}(\tau) = \sum_{j=1}^{J} A_j\ \delta(\tau - \tau_j) \quad \Leftrightarrow \quad R_{HP}(\Delta\omega) = \sum_{j=1}^{J} A_j\ \mathrm{e}^{-\mathrm{j}\Delta\omega\tau_j} \tag{7.3.54}$$

and A_j are the mean square (averaged over distance) amplitudes of the effective scatterers and τ_j are the delays to the bins. Substitution of $R_{\mathrm{H}}(\Delta\omega)$ into the integral equations gives

$$C^*(\omega_1)\sum_{j=1}^{J} A_i\ \mathrm{e}^{\mathrm{j}\omega_1\tau_j}\int_{-\Omega/2}^{\Omega/2} C(\omega_2)\ \mathrm{e}^{-\mathrm{j}\omega_2\tau_j}\ \phi_i(\omega_2)\ \mathrm{d}\omega_2 = \lambda_i\phi_i(\omega_1) \tag{7.3.55}$$

and rearranging gives the eigenfunctions

$$\phi_i(\omega) = C^*(\omega)\sum_{i=1}^{J}\psi_{ij}\ A_j\ \mathrm{e}^{\mathrm{j}\omega\tau_j} \tag{7.3.56}$$

in terms of the unknown coefficients ψ_{ij}. Substituting back and equating for all values of ω_1, there results an eigensystem comprising I independent linear equations

$$\boldsymbol{Q}\ \psi_i = \lambda_i\ \psi_i \tag{7.3.57}$$

where

$$\left[\boldsymbol{Q}\right]_{j,j'} = A_j \int_{-\Omega/2}^{\Omega/2} |C(\omega)|^2 \, \mathrm{e}^{-\mathrm{j}\omega(\tau_j - \tau_{j'})} \, \mathrm{d}\omega. \tag{7.3.58}$$

The time domain response can be generated using the Fourier transform, as noted above.

For the case where the filter is brickwall, $C(\omega) = 1$, then

$$[\boldsymbol{Q}]_{j,j'} = A_j \, \mathrm{sinc}\left(\Omega(\tau_j - \tau_{j'})\right). \tag{7.3.59}$$

For the two-path model,

$$P_{hP}^{(2)}(\tau) = |h(\tau)|^2 = \frac{1}{1+A}\delta(t) + \frac{A}{1+A}\delta(t-\tau) \tag{7.3.60}$$

the eigenvectors are

$$\lambda_1 = \frac{1}{1+A}\left(1 + \mathrm{sinc}(\Omega s)\right), \quad \lambda_2 = \frac{A}{1+A}\left(1 - \mathrm{sinc}(\Omega s)\right), \tag{7.3.61}$$

where

$$s = \frac{\sqrt{A}\tau}{1+A} \tag{7.3.62}$$

is the delay spread.

In general, when the number of paths is small such as in this two-path example, the KL expansion is not necessary for the random generation of snapshots of the channel $H(\omega, z)$, and it is easier to use the time domain technique of directly generating independent, zero-mean Gaussian variables with $E\{|h_i|^2\} = A_i$, and then Fourier transforming.

7.4 Envelope–phase relations for fading channels

In dealing with the mobile channel and its statistical description for narrowband for short-term (no shadowing) fading, the envelope, r, and phase, ϕ, are independent in the statistical sense,

$$p_{r,\phi}(r, \phi) = p_r(r) \cdot p_\phi(\phi), \tag{7.4.1}$$

i.e. for the Rayleigh fading with uniform phase,

$$p_{r,\phi} = \frac{2r}{\langle r^2 \rangle} \, \mathrm{e}^{-r^2/\langle r^2 \rangle} \cdot \frac{1}{2\pi}. \tag{7.4.2}$$

However, while the distributions exhibit independence, the envelope and phase signals are closely related, and in special cases, derived functions of them are transform pairs. Similarly, the in-phase and quadrature components, while being statistically independent, can be dependent as signals. A review of the situation will augment

an understanding of the mobile channel, and for some situations the dependency is useful in measurement systems and even offers possibilities for real time processing for improved signal reception.

Much of this section, in particular Sections 7.4.2 and 7.4.3, is from Poletti and Vaughan (1990). The wideband case is taken from the electronic circuit network case where the transfer function is also a function of frequency. However, the situation can apply to the mobile narrowband channel as well, where the frequency variable is replaced with the time (or space) variable. The envelope, r, is also referred to as the magnitude in this section and the log magnitude $m = \log r$ is also used.

7.4.1 Frequency-selective channel (network theory)

In electronic networks, various properties exist on the basis of idealised components. The frequency selective channel has a basic model which is a tapped delay line; however, this is not as general as an electronic network. An electronic network model and the channel will have the same properties, as follows. The transfer function, $H(\omega)$, and the impulse response, $h(\tau)$, are Fourier transforms, denoted with $\Leftrightarrow$, i.e.

$$H(\omega) \quad \Leftrightarrow \quad h(\tau) \tag{7.4.3}$$

and the transfer function is factored into its real and imaginary parts, denoted

$$H(\omega) = H_{\mathrm{R}}(\omega) + \mathrm{j}H_{\mathrm{I}}(\omega). \tag{7.4.4}$$

A causal impulse response can be written as a sum of odd and even terms, both of which are non-causal, as

$$h(\tau) = h_{\mathrm{e}}(\tau) + h_{\mathrm{o}}(\tau), \tag{7.4.5}$$

where

$$h_{\mathrm{e}}(\tau) = \frac{1}{2}\left(h(\tau) + h^*(-\tau)\right) \quad \Leftrightarrow \quad H_{\mathrm{R}}(\omega) \tag{7.4.6}$$

$$h_{\mathrm{o}}(\tau) = \frac{1}{2}\left(h(\tau) - h^*(-\tau)\right) \quad \Leftrightarrow \quad \mathrm{j}H_{\mathrm{I}}(\omega). \tag{7.4.7}$$

The causality means that $h(\tau)$ is zero for $\tau < 0$; the relation between h_{o} and h_{e} can be rewritten as

$$h_{\mathrm{o}}(\tau) = h_{\mathrm{e}}(\tau)\ \mathrm{sgn}(\tau) \tag{7.4.8}$$

or

$$-\mathrm{j}h_{\mathrm{o}}(\tau) = -\mathrm{j}\ \mathrm{sgn}(\tau)\ h_{\mathrm{e}}(\tau) \quad \Leftrightarrow \quad H_{\mathrm{I}}(\omega) = \left(-\frac{2}{\omega}\right) * H_{\mathrm{R}}(\omega). \tag{7.4.9}$$

This relation is the Hilbert transform[4], denoted $\mathcal{H}\{\}$,

$$H_{\mathrm{I}}(\omega) = \mathcal{H}\ \{H_{\mathrm{R}}(\omega)\}\,. \tag{7.4.10}$$

Note that $H_{\mathrm{R}}(\omega) = -\mathcal{H}\ \{H_{\mathrm{I}}(\omega)\}$ or $H_{\mathrm{R}}(\omega) = \mathcal{H}\ \{\mathcal{H}\ \{\mathcal{H}\ \{H_{\mathrm{I}}(\omega)\}\}\}$.

If the transfer function is written as

$$\tilde{h}(t) \Leftrightarrow \log H(\omega) = \log|H(\omega)| + \mathrm{j}\mathrm{Phase}\,\{H(\omega)\} \tag{7.4.11}$$

then the log magnitude and phase terms are Hilbert pairs if $\tilde{h}(\tau)$ is causal. This condition corresponds to having all the poles and zeros of $H(\omega)$ inside the unit circle of the complex z-plane. Equivalently, from the Laplace transform, all poles and zeros must be on the left hand side of the complex s plane. The one-sided Laplace transform is

$$H(s) = \int_0^\infty h(\tau)\ \mathrm{e}^{-s\tau}\ \mathrm{d}t. \tag{7.4.12}$$

The poles of $H(s)$ are in the left-hand plane and the zeros can be anywhere. Papoulis (1962) shows that if $\log|H(s)|$ is analytic the right-hand plane (i.e. no zeros there) and $H(s/\mathrm{j})$ tends to zero as s tends to infinity, then $h(\tau)$ is causal. Thus

$$\mathrm{Phase}\,\{H(\omega)\} = \mathcal{H}\ \{\log|H(\omega)|\}\,, \quad (\tilde{h}(\tau)\ \text{causal, i.e.}\ H(\omega)\ \text{minimum phase}). \tag{7.4.13}$$

This is called the minimum phase case because a change in magnitude is associated with a minimum change of phase (Balabanian and Le Page, 1956).

Voelcker (1966) shows that the Hilbert relation always holds for the derivative of the log magnitude and the derivative of the phase and this relation arises in the following section.

7.4.2 Pole-zero description of mobile signals

This section (Poletti and Vaughan, 1990) gives the mathematical development for the envelope and phase of a mobile signal in terms of the pole-zero description. The signal notation $F(f) \Leftrightarrow f(t)$ is used, and $f(t)$ can represent the time (or space) varying signal received by a narrowband receiver moving through a static scenario. $F(f)$ is in general a two-sided function of frequency and its bilateral Laplace transform in the complex time plane, denoted $f(v)$, will be discussed.

The one-sided Laplace transform used above transforms a causal time signal into a complex frequency function, which can be specified by its poles and zeros. The two-sided Laplace transform of $F(f)$ is defined by

$$f(v) = \int_{-\infty}^{\infty} F(f)\mathrm{e}^{vf}\,\mathrm{d}f, \tag{7.4.14}$$

[4] The integral transform relations are

$$H_{\mathrm{I}}(\omega) = -\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{H_{\mathrm{R}}(\Omega)}{\omega-\Omega}\,\mathrm{d}\Omega, \quad H_{\mathrm{R}}(\omega) = +\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{H_{\mathrm{I}}(\Omega)}{\omega-\Omega}\,\mathrm{d}\Omega.$$

where $v = -\sigma + \mathrm{j}2\pi t$. The $f(v)$ can be viewed as a Fourier transform of $F(f)\mathrm{e}^{-\sigma f}$ with $f(t)$ a Fourier transform of $F(f)$ as noted above. These Fourier relations are written fully as

$$F(2\pi f)\,\mathrm{e}^{-j2\pi f\sigma} \Leftrightarrow f(v) = \frac{1}{2\pi}\int F(2\pi f)\,\mathrm{e}^{-2\pi f\sigma}\,\mathrm{e}^{\mathrm{j}2\pi f t}\,\mathrm{d}(2\pi f), \quad (7.4.15)$$

$$F(2\pi f) \Leftrightarrow f(t) = \frac{1}{2\pi}\int F(2\pi f)\mathrm{e}^{\mathrm{j}2\pi f t}\mathrm{d}(2\pi f). \quad (7.4.16)$$

The two-sided Laplace transform has a region of convergence that is a vertical strip $L < \sigma < R$, where L and R define the limits of the existence of the integral. If the region of convergence includes the imaginary axis, then the Fourier transform of $F(f)$ exists. If the region of convergence includes the whole plane, the function is entire, and there are no poles, apart from possible essential singularities at infinity. Voelcker (1966) has shown that any bandlimited function is bounded, and hence entire; and since the complex fading signal is bandlimited, it has no poles. However, for completeness, both poles and zeros are included. Poles can account for multiple scattering between physical scatterers and take the form of an infinite impule response (IIR) filter, but multiple scattering can be more easily accounted for in the effective scatterers model and this takes the form of a finite impulse response (FIR) filter.

The negative sign of the time term σ in the definition above means that a causal (right-sided) function $F_\mathrm{c}(f)$, with analytic time function $f_\mathrm{a}(t)$, will have poles in the left-hand half of the $(\sigma, \mathrm{j}2\pi t)$ plane, and an anticausal (left-sided) function $F_\mathrm{ac}(f)$, with 'antianalytic' time function $f_\mathrm{aa}(t)$, will have poles in the right-hand half-plane. The positive sign of t means that the time function $f(t)$ is traced out as the imaginary axis is ascended. Other values of sign could be chosen, but the choices above were made to produce the same results as in the network theory case (Voelcker, 1966).

In order to interpret the effect of the poles and zeros, $F(f)$ is written as the convolution of a causal and an anticausal function;

$$F(f) = F_\mathrm{c}(f) * F_\mathrm{ac}(f). \quad (7.4.17)$$

The Laplace transform, $f(v)$, is then

$$\begin{aligned} f(v) &= f_\mathrm{a}(v)\; f_\mathrm{aa}(v) \\ &= \int_0^\infty F_\mathrm{c}(f)\mathrm{e}^{vf}\mathrm{d}f \int_0^\infty F_\mathrm{ac}(-l)\mathrm{e}^{-vl}\mathrm{d}l, \end{aligned} \quad (7.4.18)$$

where the substitution $l = -f$ has been made in the second integral. This shows that the transform consists of the product of two one-sided transforms, with different signs in the exponent. The poles and zeros of $f_\mathrm{a}(v)$ and $f_\mathrm{aa}(v)$ are poles and zeros of $f(v)$. (It is assumed that there are no poles and zeros in $f_\mathrm{a}(v)$ and $f_\mathrm{aa}(v)$ that cancel when $f(v)$ is formed). The antianalytic function has poles and minimum phase zeros in the right-hand half-plane, whereas the analytic function has them in the left-hand half-plane. Each function may have non-minimum phase zeros in the other respective half-plane.

The magnitude of $f_a(t)$ at time $t = t_0$ is given by the distances from the zeros to t_0 divided by the pole distances. The phase is given by the sum of pole angles minus the sum of zero angles. This is the reverse of the network theory case, caused by the change in sign in the above definition for the Laplace transform.

The magnitude of $f_{aa}(t)$ is found as for $f_a(t)$ above, but its phase is given by the sum of zero angles minus the sum of pole angles. This is shown in Figure 7.4.1.

We can write $f(v)$ in terms of its poles and zeros as

$$f(v) = \frac{\prod_{n=1}^{N}(v - v_n)}{\prod_{p=1}^{P}(v - v_p)}. \tag{7.4.19}$$

The zeros may be assigned to $f_a(v)$ or $f_{aa}(v)$. There are thus a number of degenerate decompositions of $f(v)$. The magnitude is the same in all cases and the overall phase is the same up to a multiple of π. This form ignores any essential singularities present, such as linear phase terms, and ignores constant multiples of π in phase.

Along $\sigma = 0$, $f(t)$ can be written as

$$f(t) = \frac{\prod_{n=1}^{N}(\sigma_n + j2\pi\,(t - t_n))}{\prod_{p=1}^{P}(\sigma_p + j2\pi\,(t - t_p))}. \tag{7.4.20}$$

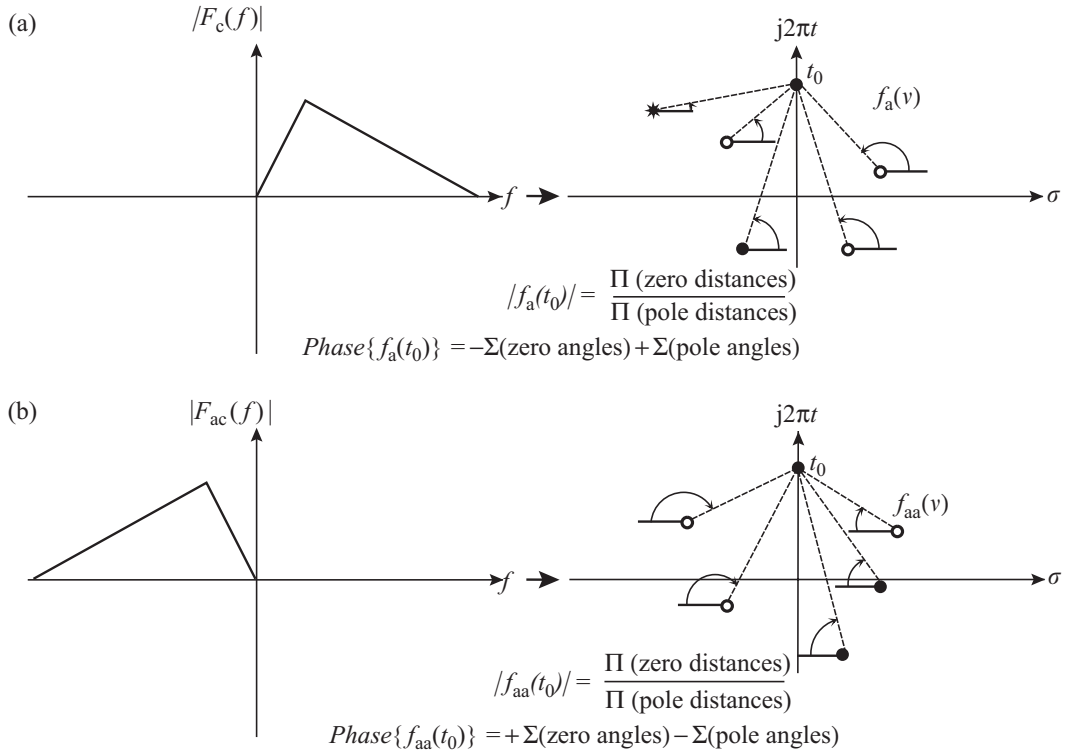


Figure 7.4.1 Causal (a) and anticausal (b) spectra and their poles and zeros (Poletti and Vaughan, 1990)

The log of $f(t)$ is thus

$$\log\left(f(t)\right) = \sum_{n=1}^{N} \log\left(\sigma_n + \mathrm{j}2\pi(t - t_n)\right) - \sum_{p=1}^{P} \log\left(\sigma_p + \mathrm{j}2\pi(t - t_p)\right), \quad (7.4.21)$$

and so the log derivative is

$$\frac{\mathrm{d}}{\mathrm{dt}}\left(\log(f(t))\right) = \sum_{n=1}^{N} \frac{\mathrm{j}2\pi}{\sigma_n + \mathrm{j}2\pi(t - t_n)} - \sum_{p=1}^{P} \frac{\mathrm{j}2\pi}{\sigma_p + \mathrm{j}2\pi(t - t_p)}. \quad (7.4.22)$$

The real and imaginary parts of these two equations are the log magnitude and phase, and log magnitude derivative and phase derivative, respectively.

For brevity, the notation is introduced

$$\begin{aligned} \log\left(f(t)\right) &= \log\left(|f(t)|\right) + \mathrm{jPhase}\left\{f(t)\right\} \\ &= m(t) + \mathrm{j}\phi(t). \end{aligned} \quad (7.4.23)$$

The phase of $f(t)$ may be written to include the nature of each zero as

$$\begin{aligned} \phi(t) = &\sum_{n=1}^{N} \left(k(n)\pi + \tan^{-1}\left(\frac{2\pi(t - t_n)}{\sigma_n}\right)\right) \\ &- \sum_{p=1}^{P} \tan^{-1}\left(\frac{2\pi(t - t_p)}{\sigma_p}\right), \end{aligned} \quad (7.4.24)$$

where $k(n) = 0$ or ± 1 and specifies whether the nth zero is minimum phase and which of the two time functions it belongs to. The phase derivative is thus

$$\phi'(t) = \sum_{n=1}^{N} \frac{2\pi\sigma_n}{\sigma_n^2 + 4\pi^2(t - t_n)^2} - \sum_{p=1}^{P} \frac{2\pi\sigma_p}{\sigma_p^2 + 4\pi^2(t - t_p)^2}. \quad (7.4.25)$$

Similarly the log magnitude is

$$\begin{aligned} m(t) = &\sum_{n=1}^{N} \frac{1}{2} \log\left(\sigma_n^2 + 4\pi^2(t - t_n)^2\right) \\ &- \sum_{p=1}^{P} \frac{1}{2} \log\left(\sigma_p^2 + 4\pi^2(t - t_p)^2\right), \end{aligned} \quad (7.4.26)$$

and the derivative of the log magnitude is

$$m'(t) = \sum_{n=1}^{N} \frac{4\pi^2(t - t_n)}{\sigma_n^2 + 4\pi^2(t - t_n)^2} - \sum_{p=1}^{P} \frac{4\pi^2(t - t_p)}{\sigma_p^2 + 4\pi^2(t - t_p)^2}. \quad (7.4.27)$$

Note that each term in the phase derivative expression has delta function characteristics at $t = t_n$ or $t = t_p$ as σ tends to zero. Similarly each term in the magnitude derivative expression is an approximation to $1/\pi t$. (Both approximations have spectra which are the exact spectra weighted by a term $\exp(-|\sigma_n| f)$ as mentioned above.)

The terms for each n and p in the phase derivative and the magnitude derivative are Hilbert transforms of each other, since each term in the $\log(f(t))$ expression has a causal transform. However, the sign of the relationship depends on which side of the imaginary axis the pole or zero is on. It is apparent, then, that if all poles and zeros of $f(v)$ are in the same half-plane, the individual phase terms will add, and a global Hilbert relationship will exist between $\phi'(t)$ and $m'(t)$, up to a constant. For poles and zeros in the left half-plane, this simply means that the two-sided spectrum, $F(f)$, is a causal function which has been shifted in a negative frequency direction to make it two-sided, i.e.

$$F(f) = F_{\mathrm{c}}(f) * \delta(f + f_0). \tag{7.4.28}$$

More complicated decompositions may be generated by assigning zeros to $f_{\mathrm{aa}}(t)$. In such a case $f_{\mathrm{a}}(t)$ is analytic and minimum phase and $f_{\mathrm{aa}}(t)$ is bandlimited and maximum phase.

For all poles and zeros in the right-hand half-plane the spectrum is an anticausal function which has been shifted right in frequency,

$$F(f) = \delta(f - f_0) * F_{\mathrm{ac}}(f). \tag{7.4.29}$$

Again, some zeros may be reassigned to the analytic function, so that $f_{\mathrm{a}}(t)$ is maximum phase and bandlimited (Papoulis, 1962). It must be remembered that a bilateral Hilbert signal has an essential singularity which will contribute an additional linear phase term to $f(t)$. This produces an additional constant term in the phase derivative. When only short-term (a fade) fluctuations in phase and envelope are of importance, the constant term is not so important.

For a general bilateral signal without the properties stated above there is no global Hilbert relationship between log magnitude and phase derivatives. However, it is apparent from the expressions for $m'(t)$ and $\phi'(t)$ that the effect of poles and zeros at large distances from $t = t_0$ is negligible. These expressions may then be truncated to include only local poles and zeros. If all poles and zeros near the current point in time are on one side of the time axis, then locally, an approximate Hilbert relation will hold. This can be useful for signal processing in order to reduce the effects of the deep fades.

7.4.3 Zero description of fading signal

The zeros of the fading signal are considered in the $(\sigma, 2\pi Vt)$ plane, where the time axis has been scaled by a factor V, the speed of the mobile receiver. This means that the imaginary axis is equivalent to distance travelled, and for a static spatial field the zeros will be at fixed positions, $(\sigma_n, 2\pi d_n)$, giving the same zero pattern for all speeds.

The formulas for the phase derivative and the log magnitude derivative can be modified to include the speed of the moving terminal, V,

$$\phi'(t) = \sum_{n=1}^{N} \frac{2\pi V \sigma_n}{\sigma_n^2 + 4\pi^2 V^2 (t - t_n)^2}, \tag{7.4.30}$$

$$m'(t) = \sum_{n=1}^{N} \frac{4\pi^2 V^2 (t - t_n)}{\sigma_n^2 + 4\pi^2 V^2 (t - t_n)^2}. \tag{7.4.31}$$

These show how the peak instantaneous frequency and the shape of the magnitude function depend on the velocity.

Deep fades may be associated with the presence of one or more zeros near the time axis. A 'first-order fade' is caused by a single dominant zero, and the phase and magnitude derivatives are always approximately Hilbert related, up to an ambiguity in sign. A 'second-order fade' is from two dominant zeros. If the two zeros are on the same side, the approximate Hilbert relations hold, again up to an ambiguity in sign. If the zeros are on opposite sides, their phase contributions tend to cancel. If the zeros are completely symmetric, the fading signal is locally constant in phase.

7.4.4 Application to single variable modulation correction in narrowband systems

This section is from Poletti and Vaughan (1990), and a similar technique is described by Treichler and Agee (1983).

7.4.4.1 Magnitude correction

In general, the mobile channel is not minimum phase. If a constant phase (AM) radio signal is transmitted, the received phase will be modulated by the multipath channel. The phase derivative of the radio signal may then be measured to indicate the phase derivative of the fading envelope. If the Hilbert transform of the phase derivative is taken, those portions where there are 'Hilbert fades' will correspond closely to the derivative of the log magnitude. However, the sign of the relationship will be ambiguous. The magnitude of the signal is always negative going approaching a fade, but the phase derivative may be negative or positive going. This is not a problem in the AM case. The absolute value of the Hilbert transform of the phase derivative is always positive going. When added to the log magnitude of the received signal it will always oppose the drop in the received signal magnitude.

A simulation demonstrating the approach used the following technique (Poletti and Vaughan, 1990). The phase derivative is Hilbert transformed and its absolute value taken. Small phase variations are suppressed, and the resulting correction signal is scaled using a running estimator and applied to the distorted magnitude. Much of the short-term, Rayleigh-like fading can be removed, leaving the longer-term variations.

Typically AM is not used at high frequencies where multipath effects are significant. Hence the magnitude correction technique has limited applicability for mobile communications.

7.4.4.2 Phase correction

If a constant envelope modulation is transmitted, the envelope of the received signal will represent the magnitude of the channel gain. The derivative of the log magnitude of the fading signal is normally straightforward to measure from the RSSI. The Hilbert transform of this yields an unsigned phase derivative correction signal. The sign must be derived by sensing the phase derivative of the received signal at the time of the fade, and detecting whether it has become more positive or more negative. This is easier to do if the fade is deep, because in such a case the FM spike caused by the fade may be large compared to the baseband signal excursion. For shallow fades it may not be possible to detect the direction of the phase derivative excursion, but this is not serious because the main objective is to correct the signal at the deeper fades. (Recall that in digital communications it is the deep fades that are responsible for the so-called irreducible bit error rate.)

The correction of the baseband signal involves the following:

- Detection of a deep fade by observing the magnitude drop below a certain threshold (or observing the derived correction signal rise above a certain threshold).
- Observing the received baseband signal and sensing whether the phase derivative is positive, negative, or negligible (in which case no correction is applied).
- Correction of the baseband signal.

A simulation is easy to produce for demonstrating the promise of such a system (Poletti and Vaughan, 1990). Treichler and Agee (1983) introduced an adaptive LMS algorithm for correcting the phase of constant modulus modulations corrupted by multipath although their system does not directly use the Hilbert relations. The above systems are types of blind correction methods because no sounding pulse (e.g. a training sequence) is required to characterise the channel.

Finally, it is possible to recover the channel transfer function by using special communications signal properties (Tong *et al*., 1994). The constant modulus signals themselves are not used, so the method is still referred to as blind, but the higher-order statistics of the signals, which are arranged to be periodic (cyclostationary), are used. The periodic correlation of the communications signals is effectively the training sequence in such a scheme. Again, this scheme does not use the Hilbert properties of the channel.

7.4.5 Channel characterisation using the Hilbert relations

Donaldson *et al*. (1996) took indoor measurements in the frequency domain and applied the Hilbert transform relations for channel characterisation, in particular to ascertain the extent to which the channel was minimum phase. Their technique is to measure $H_{\mathrm{I}}(\omega)$ and $H_{\mathrm{R}}(\omega)$, and compare, by correlation coefficient, the measured $H_{\mathrm{I}}(\omega)$ with that calculated by Hilbert transform of $H_{\mathrm{R}}(\omega)$. Similarly, a comparison of the phase as found from the Hilbert transform of the magnitude, and the measured phase, can be made. Favourable comparisons indicate a mainly minimum phase channel. The degree of minimum phase behaviour is also taken as the proportion of zeros

inside the unit circle in the z-domain, or equivalently on one side of the complex plane. For indoor channels, where there is often a line-of-sight path which is dominant, the channels are often minimum phase. Donaldson *et al*. also note that the minimum phase channel results in the channel having the lowest delay spread,[5] and suggest that the maximum phase case should provide an upper limit on the delay spread. The ramification is that a magnitude-only measurement can be used to find upper and lower limits for the delay spread. This is demonstrated below using the two-path model.

7.4.6 Delay spread limits from magnitude-only frequency transfer function

In taking a measurement of a channel transfer function, the complex value is often difficult or expensive to obtain, and typically, only the magnitude (from the power) is obtained. If there are N distinct (complex) zeros in the magnitude function, then since the phase change can be in either direction for each zero, there are 2^N possible transfer functions. One of these is minimum phase (all the individual fades have an associated minimum phase change), one is maximum phase (all the fades have a maximum phase change because each zero is on the right-hand side of the axis), and the others are called non-minimum phase. The two-path model for a static impulse response,

$$h^{(2)}(\tau) \propto \delta(\tau) + a_2\delta(\tau - \tau_2) \Leftrightarrow H^{(2)}(\omega) = 1 + a_2 \mathrm{e}^{-\mathrm{j}\omega\tau_2} \tag{7.4.32}$$

is minimum phase when $|a_2| < 1$, and non-minimum phase when $|a_2| > 1$, as discussed in Chapter 5. The zeros are evenly spaced along (or rather beside) the frequency axis. For a given τ_2, the delay spread is a maximum when $a_2 = 1$, and for this case the zeros of $H(\omega)$ are real, i.e. they are on the axis between the left- and right-hand planes of $H(s)$. There is an infinite number of zeros in $H^{(2)}(s)$ and all of these affect the local behaviour of the channel, although for deep fades the corresponding (closest) zero dominates as discussed above. A particular zero at $s = s_z$ contributes the following behaviour to the transfer function. Putting $s = \alpha + \mathrm{j}\omega$, the single zero transfer function is, from the previous section,

$$H(s) = s - s_z = (\alpha - \alpha_z) + \mathrm{j}(\omega - \omega_z). \tag{7.4.33}$$

It is emphasised here that this approximation is only good near a deep fade. The magnitude and phase are

$$|H(s)|^2 = (\alpha - \alpha_z)^2 + (\omega - \omega_z)^2 \tag{7.4.34}$$

$$\text{Phase}\,\{H(s)\} = \pi + \tan^{-1}\left(\frac{\omega - \omega_z}{\alpha_z}\right) \tag{7.4.35}$$

[5] Based on minimum and maximum delay sequences as properties of minimum and maximum phase transfer functions respectively (A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall, 1975).

so that as the mobile terminal moves past the zero (from position $\omega \rightarrow -\infty$ to $\beta \rightarrow \infty$), the phase of the transfer function moves from $-\pi/2$ to $\pi/2$. The phase direction is the opposite for the zero being on the other side of the plane. The zero can be flipped to the other plane by the all-pass operation, with $\alpha_z = 1 - |a_2|$,

$$H(\omega) \leftarrow H(\omega)\mathrm{e}^{-\mathrm{j}\left(\pi+\tan^{-1}((\omega-\omega_z)/\alpha_z)\right)}. \tag{7.4.36}$$

This operation does not affect the magnitude of the transfer function, just the phase. The effect of flipping individual zeros of the transfer function, i.e. changing individual fades from a minimum phase to maximum phase character, can now be estimated. Zero flipping is discussed by Voelcker (1966), Al-Jalili (1991) and Poletti (1997).

A finite bandwidth of the two-path transfer function, with $a_2 = 0.8$ (see Chapter 5), is considered. The following examples were developed with assistance from M. Leonard-Taylor (Leonard-Taylor, 1998). The magnitude, phase, complex plane trajectory, and impulse response (found by Fourier transform of the transfer function), limited to a dynamic range of 50 dB, is given in Figure 7.4.2. The transfer function has zeros extending to outside of the band shown, and the effect of these zeros is not included here. The main result of this is inaccurate phase representation over the finite band, and the corresponding inaccuracies in the derived impulse response and delay spread. The two-path delay spread, from Chapter 5, is $a^2\tau/(1 + a^2) = 0.4878\tau_2$. The situation is worse for smaller a_2, because the local zeros are less dominant in the local transfer function since they are further from the axis (α_z is larger).

The fully minimum phase character is now disturbed by converting the centre-most zero to be a 'maximum phase' fade, and the corresponding quantities are given in Figure 7.4.3. The magnitude is the same, the complex trajectory changes drastically, and the impulse response changes very slightly, with a corresponding very slight change in delay spread, again calculated over the 50 dB dynamic range.

By flipping all the zeros, the transfer function becomes maximum phase. This is given in Figure 7.4.4. The impulse response now features dominantly the time-reversed two-path model, $h^{(2)}(\tau) = a_2\delta(\tau) + \delta(\tau - \tau_2)$, as expected, and the delay spread is $0.492\tau_2$, which should ideally be the same as for the minimum phase case. But there are also some low level extra discrete paths at multiples of τ_2. These become much larger for deceasing a_2 values, which is when the approximation for the individual zero contribution is getting worse. For fewer zeros flipped, the skirts of the impulse response are higher and the discrete-like contributions are smaller.

The situation can be compared to calculating the delay spread directly from the time domain for the two-path model. The delay spread for the two-path model is a maximum, for a given τ_2, when both paths have the same energy, i.e. when $|a_2| = 1$. The minimum delay spread occurs when the ratio of amplitudes, a_2, is minimised for the minimum phase case, or maximised, for the maximum phase case. Instead of changing a_2 in the time domain in this way, we are here altering the frequency domain phase and the result is a many-path model which has inaccuracies caused by approximating an infinite zero transfer function with a finite zero model. As the transfer function goes from minimum phase to maximum phase, the delay spread reaches a peak (cf. when $a_2 = 1$) and reduces again. The inaccuracy becomes greatest

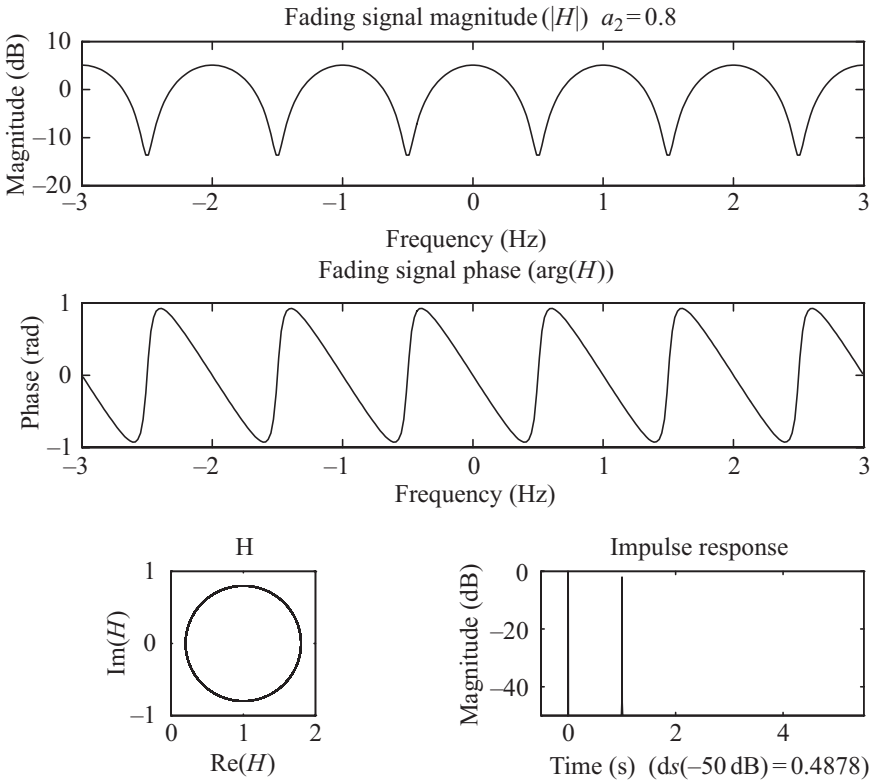


Figure 7.4.2 The log magnitude, phase, complex trajectory and impulse response for the minimum phase two-path model with $a_2 = 0.8$. The function is periodic and is minimum phase. The delay spread is $0.4878\tau_2$

when the maximum phase case is reached and there results a many-path model, which ideally reduces exactly to the maximum phase two-path model with $h^{(2)}(\tau) = a_2\delta(\tau) + \delta(\tau - \tau_2)$. Nevertheless, the behavioural characteristics of the channel are clarified by the approach. To summarise, we start with a minimum phase two-path model, and by flipping successive individual zeros so that successive individual fades become maximum phase, the model gradually loses its two-path, and even its discrete path, nature. When all the zeros have been flipped, we have returned (except for errors in the approximation) to the two-path model, which is now time-reversed and is maximum phase. The minimum phase channel corresponds to minimum delay spread, as suggested by the concept of minimum sequences. The maximum phase channel, corresponding to the maximum sequence, also has minimum delay spread because the delay spread metric does not distinguish between the 'direction' of the sequence, i.e. the delay profile. The maximum delay spread occurs between the minimum phase and maximum phase case, i.e. when 50 per cent of the fades are minimum. This type of signal is found from the Clarke scenario, for example.

7.4.7 Angular distribution from spatial transfer function

In the spatial domain, a channel $H(z)$ (z here denotes distance) comprises contributions from scatterers from various directions $u_i (= k_C \cos\theta_i)$ given by a distribution,

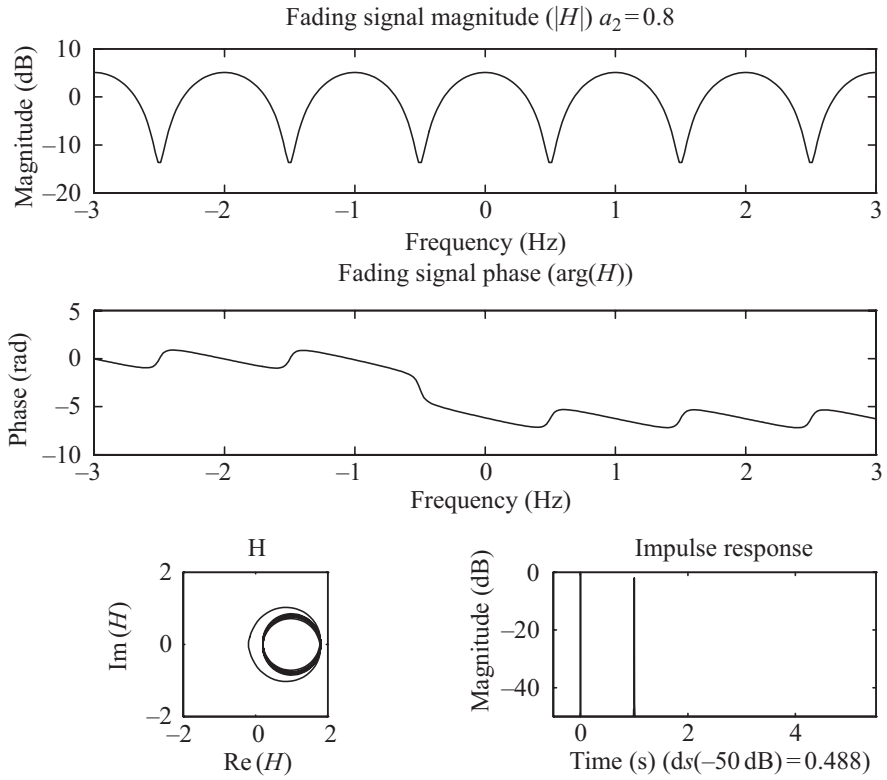


Figure 7.4.3 *The log magnitude, phase, complex trajectory and impulse response for the two-path model with $a_2 = 0.8$, but with one fade in the band forced to be 'maximum phase'. The magnitude of the transfer function is unaltered*

$h(u) \Leftrightarrow H(z)$ (see Chapter 2), where the notation is $a(u) \Leftrightarrow h(z)$. There are no poles in the spatial transfer function and so only zeros need be considered in analysis, as discussed above. Such a channel, if generated by simulation from a discretised Clarke scenario can be expected to be '50 per cent minimum phase', i.e. on average, half of the zeros will be in the left-hand plane and half in the right-hand plane. If a single, dominant effective source is included in the scenario (which now becomes directional and the channel becomes Rician with a large Rice k-factor) then the channel will be 100 per cent minimum phase if k_{Rice} is large enough. Only in this situation can the Hilbert relations be applied to the whole spatial channel (in the region of static scatterers). The directions and amplitudes of the effective sources, $h(u)$, can be found by Fourier transform of $H(z)$, and thence the angular spread, etc., in the same way as for the delay spread of the frequency transfer function case. When the phases of $H(z)$ are altered in the same way as above, the transform, giving the angular power distribution, becomes smeared and takes on a continuous rather than discrete character, as with the skirts of the impulse responses above. However, envelope–phase relations can still hold, even in the Clarke scenario (50 per cent minimum phase).

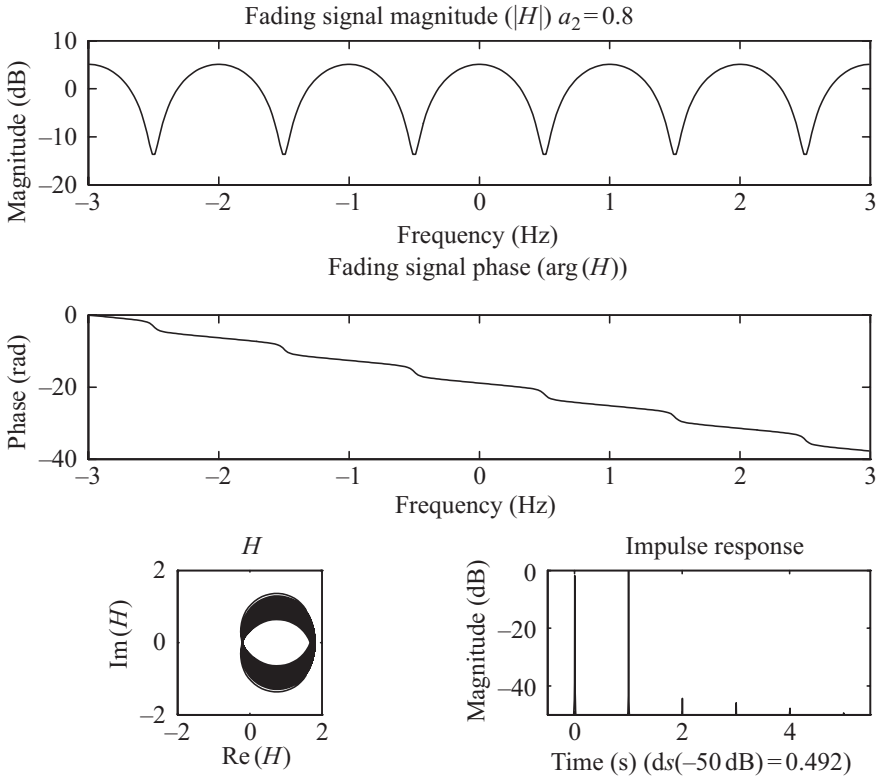


Figure 7.4.4 The two-path model with all fades in the band forced to be maximum phase. The dominant part of the impulse response is the time-reversed form of the original minimum phase two-path model, and there are also small contributions at later values of τ_2

7.4.8 Suppressing the effect of the phase ambiguity in Hilbert relations

It is noted above that the direction of the phase excursion calculated from the magnitude change is ambiguous in sign. This sign ambiguity can be suppressed by considering either the square of the phase or the magnitude of the phase. For example, the correlation coefficient between the Hilbert transform of the log magnitude derivative and the absolute phase derivative, calculated over a long record length of several tens of fades, reveals nearly perfect likeness between these quantities, i.e. (the correlation coefficient calculation is denoted $\odot$),

$$\rho_{m'|\phi'|} = \mathcal{H}\left\{\frac{m'(t)}{m(t)}\right\} \odot \left\{|\phi'(t)|\right\} \approx 1. \tag{7.4.37}$$

The correlation coefficient function, with correlation displacement t_d,

$$\rho_{m'|\phi'|}(t_\mathrm{d}) = \mathcal{H}\left\{\frac{m'(t)}{m(t)}\right\} \odot \left\{|\phi'(t+t_\mathrm{d})|\right\}, \tag{7.4.38}$$

drops to essentially zero quickly (by $Vt_\mathrm{d} \approx 0.1$ wavelength) in a 50 per cent minimum phase signal. If the channel is Rician and from a purely minimum phase (from a

sufficiently dominant source), the absolute sign for the phase becomes unnecessary and the correlation coefficient is always unity. The lobe of the correlation coefficient function broadens with increasing Rice k-factor.

Finally, following the frequency domain technique of the previous section, by correlating a measured imaginary part with the imaginary part calculated by Hilbert transform of the real part (or the magnitude and measured/calculated phase), the extent to which the spatial channel is minimum phase can be established.

7.4.9 Summary

The envelope and phase signals in mobile channels feature special dependence, despite their independence in the statistical sense. The development here follows that of the network theory case, which is directly applicable to frequency selective channels. A pole zero description produces the requirements for a Hilbert relation between the derivatives of the log magnitude and phase. The fast fading channel of mobile communications does not, in general, meet these requirements, but some fades – those with all zeros on the same side of the time axis – have Hilbert relations. The relations offer potential for reducing the detrimental effects of deep fades by extra signal processing at a receiver and can be applied to channel characterisation.

7.5 Testing the multipath model – the prediction of short-term fading

7.5.1 Introduction

The model for the short-term channel behaviour featuring the Rayleigh-like fading of the envelope comprises a set of effective scatterers whose discrete contributions are summed. Recall that the model is written in terms of the baseband equivalent transfer function which depends on position z and baseband frequency ω

$$H(z, \omega) = \sum_{i=1}^{N} a_i \mathrm{e}^{\mathrm{j}u_i z} \mathrm{e}^{-\mathrm{j}\omega\tau_i}. \tag{7.5.1}$$

In reality, the distribution is continuous as discussed earlier, except for the simplest of situations, and the discrete model comprising effective point sources (scatterers) can only be tested for accuracy by using real-world experimental results.

The modelled scatterers are constant over a neighbourhood of the receiver position. They are characterised by their number, N, complex amplitudes, a_i, their directions given by the spatial Doppler frequencies, $u_i = k_C \cos\theta_i$, and delay times, τ_i. The size of the neighbourhood in space and frequency over which the effective scatterers are constant is not known from a snapshot receiver signal, but in the development of the above model in Chapter 2, it was noted that if the physical scatterers are in the far-field of a linear locus of the mobile then for narrow relative bandwidths, the effective scatterers will be constant. The proposition seems reasonable for at least a few wavelengths at microwave frequencies. In dealing with effective scatterers, recall

that the effect of the antenna is included. If the antenna is rotated, then the effective scatterers change accordingly. If the pattern is omnidirectional, however, then a rotation has no effect. With vehicular antennas, linear motion dominates rotation, but with personal systems, in particular handheld terminals, rotation may, at times, dominate the motion. At frequencies as high as tens of GHz, typical rotations of handsets can be linearised over distances of several wavelengths, and the model of equation (7.5.1) becomes valid over this range. This concept can be put to a test as follows.

7.5.2 Using the propagation model for signal prediction

The scattering model, with known static scatterers, gives the intriguing possibility of knowing the channel transfer function (in the neighbourhood) before the mobile has reached that position (Vaughan and Andersen, 1995; Eyceoz *et al*., 1998; Hwang and Winters, 1998; Andersen *et al*., 1999; Duel-Hallen *et al*., 2000; Vaughan *et al.*, 2000). How this could be done is demonstrated by addressing only the narrowband (i.e. flat-fading) channel, i.e.

$$H(z) = \sum_{i=1}^{N} a_i \mathrm{e}^{\mathrm{j}u_i z}. \tag{7.5.2}$$

The re-inclusion of frequency does not alter the prediction concept. Firstly the scatterers must be characterised. Then scatterers can be tracked, in principle, since they are (by assumption) slowly changing with respect to the mobile position. The concept is illustrated in Figure 7.5.1. Two steps of processing are required for the prediction. The first step is to get an estimate of the effective scatterers and 'lock on' to the correct solution, and the second step is to track the relatively slow-changing effective scatterers. Tracking algorithms are not used here; the subspace algorithms applied only use the past samples of the channel.

How well the signal can be predicted serves to check the discrete sources model of the source continuum. Real-world signals are needed for such a check. Successful prediction of real-world signals also puts the signal processing to the test, but the processing limitations can be first characterised by using computer simulation. Once the performance of the prediction is known for the synthetic, discrete source case, real-world signals can be tested and this test indicates the realism (or otherwise) of the source discretisation in the model. Results using just three sources, imaged from a synthetic circular array, indicate some promise for estimating the signal envelope indoors (Raich and Vaughan, 1999).

If the prediction of the short-term channel behaviour is possible, then it would serve more than to confirm the propagation model. The use of feedforward diversity (space or time) combiners would be possible, rather than the use of feedback combiners. Feedback combiners have the disadvantage of having a time delay between the weights solution for the 'out-of-date' channel and the optimum solution for the current channel. For fast fading channels, i.e. those with higher carrier frequencies and/or quickly moving terminals, the fading rate limits the channel capacity because of this latency in the feedback combination process. Successful prediction allows,

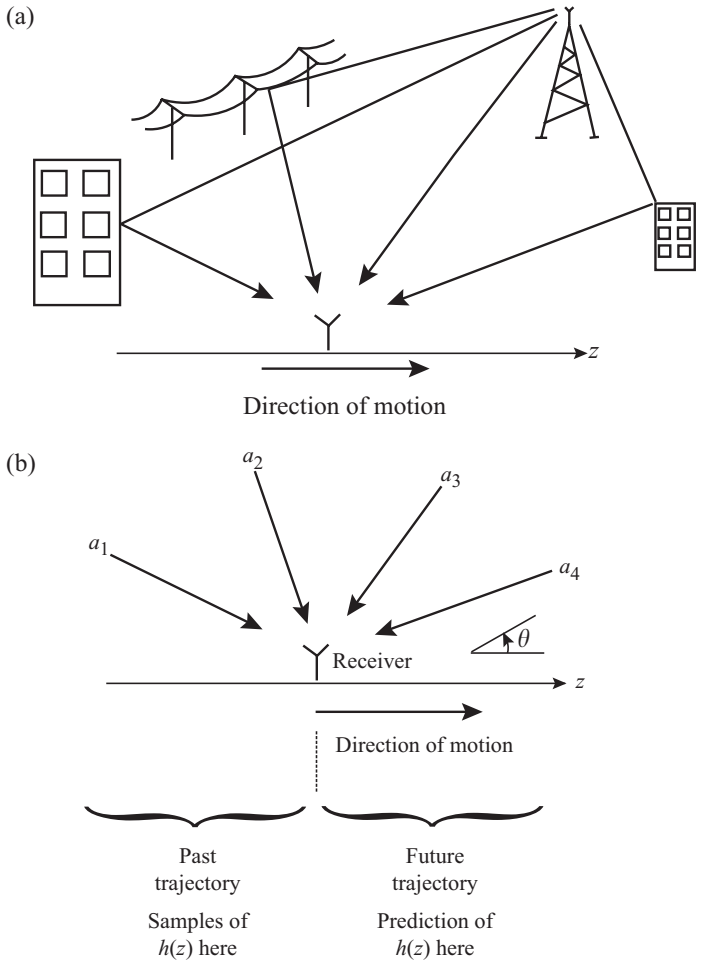


Figure 7.5.1 *(a) The physical scattering situation and (b) the effective scatterers with a 'past' trajectory over which the effective scatterers are characterised and a 'future' trajectory over which the channel can be predicted from the channel model. A comparison of the predicted channel and the measured channel tests the channel model and its formulation, including scattering and propagation assumptions, and the numerical solution of the nonlinear simultaneous equations which solve for the effective scatterers*

in principle, a way to avert this fading rate limitation. This could take the form of a change of time-slot in a TDMA system (avoiding time-slots occupying predicted fadetimes/locations, or seeking time-slots at predicted high SNR times/locations); a change of frequency in FDMA (avoiding frequency bins which occupy a predicted frequency fade location); a change of transmitted power negotiated by the receiver, or a negotiated change of weights in transmit diversity.

In the remainder of this section, the theory and signal processing for signal prediction (see also Rosenblatt,1960) are developed. Numerical experiments from subspace techniques characterise the signal processing limits and finally some real-world experimental results are presented.

7.5.3 Noise-free case: deterministic method

The number of effective scatterers, N, is unknown, but some limit for N can be estimated. Specifically, for a given mobile trajectory over which the channel is sampled, N is limited by angular resolvability criteria that can define the bins of the u_i.

To calculate the effective scatterers, samples are taken of the channel as a function of position, as the mobile moves. In practice a position sensor is therefore required, or else the samples are time-based and we make assumptions about the velocity of the receiver, $V = z/t$. In fact, the velocity need not be known, as long as it is known to be constant, in which case known regular time samples are sufficient.

The samples can be written as

$$\begin{aligned} h_1 &= h(z_1) = a_1 \mathrm{e}^{\mathrm{j}u_1 z_1} + a_2 \mathrm{e}^{\mathrm{j}u_2 z_1} \cdots + a_N \mathrm{e}^{\mathrm{j}u_N z_1} \\ h_2 &= h(z_2) = a_1 \mathrm{e}^{\mathrm{j}u_1 z_2} + a_2 \mathrm{e}^{\mathrm{j}u_2 z_2} \cdots + a_N \mathrm{e}^{\mathrm{j}u_N z_2} \\ &\vdots \\ h_M &= h(z_M) = a_1 \mathrm{e}^{\mathrm{j}u_1 z_M} + a_2 \mathrm{e}^{\mathrm{j}u_2 z_M} \cdots + a_N \mathrm{e}^{\mathrm{j}u_N z_M} \end{aligned} \tag{7.5.3}$$

in which M complex samples are taken to characterise N effective scatterers. There are $3N$ real unknowns (each of the u_i, $\mathrm{Re}\{a_i\}$ and $\mathrm{Im}\{a_i\}$) and for each equation there are two real knowns ($\mathrm{Re}\{h_i\}$, $\mathrm{Im}\{h_i\}$). The number of independent equations required for solution is $M \geq 3N/2$. If the samples are truly uncorrelated, then $M = 3N/2$ is sufficient. For correlated samples in the presence of noise, more sophisticated methods are required which produce a solution which is optimal in some sense. The solution technique is not pursued further here, except to recall that the system of equations comprises the Fourier transform where the Fourier variables are u and z (see Chapter 2). Instead of regularly spaced bins in u, however, as in the usual FFT or DFT algorithms, it is the spacings between the effective scatterers (the values of u_i with the range $[-k_\mathrm{C}, k_\mathrm{C}]$) in $a(u)$ which contain much of the information which defines the unbounded $h(z)$. The evenly spaced bins, here in u, of the usual FFT also give a well-defined periodicity in the transform, i.e. in the spatial domain, namely the spatial function has periodicity $1/\Delta u$, where Δu is the size of each of the bins. When the bins are unequal, the period increases. This allows prediction of the spatial signal which is not simply a replica of the spatial samples used to establish the N effective scatterers.

7.5.3.1 Partitioning of u for equal Doppler contributions

The trajectory of the moving mobile while it samples the channel gives a guide to the angular resolution to the sources. If the mobile moves a linear distance of L metres, then the Fourier resolution is recalled as

$$\begin{aligned} \Delta\theta &\approx \lambda/L, \qquad \text{broadside } (\theta = \pi/2), \\ &\approx \sqrt{2\lambda/L}, \qquad \text{endfire } (\theta = 0). \end{aligned} \tag{7.5.4}$$

The broadside resolution is the finer and this angle size can be used as a simple guide for defining bins for all the u_i. For example, a 5λ trajectory for an antenna means

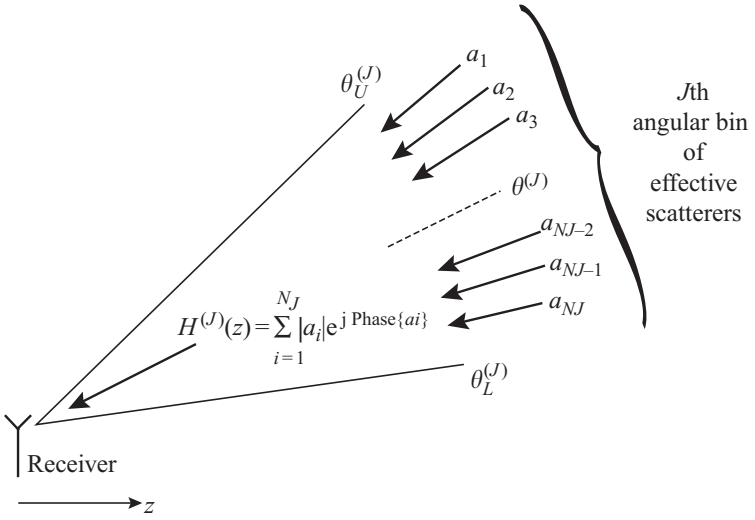


Figure 7.5.2 The J th angular partition of effective scatterers, defined by some angular resolution criteria from the finite trajectory length of the mobile receiver

that sources can be resolved (to the Fourier resolution) to about 11° at broadside and about 37° at endfire. So an initial partitioning guide would be to assume uniformly spaced effective sources spaced by 5°.

Other partitioning options are available, however. These are of interest because it is desirable to minimise the number of bins since this minimises the number of equations and solution time for the effective scatterers from the equation set (7.5.3). The contribution to the channel from the Jth bin has mean direction $u^{(J)}$ and it contains N_J effective scatterers, as shown in Figure 7.5.2.

The receiver is assumed to move a distance $\Delta z = z_2 - z_1$, which is sufficiently small so that the contribution from the bin of effective scatterers does not change in amplitude (but does change in phase) as the receiver moves Δz. This difference in the contribution to the channel transfer function at the adjacent positions, for the Jth bin, which contains N_J effective scatterers, is written as

$$h^{(J)}(z_2) - h^{(J)}(z_1) = \sum_{i=1}^{N_J} a_i e^{ju_i z_2} - \sum_{i=1}^{N_J} a_i e^{ju_i z_1}, \tag{7.5.5}$$

where the two terms are now phasors since they have a common amplitude. The effect of any directionality in the pattern is included here because the a_i include the pattern effect. Denoting the phase difference between the two phasors as $\Psi^{(J)}$, we have

$$\Psi^{(J)} = u^{(J)} \Delta z = k_C \cos(\theta^{(J)}) \Delta z, \tag{7.5.6}$$

where $\theta^{(J)}$ is the mean angle to the Jth resolvable bin. If the amplitude of the phasors is $a^{(J)}$, then the scalar difference between the phasors is (Nørklit and Vaughan, 1999)

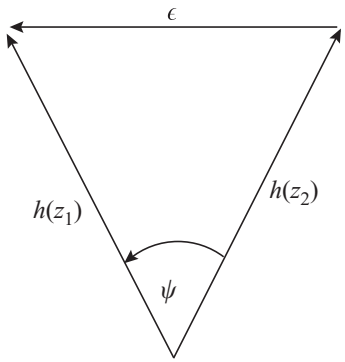


Figure 7.5.3 The phasor contributions to the transfer function from a single bin of effective scatterers at different spatial positions of the moving receiver

$$\epsilon^{(J)} = \left| h^{(J)}(z_2) - h^{(J)}(z_1) \right| = 2a^{(J)} \sin\left(\frac{\Psi^{(J)}}{2}\right), \tag{7.5.7}$$

as indicated in Figure 7.5.3.

Assuming that there are many effective sources of similar amplitude in each bin ($N_J > \sim 10$) then the $\epsilon^{(J)}$ are approximately Rayleigh distributed.

If the Jth bin is defined as occupying the range $[\theta_{\mathrm{L}}^{(\mathrm{J})}, \theta_{\mathrm{U}}^{(\mathrm{J})}]$, then the expected value of the energy in the change in the transfer function caused by the contribution from the single bin (the 'Jth' notation is dropped for now) of effective scatterers is

$$\begin{aligned} E\{\epsilon^2\} &= E\left\{\left(2\, a \sin\left(\frac{\Psi}{2}\right)\right)^2\right\} \\ &= 4\; E\left\{a^2\right\} \cdot E\left\{\sin^2\left(\frac{\Psi}{2}\right)\right\} \\ &= 4 \int_{\theta_{\mathrm{L}}}^{\theta_{\mathrm{U}}} p_\theta(\theta)\mathrm{d}\theta \cdot \int_{\theta_{\mathrm{L}}}^{\theta_{\mathrm{U}}} p_\theta(\theta) \sin^2\left(\frac{1}{2} k_{\mathrm{C}} \Delta z \cos\theta\right)\, \mathrm{d}\theta. \end{aligned} \tag{7.5.8}$$

The pdf for the incident power in a uniform scenario (including an omnidirectional antenna) of support $\theta = [0, \pi]$, is $p_\theta(\theta) = 1/\pi$, and substituting this gives

$$E\{\epsilon^2\} = \frac{4\,(\theta_{\mathrm{U}} - \theta_{\mathrm{L}})}{\pi^2} \int_{\theta_{\mathrm{L}}}^{\theta_{\mathrm{U}}} \sin^2\left(\frac{1}{2} k_{\mathrm{C}} \Delta z \cos\theta\right)\, \mathrm{d}\theta. \tag{7.5.9}$$

Note that $(\theta_{\mathrm{U}} - \theta_{\mathrm{L}})/\pi = N_j^2$ is simply the power of the contribution from the bin, i.e. proportional to the power of the (uniform) effective sources in the bin. This integral has no closed form solution. However, if we use the small angle approximation $\sin x \approx x$,

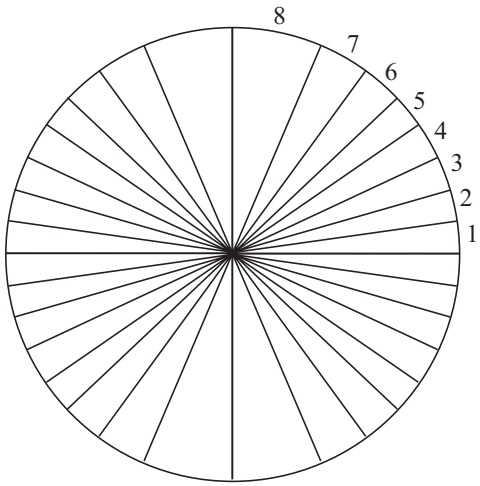


Figure 7.5.4 The partitioning in polar coordinates for the example of eight bins in a uniform, full-circle scenario

then the equation simplifies to

$$E\{\epsilon^2\} \approx \frac{1}{2}\left(\frac{k_C \Delta z}{\pi}\right)^2 (\theta_U - \theta_L)\left((\theta_U - \theta_L) + \cos(\theta_U + \theta_L)\sin(\theta_U - \theta_L)\right). \tag{7.5.10}$$

Numerical validation of this approximation for the partitioning (below) indicates that it holds for $\Delta z/\lambda_C < 0.3$. Equating this energy for each bin results in the equations

$$\begin{aligned}
&\Big(\theta_1 - \theta_0\Big)\Big(\theta_1 - \theta_0 + \cos(\theta_1 + \theta_0)\sin(\theta_1 - \theta_0)\Big) \\
&= \Big(\theta_2 - \theta_1\Big)\Big(\theta_2 - \theta_1 + \cos(\theta_2 + \theta_1)\sin(\theta_2 - \theta_1)\Big) \\
&= \Big(\theta_3 - \theta_2\Big)\Big(\theta_3 - \theta_2 + \cos(\theta_3 + \theta_2)\sin(\theta_3 - \theta_2)\Big) \\
&\quad\vdots \\
&= \Big(\theta_M - \theta_{M-1}\Big)\Big(\theta_M - \theta_{M-1} + \cos(\theta_M + \theta_{M-1})\sin(\theta_M - \theta_{M-1})\Big)
\end{aligned} \tag{7.5.11}$$

in which the M bins are defined as being from θ_0 to θ_1, θ_1 to θ_2, etc., with $\theta_0 = 0$ and $\theta_M = 90°$. The equations can be solved iteratively. The partitioning is symmetrical about $\theta = 0$ and $\theta = 90°$. An example for eight bins is given in Figure 7.5.4 (cf. Nørklit and Vaughan, 1999).

The bins are larger in the broadside direction, where the Doppler shift is smallest. This is the opposite trend to the partitioning resulting from considering the angular resolution criteria. This type of partitioning can be used as an 'average best guess' initial partition for solving for the scatterer positions' amplitudes and phases.

7.5.4 Additive noise case including subspace/super-resolution methods

Although the noise-free case demonstrates the idea of signal prediction, and some fundamental limitations, the inevitable presence of noise in real-world systems calls for a different approach to the solution.

7.5.4.1 Point source model

The model for the sampled transfer function can be written, here for the mth sample, with a slight change in notation, as

$$h_m = \sum_{n-1}^{N} s_n \mathrm{e}^{\mathrm{j}u_n m} + \eta_m \tag{7.5.12}$$

in which the nth of N point-sources has complex amplitude s_n and η_m is the noise contribution.

The model is simply a sum of sinusoids in the presence of noise, and so the established spectral estimation techniques are called on. The basic techniques are reviewed below. The samples are considered as signals from a synthetic receiving array antenna of M elements. The subspace methods give super-resolution of the angular positions of the sources.

7.5.4.2 Vector notation

The vector notation is as follows. The directions are expressed as a function of zenith angle, θ, or as a function of the spatial Doppler frequency, $u = k_{\mathrm{C}} \cos\theta$. The M spatial samples are written

$$\boldsymbol{h} = [h_1, h_2, \ldots, h_M]^T , \tag{7.5.13}$$

and the N effective sources are assembled as a vector of their complex amplitudes

$$\boldsymbol{s} = [s_1, s_2, \ldots, s_N]^T , \tag{7.5.14}$$

and a separate vector of their directions expressed as

$$\boldsymbol{u} = [u_1, u_2, \ldots, u_N]^T . \tag{7.5.15}$$

The directional steering vector of the array to a source at u is

$$\begin{aligned} \boldsymbol{a}(\boldsymbol{u}) &= \left[\mathrm{e}^{\mathrm{j}uz_0}, \mathrm{e}^{\mathrm{j}uz_1}, \mathrm{e}^{\mathrm{j}uz_2}, \ldots, \mathrm{e}^{\mathrm{j}uz_{(M-1)}}\right]^T \\ &= \left[1, \mathrm{e}^{\mathrm{j}u\Delta z\cdot 1}, \mathrm{e}^{\mathrm{j}u\Delta z\cdot 2}, \ldots, \mathrm{e}^{\mathrm{j}u\Delta z\cdot(M-1)}\right]^T \end{aligned} \tag{7.5.16}$$

with the second line for the case of equi-spaced samples. The steering vectors to all of the sources are combined into the $M \times N$ matrix

$$\boldsymbol{A}(\boldsymbol{u}) = [\boldsymbol{a}(u_1)\ \boldsymbol{a}(u_2)\ \boldsymbol{a}(u_3)\ \ldots .\ \boldsymbol{a}(u_N)] \tag{7.5.17}$$

so that the vector form of equation (7.5.12) is

$$\boldsymbol{h} = \boldsymbol{A}(\boldsymbol{u})\boldsymbol{s} + \boldsymbol{\eta} \tag{7.5.18}$$

where $\boldsymbol{\eta}$ is the M-vector of noise contributions at the spatial sample points.

7.5.4.3 Gaussian continuous source model

For an uncorrelated scenario with directional power $p_\theta(\theta)$, the $M \times M$ correlation matrix of the samples is

$$\boldsymbol{R}_h = E\left\{\boldsymbol{h}\boldsymbol{h}^H\right\} = \int_{-\pi}^{\pi} \boldsymbol{a}(\theta)\, \boldsymbol{a}^H(\theta)\; p_\theta(\theta)\, \mathrm{d}\theta \tag{7.5.19}$$

and the model can be described statistically as the samples being distributed as complex Gaussians, i.e.

$$\boldsymbol{h} \sim \mathcal{CN}\,(\boldsymbol{0}, \boldsymbol{R}_h) \tag{7.5.20}$$

where $\boldsymbol{0}$ is an M-vector of zeros representing the zero means. The M-vector of samples ($\boldsymbol{h}$, the measurement segment) is to be used to predict the signal at the succeeding sample positions, written as the samples $\boldsymbol{h}_{M2} = h_{M+1}, h_{M+2}, \ldots, h_{M2}$.

7.5.4.4 Basic linear estimator for the continuous model

The basic linear predictor with minimum mean square error is as follows. The data are assembled as

$$\boldsymbol{h}_\mathrm{L} = \begin{bmatrix} \boldsymbol{h}_{M2} \\ \boldsymbol{h} \end{bmatrix} \tag{7.5.21}$$

and this has the complex normal distribution, i.e.

$$\boldsymbol{h}_\mathrm{L} \sim \mathcal{CN}\Big(\boldsymbol{0}, \boldsymbol{R}_{h\mathrm{L}}\Big), \quad \boldsymbol{R}_{h\mathrm{L}} = \begin{bmatrix} \boldsymbol{R}_h & \boldsymbol{R}_{h\,hM2} \\ \boldsymbol{R}^H_{h\,hM2} & \boldsymbol{R}_{hM2} \end{bmatrix}. \tag{7.5.22}$$

(An alternative formulation is to replace the prediction segment $\boldsymbol{h}_{M2}$ by the measurement plus prediction segment, $\boldsymbol{h}_\mathrm{L}$.) The MMSE for $\boldsymbol{h}_{M2}$ given $\boldsymbol{h}$ is the conditional expectation:

$$\hat{h} = E\,\{\boldsymbol{h}_{M2}|\boldsymbol{h}\} = \boldsymbol{R}_{hM2\,h}\boldsymbol{R}_h^{-1}\boldsymbol{h} \tag{7.5.23}$$

and the pseudo inverse is used if $\boldsymbol{R}_h$ cannot be inverted. The correlation functions require a model. For example from the Clarke uniform scenario, the ijth element of both of the correlation matrices is $J_0\,((i - \mathrm{j})2\pi d/\lambda)$. The basic linear estimator with the Gaussian model serves as a benchmark for the subspace methods described below, and does not predict the signal behaviour for more than the decorrelation distance of the signal (a fraction of a wavelength). The basic linear estimator serves only as a benchmark for the subspace methods described below, and predicts well in the presence of noise only for very short distances (less than 0.1 wavelengths), which is similar behaviour to that of polynomial extrapolation.

7.5.4.5 Estimating a correlation matrix from samples

There is a full set of samples for the correlation coefficient, although the samples are not taken simultaneously since the array is synthetic. It is the assumption that the scenario is static that allows the samples to be gathered and processed by a statistically consistent operation to produce a correlation matrix with appropriate properties.

The estimate of the correlation matrix is given by the 'covariance method', and the 'modified covariance method' (see, for example, Therrien, 1992). In the covariance method, the samples are arranged normally as the $P \times (M - P)$ matrix

$$\boldsymbol{H} = \begin{bmatrix} h_{M-P} & \dots & h_2 & h_1 \\ h_{M-P+1} & \dots & h_3 & h_2 \\ . & . & . & . \\ h_{M-1} & \dots & . & h_P \end{bmatrix} \tag{7.5.24}$$

and the $P \times P$ correlation matrix is

$$\hat{\boldsymbol{R}} = \left(\frac{1}{M - P + 1} \right) \boldsymbol{H}^H \boldsymbol{H}. \tag{7.5.25}$$

The allocation of rows and columns (i.e. the size of P) gives a trade-off between the number of samples used for averaging, and the size of the resulting correlation matrix, which in turn gives the number of eigenvalues from which the sources are selected (see below). In the limit with $P = M$, there is no averaging in the elements of $\hat{\boldsymbol{R}}$, but $\hat{\boldsymbol{R}}$ is as large as possible, being $M \times M$, and has the maximum number, M, of eigenvalues. The other limit is $P = 1$ and the correlation matrix is a scalar, averaged over all the product combinations of the data samples. Often, P is taken to be about the same as $M/2$ so that $\boldsymbol{H}$ is square, or almost square.

In the application here, the spatial frequencies are limited to less than $\pm k_C$, and this can be used to assist with the estimation algorithm. The samples can be taken to be denser than strictly necessary for estimation, and this can be used for computational advantage. The spatial samples are a distance $z = d$ apart. The maximum spatial Doppler frequency is therefore $\pm k_C d$. (More correctly, this is the spatial Doppler phase; the units are rads per inter-element spacing and sometimes referred to as rads per sample.) We construct a matrix of samples which overlap by $K < \lambda/(2d)$ samples further than the original M dimension; the $(M - PK + K) \times P$ matrix of samples is assembled as

$$\boldsymbol{H} = \begin{bmatrix} h_{(P-1)K+1} & \dots & h_{K+1} & h_1 \\ h_{(P-1)K+2} & \dots & h_{K+2} & h_2 \\ . & . & . & . \\ h_M & \dots & . & h_{M-(P-1)K} \end{bmatrix}. \tag{7.5.26}$$

The columns contain consecutive samples and the rows are samples which are many (K) spatial samples apart. The spatial Nyquist frequency is $\lambda/2$ (recall the assumption that the scatterers are in the far field and are static) and the adjacent samples in the rows should not exceed this, and this in turn limits the value of K, given above.

The $P \times P$ correlation matrix, from the covariance method, is

$$\hat{\boldsymbol{R}} = \hat{\boldsymbol{R}}_1 = \left(\frac{1}{M - PK + K} \right) \boldsymbol{H}^H \boldsymbol{H}. \tag{7.5.27}$$

The modified covariance uses an average of the forward direction (of the time or space samples) and backward direction calculations of the covariance:

$$\hat{\boldsymbol{R}} = \hat{\boldsymbol{R}}_2 = \frac{1}{2} \left(\hat{\boldsymbol{R}}_1 + \boldsymbol{J}\ \hat{\boldsymbol{R}}_1^* \boldsymbol{J} \right) \tag{7.5.28}$$

where $\boldsymbol{J}$ is a $P \times P$ matrix of ones and zeros with elements $(\boldsymbol{J})_{i,j} = \delta_{i,P+1-j}$, which is used for the backward calculation. Although the forward and backward calculations have the same statistical properties, the two snapshot estimates are different, so the modified covariance offers a more accurate estimation from the same data and removes the dilemma as to which direction gives the more correct correlation.

7.5.4.6 Estimating the sources

Estimating the number of sources is a classical problem in spectral estimation and array processing. The maximum likelihood method is not to be used for this estimate because increasing the degrees of freedom (i.e. the number of sources) will increase the likelihood function without bound. However, the Akaike criterion (Akaike, 1974), and the minimum description length (MDL) criterion (Wax and Kailaith, 1985) introduce a penalty for the number of sources in order to get round this problem. The minimum description length, which is known to be unbiased and consistent, is given here.

The correlation matrix is eigen-decomposed as

$$\hat{\boldsymbol{R}} = \hat{\boldsymbol{E}} \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{E}}^H.$$

From now on the hats on the quantities in this equation denoting estimates are dropped. Λ is the diagonal of P eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_P$. The MDL is the value of N which minimises

$$MDL(N) = -(P - N)(M - PK + K) \log \left(\frac{\prod_{i=N+1}^{P} \lambda_i^{1/(P-N)}}{(1/(P - N)) \sum_{i=N+1}^{P} \lambda_i} \right) + \frac{N}{2}(2P - N) \log(M - PK + K) \tag{7.5.29}$$

and this is denoted $\hat{N}$.

Once the number of sources is estimated, the matrices can be assembled and the source complex amplitudes are estimated by (see equation 7.5.18)

$$\hat{\boldsymbol{s}} = \left(\boldsymbol{A}^H(\hat{u})\ \boldsymbol{A}(\hat{u}) \right)^{-1} \boldsymbol{A}^H(\hat{\boldsymbol{u}})\ \boldsymbol{h} \tag{7.5.30}$$

where the $\hat{N}$ estimates of the directions, $\hat{\boldsymbol{u}}$, can be found from a variety of subspace techniques, as follows.

7.5.4.7 MUSIC

In the Multiple Signal Classification (MUSIC) algorithm (Schmidt, 1986), the eigenvectors from the above eigen-decomposition are separated into signal and noise contributions, according to the descending order of the eigenvalue size. The $\hat{N}$ largest eigenvalues correspond to the $\hat{N}$ sources and the remaining eigenvalues are assigned to be noise. The eigenvectors are partitioned as

$$\boldsymbol{E} = [\boldsymbol{E_1}, \boldsymbol{E_2}, \ldots \boldsymbol{E_M}] = [\boldsymbol{E_s}\ \boldsymbol{E_n}], \tag{7.5.31}$$

where

$$\boldsymbol{E_s} = \left[\boldsymbol{E_1}, \boldsymbol{E_2}, \ldots, \boldsymbol{E}_{\hat{N}}\right], \tag{7.5.32}$$

$$\boldsymbol{E_n} = \left[\boldsymbol{E}_{\hat{N}+1}, \boldsymbol{E}_{\hat{N}+2}, \ldots, \boldsymbol{E_M}\right]. \tag{7.5.33}$$

The $\hat{N}$ largest peaks of the function (sometimes called the pseudo spectrum)

$$\hat{P}(u) = \frac{1}{\boldsymbol{a}^H(u)\ \boldsymbol{E_n}\ \boldsymbol{E_n^H}\ \boldsymbol{a}(u)} \tag{7.5.34}$$

give the directions of the sources, $\hat{\boldsymbol{u}}$. These peaks occur where the steering vector $\boldsymbol{a}(u)$ corresponds to a direction u which is orthogonal (in the eigenvector sense) to the 'directions' of the noise contributions which are given by the noise eigenvectors.

7.5.4.8 Minimum norm algorithm

The minimum norm procedure (Kumaresan and Tufts, 1983) uses the same eigen-decomposition as MUSIC, but the noise subspace is represented by a single vector $\boldsymbol{e}$. The first component of $\boldsymbol{e}$ is scaled to 1, and the algorithm minimises the norm, $\sum \text{diag}\left(\boldsymbol{e}^H\boldsymbol{e}\right)$. The formulation is as follows. The noise is partitioned as

$$\boldsymbol{E_n} = \begin{bmatrix} \boldsymbol{c}^H \\ \boldsymbol{E_n^{\downarrow}} \end{bmatrix} \tag{7.5.35}$$

where $\boldsymbol{c}$ contains the top row of $\boldsymbol{E_n}$. The superscript $\downarrow$ denotes that the top row has been removed from the matrix. The scaling gives (Therrien, 1992)

$$\boldsymbol{e} = \begin{bmatrix} 1 \\ \boldsymbol{E_n^{\downarrow}}\boldsymbol{c}/\left(\boldsymbol{c}^H\boldsymbol{c}\right) \end{bmatrix} \tag{7.5.36}$$

and a pseudo spectrum is

$$\hat{P}(u) = \frac{1}{|\boldsymbol{a}^H(u)\boldsymbol{e}|^2}. \tag{7.5.37}$$

An alternative is to find the roots of the polynomial using the elements of $\boldsymbol{e}$,

$$G(w) = 1 + \sum_{p=1}^{M-\hat{N}} \boldsymbol{e}_i\, w^{-p} \tag{7.5.38}$$

where $w = e^{\mathrm{j}uz}$ and projecting these onto the unit circle along their radial direction.

7.5.4.9 Principal Components Linear Prediction (PCLP)

The PCLP (Tufts and Kumaresan, 1982a,b) is an autoregressive-type prediction, although instead of using the prediction filter as is, the roots of the prediction filter lead to the estimates of the directional cosines. The correlation matrix uses only the components corresponding to the signal subspace. The partitioning is

$$\hat{\boldsymbol{R}} = \begin{bmatrix} f & \boldsymbol{r}^H \\ -\boldsymbol{r} & \boldsymbol{R}^{\searrow} \end{bmatrix} \tag{7.5.39}$$

where $\boldsymbol{R}^{\searrow}$ is $\hat{N} \times \hat{N}$ and contains the $\boldsymbol{E}_s$ information, and f is not used further. The coefficients of the prediction filter are in $\boldsymbol{e} = \left[1\ \boldsymbol{d}^{\downarrow \mathrm{T}}\right]^{\mathrm{T}}$ where

$$\boldsymbol{d}^{\downarrow} = \sum_{i=1}^{\hat{N}} \frac{1}{\lambda_i} (\boldsymbol{e}_i^H \boldsymbol{r}) \boldsymbol{e}_i \tag{7.5.40}$$

and, as above, the frequency estimates can be from projecting the zeros of the filter onto the unit circle.

7.5.4.10 ESPRIT

The Estimation of Signal Parameters via Rotational Invariance Technique (Roy and Kailaith, 1989) partitions the source matrix $\boldsymbol{A}$, which spans the same subspace as the signal eigenvectors, $\boldsymbol{E}_s$. Submatrices are denoted as follows. Deleting the top row of $\boldsymbol{A}$ leaves $\boldsymbol{A}^{\downarrow}$, and deleting the bottom row of $\boldsymbol{A}$ leaves $\boldsymbol{A}^{\uparrow}$. Similarly, the notation holds for $\boldsymbol{E}_s^{\downarrow}$ and $\boldsymbol{E}_s^{\uparrow}$. Writing the elements of the steering vector in a diagonal matrix as

$$\boldsymbol{\gamma} = \mathrm{diag}\left(\mathrm{e}^{\mathrm{j}u\Delta z \cdot 1}, \mathrm{e}^{\mathrm{j}u\Delta z \cdot 2}, \ldots, \mathrm{e}^{\mathrm{j}u\Delta z \cdot \hat{N}}\right) \tag{7.5.41}$$

allows the expression $\boldsymbol{A}^{\downarrow} = \boldsymbol{\gamma}\boldsymbol{A}^{\uparrow}$. Also, we can write $\boldsymbol{E}_s^{\downarrow} = \boldsymbol{\Phi}\boldsymbol{E}_s^{\uparrow}$ and because $\boldsymbol{\Phi}$ and $\boldsymbol{\gamma}$ are similar matrices (the source matrix and the signal matrices span the same space), then their eigenvalues are the same. So the eigenvalues of $\boldsymbol{\Phi}$ contain the directional information of the sources. The least squares estimate of $\boldsymbol{\Phi}$ is

$$\hat{\boldsymbol{\Phi}} = \left(\boldsymbol{E}_s^{\uparrow H} \boldsymbol{E}_s^{\uparrow}\right)^{-1} \boldsymbol{E}_s^{\uparrow H} \boldsymbol{E}_s^{\downarrow}. \tag{7.5.42}$$

Again, the roots are projected onto the unit circle which suits the model of constant (non-decaying or growing in amplitude) far-field sources. (Retaining the non-unity magnitude of the roots gives a lower mean square error, but the prediction performance for this particular problem decreases.)

7.5.4.11 Simulation results

The performance of the signal prediction depends on its specific application. Simulation studies have been reported (Andersen *et al.*, 1999; Vaughan *et al.*, 2000) which use a *prediction distance*, defined as the distance for which the predicted and actual

signal envelopes depart by some defined quantity. The results below use a value of 20 per cent of the RMS value of the envelope in the measurement segment. Figure 7.5.5 gives an example of the prediction behaviour.

The subspace techniques have reasonably similar performances. An example from a selection of the best performances is given in Figure 7.5.6. For this, the finite number of sources were Rayleigh distributed, although no significant difference arises if constant amplitudes are used. The Minimum Descriptor Length (MDL) criterion proved reliable. From the figure, as long as the number of sources is small, then the prediction works for measurement segments greater than a couple of wavelengths. A fixed number of samples (40) was used here, but it makes little difference to sample more densely than about a half wavelength (although the SNR may be improved this way). From the above section, there are three unknowns for each source and so a minimum of some three real samples, or on average one and a half complex samples which may occupy some three quarters of a wavelength, is required for each source to be solved. This is evident in the plot, where the prediction starts to work after a measurement distance proportional to a measurement distance following this rule of thumb.

The effect of the SNR is also of interest. This is shown in Figures 7.5.7 and 7.5.8. The prediction does not work well for many sources (several tens) unless the measurement segment is correspondingly long, or the SNR is unrealistically high. From

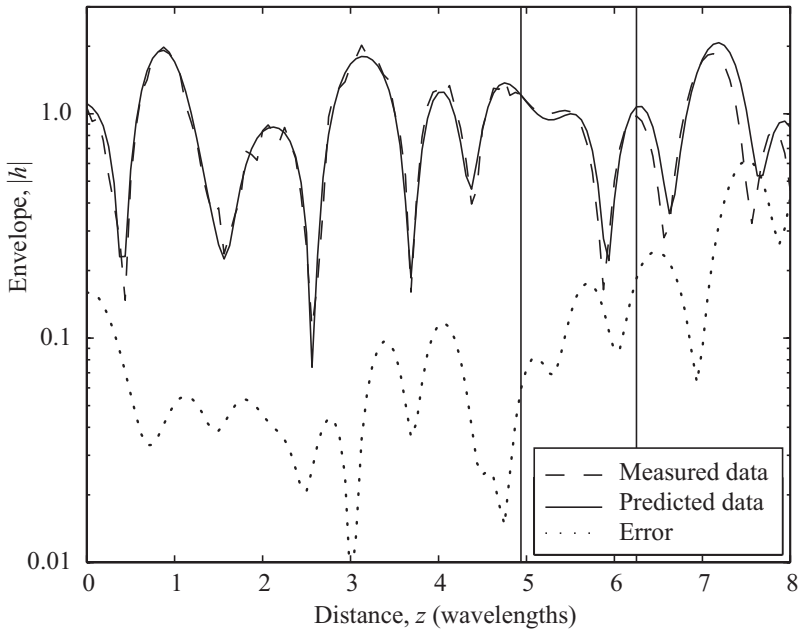


Figure 7.5.5 *An example of signal prediction behaviour. The data to the left of the left line are the measurement segment and the data in the region to the right are the prediction segment. The region between the lines is where the predicted envelope is within 20 per cent of the actual envelope. The RMS value of the measured envelope is normalised to 1, so the prediction fails when the error waveform first reaches 0.2*

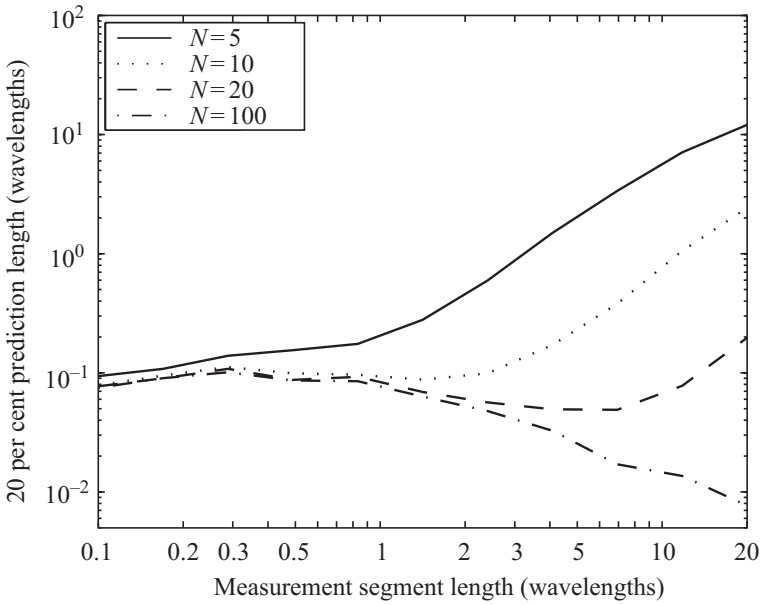


Figure 7.5.6 The 20 per cent prediction length in wavelengths against the length of the measurement segment for the best of the subspace techniques MUSIC, ESPRIT, PCLP, and Minimum norm. The number of samples in the measurement segment is fixed at 40, and the SNR is kept to 20 dB. The sources have uniformly distributed directions and phases, and Rayleigh distributed amplitudes. The Monte-Carlo run size was 3000

these simulation results, the length of the measurement segment limits the prediction application for realistic SNRs. This is because in reality it would be unlikely to get a sufficiently long linear trajectory over which the sources appear static for a large number of sources and a single antenna. An interesting use for the prediction technique may be to serve as a test for real-world data being produced by a small number of sources. The pdf of the prediction length is similar to an exponential distribution, although the Kolmogorov–Smirnov (see Appendix B) test fails for the hypothesis that it is exponential.

7.5.5 Real-world experimental results

Experimental results were reported by Andersen *et al*. (1999), using a roof-mounted car antenna in a suburban environment at 1.8 GHz and a speed of about 20 km h^{-1}. The measurement segment length was about one-third of a wavelength, comprising 20 samples. Using a modified ESPRIT method, the average prediction length was reported to be 0.95λ. This result well exceeds the possible prediction length limit indicated by the simulation results of Figure 7.5.6, where a maximum prediction length of $\sim 10^{-1}\lambda$ corresponds to the $\lambda/3$ measurement segment. It agrees, however, with simulation results based on the modified ESPRIT technique. The theoretical results are given in Figure 7.5.9 (Andersen *et al.*, 1999), which shows the performance

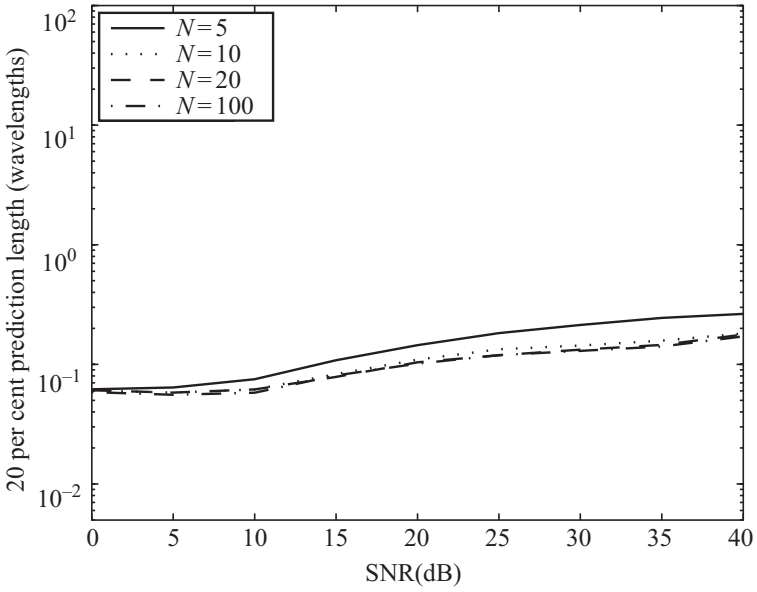


Figure 7.5.7 The prediction length from Figure 7.5.6 against the signal-to-noise ratio for a measurement segment of 1/3 wavelengths

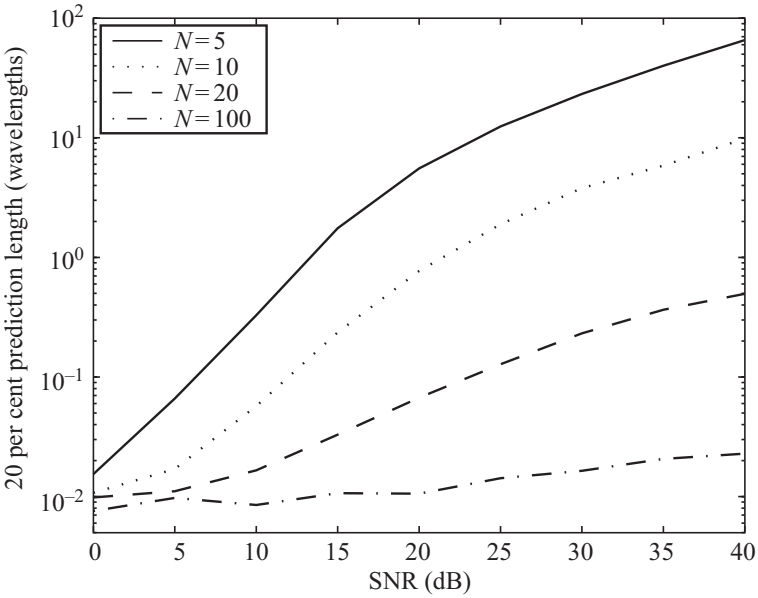


Figure 7.5.8 The prediction length from Figure 7.5.6 against the signal-to-noise ratio for a measurement segment of ten wavelengths

(prediction length) vs the analysis length for number of sources equal to 10, 20 and 300. For a large number of sources the prediction length is close to one wavelength in agreement with the experimental results.

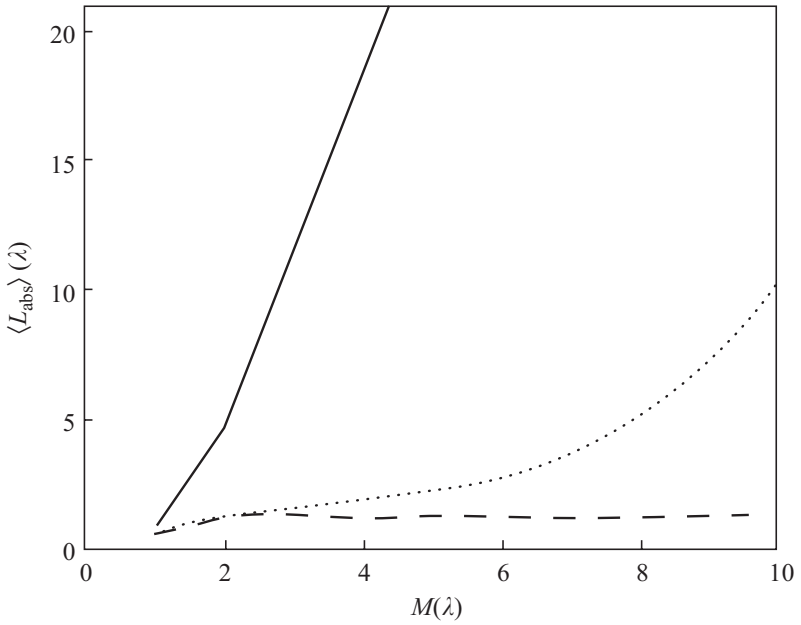


Figure 7.5.9 Performance vs analysis segment length. Solid, $N = 10$; dotted, $N = 20$; dashed, $N = 300$

Further information and results on this type of signal prediction are in Ekman (2000), Teal and Vaughan (2001), and Teal (2001); and Fleury *et al*. (1999) give super-resolution techniques for sources in mobile scenarios. P. Teal and R. Raich are acknowledged for their assistance in the review of super-resolution methods presented in this section.

7.6 Correlation analysis for measured narrowband signals

This, and the following section, give a theoretical background of sounding the mobile channel. The narrowband case is treated in this section, with particular attention to the problem of the large dynamic range encountered in flat fading, and its ramification for processing the measured data used to yield correlation information. The topic of finite samples is addressed first, dealing with the Gaussian quadrature signals. The envelope correlations are then addressed, with the effect of clipping caused by the large dynamic range of mobile signals, and the effect of demeaning is included, with a measured data example.

7.6.1 Estimating the correlation coefficient of Gaussian signals from finite samples

Related to the estimation of the mean using a finite distance or number of samples (Chapter 6), is the topic of estimating the correlation coefficient. This is required,

for example, for evaluating diversity performance by time- (or distance-) series correlation, and is also used extensively in real-time processing for communications.

7.6.1.1 Independent samples

The non-parametric case has been treated by Withers (1977). For n independent samples, the variance of the correlation coefficient is given as a series as $\sigma_\rho^2 = A/n + B/n^2 + \ldots$, where (Withers, 1977)

$$A = \frac{1}{\mu_{20}\mu_{02}}\left(\mu_{22} - \mu_{11}\left(\frac{\mu_{31}}{\mu_{20}} + \frac{\mu_{13}}{\mu_{02}}\right) + \frac{\mu_{11}^2}{4}\left(\frac{2\mu_{22}}{\mu_{20}\mu_{02}} + \frac{\mu_{40}}{\mu_{20}^2} + \frac{\mu_{04}}{\mu_{02}^2}\right)\right), \tag{7.6.1}$$

and this first term alone is used for the asymptotic case of large n. The μ_{ij} are the joint ijth moments (or their sample estimates) of the joint distribution of the signals being sampled. For finite samples the formulation is better suited to distributions which do not have long tails, since higher-order moments are required.

The case for the correlation coefficient between bivariate normally distributed samples has been treated by Fisher (1921), who gives the exact pdf of the correlation coefficient for the case of $n > 3$. Denoting the estimate of the correlation coefficient as r, where r is calculated in the usual way using

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_x)\left(y_i - \mu_y\right)}{\sqrt{\left(\sum_{i=1}^{N}(x_i - \mu_x)^2\right)\left(\sum_{i=1}^{N}\left(y_i - \mu_y\right)^2\right)}} \tag{7.6.2}$$

where μ is the mean, then Fisher's pdf can be expressed (Hotelling, 1953; Anderson, 1958) as

$$\begin{aligned} p_r(r) = {} & \frac{n-1}{\sqrt{2\pi}}\frac{\Gamma(n)}{\Gamma(n+1/2)}\left(1-\rho^2\right)^{n/2}\left(1-r^2\right)^{(n-3)/2} \\ & \cdot (1-\rho r)^{-n+1/2}\, F\left(\frac{1}{2}, \frac{1}{2}; n+\frac{1}{2}; \frac{1+\rho r}{2}\right) \end{aligned} \tag{7.6.3}$$

where F is the hypergeometric function,

$$F(a, b; c; x) = \sum_{i=0}^{\infty}\frac{\Gamma(a+i)}{\Gamma(a)}\frac{\Gamma(b+i)}{\Gamma(b)}\frac{\Gamma(c)}{\Gamma(c+i)}\frac{x^i}{\Gamma(i+1)}, \tag{7.6.4}$$

which converges quickly (typically less than 10–30 terms). Some examples of the pdf are given in Figure 7.6.1 to provide a feel for the behaviour. The variance reduces when the true correlation coefficient, ρ, increases from 0, and becomes zero when $\rho = \pm 1$. The case $\rho = 0$ has a symmetric pdf, but the pdf is asymmetric for all other ρ. Note that the mode of the asymmetric pdfs does not equate to the expected value. The cases for $\pm\rho$ are the same in the sense that they are mirror images on the plot.

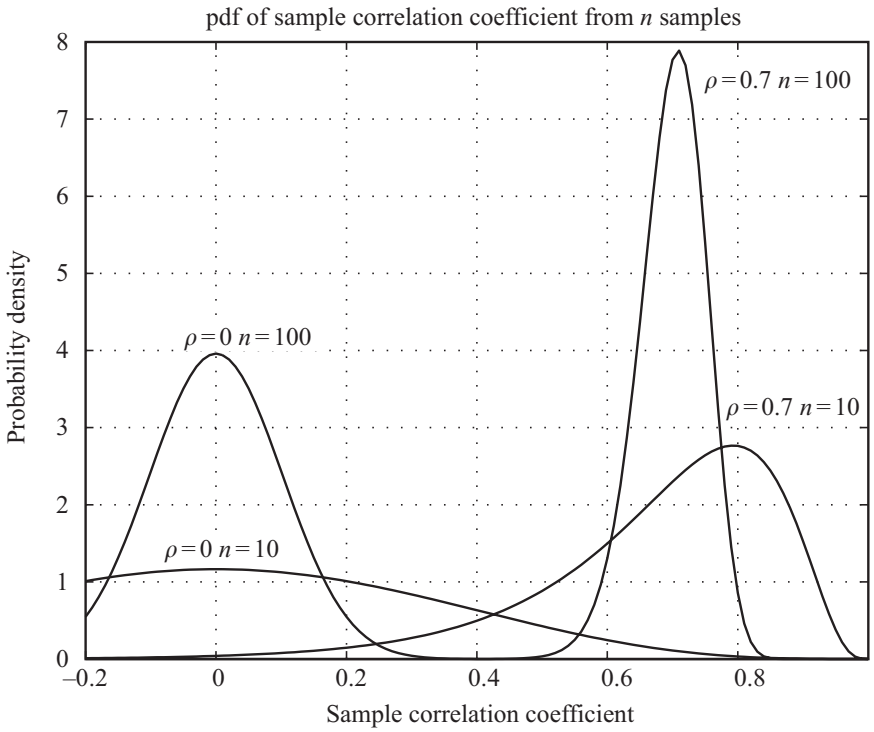


Figure 7.6.1 Pdfs for the correlation coefficient estimate from n independent samples. The variance is a maximum when the true correlation coefficient is $\rho = 0$, and is zero when $\rho = \pm 1$

The above expression for the pdf is complicated, but applying Fisher's transformation to ρ, and to its estimate, r, namely

$$\eta = \tanh^{-1}(\rho), \quad z = \tanh^{-1}(r) \tag{7.6.5}$$

provides an asymptotically (large n) normal distribution for z as $\mathcal{N}\left(\eta, 1/\sqrt{n-3}\right)$. This can be written that the asymptotic pdf of $\sqrt{n-3}\,(z-\eta)$ is $\mathcal{N}(0, 1)$. By taking η as always normally distributed (as can be seen from Figure 7.6.1, this is not a good approximation for low values of n), normal confidence limits can be written directly for the correlation coefficient as a function of the number of independent samples. For example, the 95 per cent limits are expressed

$$P(-1.96 \leq \sqrt{n-3} \leq 1.96) = 0.95 \tag{7.6.6}$$

i.e. the 95 per cent lower and upper limits (i.e. $\pm 2\sigma$) are, respectively,

$$[\rho_{\mathrm{U}}, \rho_{\mathrm{L}}] = \left[\tanh^{-1}\left(\frac{-1.96}{\sqrt{n-3}}\right) + \tanh^{-1}\rho, \tanh^{-1}\left(\frac{+1.96}{\sqrt{n-3}}\right) + \tanh^{-1}\rho\right] \tag{7.6.7}$$

and these are depicted in Figure 7.6.2. For a given sample correlation coefficient, r, the 95 per cent range of values for ρ is given on the ordinate. For example, for

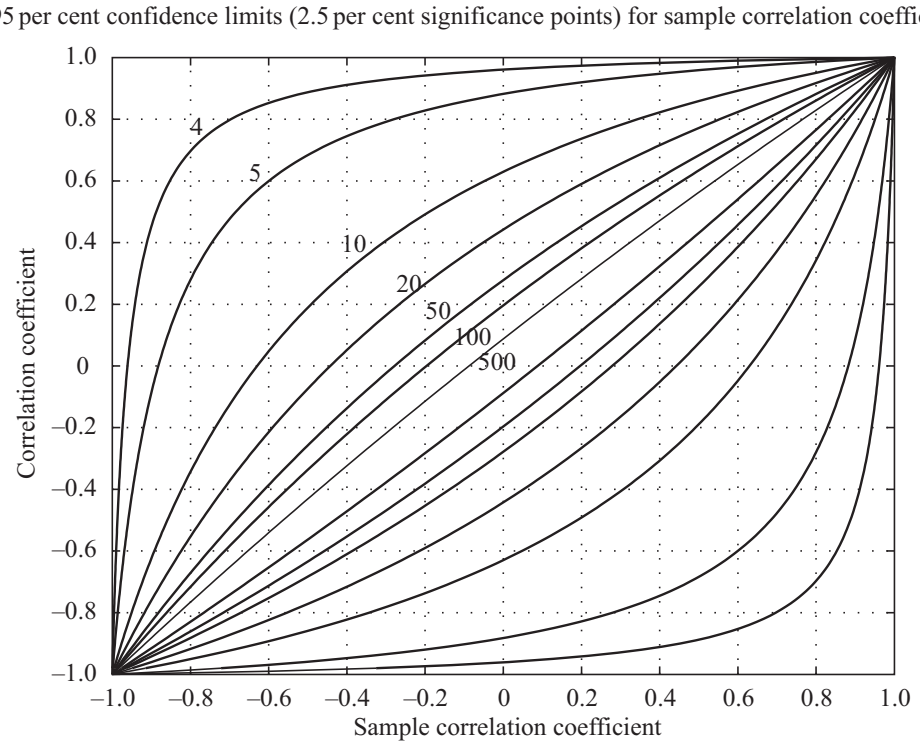


Figure 7.6.2 The 95 per cent confidence intervals for the sample correlation coefficient with the number of independent samples as a parameter. The plot assumes a normal distribution, which holds with increasing accuracy for increasing n

$r = -0.2$, calculated from $n = 20$ samples, then 95 per cent of such calculations drawn from the same data distribution, will lie between $\rho = -0.6$ and $+0.25$. The zero variance for $\rho = \pm 1$, and the maximum variance for $\rho = 0$, cases are evident.

This information can be applied to the correlation functions used in diversity evaluation, for example, in order to estimate the number of samples required for an accurate correlation calculation. The limits on the spatial diversity correlation coefficient for a uniform scenario are shown in Figure 7.6.3 (Vaughan, 2000). Here the J_0 form is enveloped by the 95 per cent limits for various n. The interpretation is that 95 per cent of independently measured correlation coefficient functions will fit into the envelope. In order to get a reliable estimate which is close to the proper function, the figure shows explicitly that a very large number of samples is required, in particular away from the origin where the function becomes small. A crude estimate for distance-series in the omnidirectional scenario, is to take samples every half wavelength, as discussed in the above section, so dividing the number of samples n by two gives the approximate distance required for a measurement.

7.6.2 Envelope correlations

The envelope, or the power, is the parameter of main interest for the narrowband channel, because it defines the radio coverage and because envelope correlations and

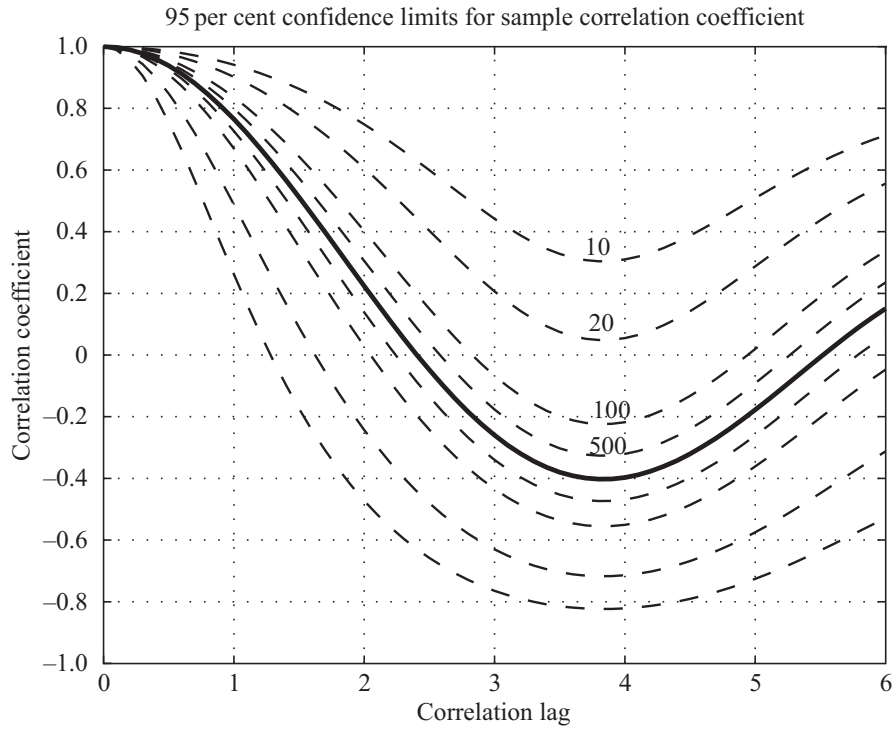


Figure 7.6.3 The 95 per cent confidence intervals applied to the correlation coefficient for spatial diversity in an omnidirectional scenario, with the number of samples used for the estimate as a parameter; 95 per cent of calculations fall within the envelope. The true correlation coefficient for the omni case is the J_0 form, shown with the thicker line

cross-correlations can reveal virtually all information for the spatial behaviour and diversity evaluation. The envelope can be obtained with relative ease from a modestly priced receiver. Simple equipment is an advantage when measurements are taken for the mobile channel, where often the measurement equipment itself must be mobile. The complex envelope, used for more complete channel characterisation, is more difficult to measure and needs considerably more expensive equipment. It is normally measured by mixing down to complex baseband and sampling the bandlimited in-phase and quadrature signals independently. The envelope and phase can then be calculated directly if these are required as explicit quantities as well.

Most receivers feature a signal which is called the Received Signal Strength Indicator, or RSSI. This signal is a voltage which is approximately proportional to the log of the power (signal plus noise) in the bandwidth of the receiver. The proportionality is typically nonlinear with a quasi-linear region. An example of the transfer function is given in Figure 7.6.4.

For a measurement setup, the RSSI voltage is typically passed through an analogue-to-digital converter (ADC) and the samples are logged as a time series. It is useful to have a position sensor to trigger the samples so that the data are a properly defined spatial series (i.e. the spacing between samples is well-defined).

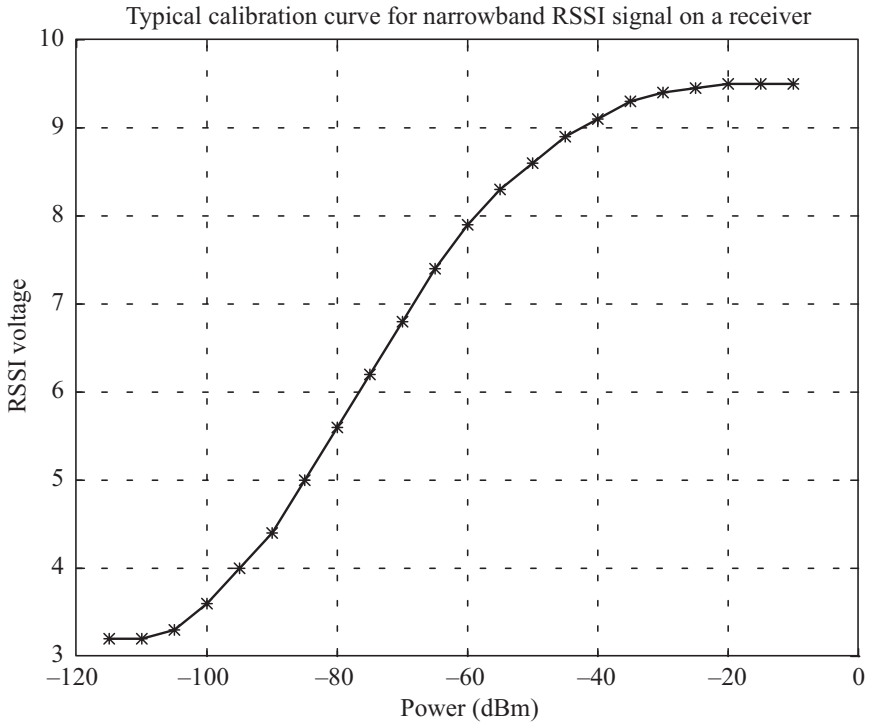


Figure 7.6.4 *A typical calibration curve representing the transfer function for the RSSI signal and the signal power. The stars are measured points and the line is the linear interpolation*

Otherwise a simple clock timer can be used to trigger the sampling times and, if possible, some control applied to the speed of the mobile terminal so that the spacing between samples can be estimated.

A measurement system is calibrated by feeding a signal generator with known power levels into the receiver to produce a table of values such as those in Figure 7.6.4. The measurement values can then be passed through such a look-up table to record the correctly calibrated power values for the time series. The envelope, if required, needs a square root operation to be applied to the power data. Finally, most receivers have an automatic gain controller (AGC) and care should be taken that its damping time-constant does not adversely affect the variation of the RSSI.

For coverage measurements, spaced samples are required at spacings which result in essentially uncorrelated samples. For regenerating the envelope from samples, the minimum sampling requirement can be gauged from the spectrum of the signal. The spectrum of the power signal is simpler than that of the envelope in the sense that it can be derived directly from Gans' power spectral model. For example, the power spectrum of the electric field incident at a vertical electrical dipole antenna in a uniform scenario is (Jakes, 1974)

$$S_{\mathrm{E}}^{(\mathrm{U})}(f) = \frac{3b_0}{\omega_{\mathrm{m}}}\left(1 - (f/f_{\mathrm{m}})^2\right)^{-1/2} \tag{7.6.8}$$

where $f = \omega/(2\pi)$ is the baseband frequency in Hz, $f_\mathrm{m} = V/\lambda$ is the maximum Doppler frequency shift at an antenna moving at velocity $V\,\mathrm{m\,s^{-1}}$ in a static scenario, and b_0 is the power available from the scenario to an ideal, omnidirectional, unity gain antenna. This spectrum is always (i.e. for any fixed-beam antenna and static scatterers) limited to the maximum Doppler frequency. For the omnidirectional case, the cut-off is at a singularity, i.e. where $f = \pm f_\mathrm{m}$. The bandwidth is therefore never greater than $2f_\mathrm{m}$. In principle, a minimum sampling frequency of greater than $4f_\mathrm{m}$ is thus required to digitise the signal without aliasing. In practice, moving scatterers can extend this bandwidth. For example, a vehicle moving at velocity V may receive an rf component from another vehicle, in the reverse direction, travelling at $-V$, resulting in a baseband Doppler frequency component at $-2f_\mathrm{m}$ and giving an rf spectrum bandwidth of $3f_\mathrm{m}$.

The bandwidth of the in-phase and quadrature components of the channel, I and Q, is given directly from the generation of these signals. The baseband signal, in the spatial domain, is

$$s(z) = I(z) + \mathrm{j}Q(z) = \sum_i a_i \mathrm{e}^{\mathrm{j}\alpha_i} \mathrm{e}^{\mathrm{j}u_i z}, \tag{7.6.9}$$

where a_i, α_i and u_i describe the ith effective scatterer. So the spectra of both components are confined to the spectrum of the effective scatterers, i.e. $\pm k_\mathrm{C}$ rad m^{-1} or $\pm f_\mathrm{m}$ Hz. The power of the signal is the envelope squared, $r^2 = |s|^2 = I^2 + Q^2$, and the squaring action doubles the bandwidth to $\pm 2f_\mathrm{m}$ Hz.

The envelope spectrum is found from the Fourier transform of a truncated series approximation to the envelope autocorrelation function (cf. Jakes, 1974). The autocorrelation function is given by (cf. Uhlenbeck, 1943)

$$\begin{aligned} R_r(\tau) &= b_0 \frac{\pi}{2} F\left[-\frac{1}{2}, -\frac{1}{2}; 1, \rho^2(\tau)\right] = b_0\left[2E\left(\rho^2(\tau)\right) - \left(1-\rho^2(\tau)\right)K\left(\rho^2(\tau)\right)\right] \\ &= b_0 \frac{\pi}{2}\left[1 + \frac{1}{4}\rho^2(\tau) + \frac{1}{64}\rho^4(\tau) + \cdots\right] \\ &\approx b_0 \frac{\pi}{2}\left[1 + \frac{1}{4}\rho^2(\tau)\right], \end{aligned} \tag{7.6.10}$$

where F is the Gaussian hypergeometric function, K and E are the complete elliptic integrals of the first and second kind respectively,

$$K(m) = \int_0^{\pi/2} \left(1 - m\sin^2\theta\right)^{-1/2} \mathrm{d}\theta, \quad E(m) = \int_0^{\pi/2} \left(1 - m\sin^2\theta\right)^{1/2} \mathrm{d}\theta; \tag{7.6.11}$$

and $\rho^2 = \rho_{II}^2 + \rho_{IQ}^2$ is the square magnitude of the complex correlation coefficient function of the narrowband signals. This series approximation is accurate to 1.8 per cent at the origin. For an omni antenna in a uniform scenario, $\rho^2 = \rho_{II}^2 = J_0^2(k_\mathrm{C} V\tau)$,

and the spectrum of the envelope, expressed again in Hz, is

$$\begin{aligned} S_r^{(\mathrm{U})}(f) &= \int_{-\infty}^{\infty} R_r(\tau)\mathrm{e}^{-\mathrm{j}\omega\tau}\,\mathrm{d}\tau \\ &\approx \int_{-\infty}^{\infty} \frac{\pi}{2} b_0 \left(1 + \frac{1}{4} J_0^2(k_\mathrm{C} V\tau)\right) \mathrm{e}^{-\mathrm{j}\omega\tau}\,\mathrm{d}\tau \\ &= b_0 2\pi^3 \delta(f) + \frac{b_0}{8\pi(2f_\mathrm{m})} K\left(\sqrt{1-(f/(2f_\mathrm{m}))^2}\right). \end{aligned} \tag{7.6.12}$$

The first term in the spectral expression is a d.c. component and the second term is the approximation (series truncation) to the continuous spectrum.

The spectrum is confined to frequencies between 0 and $2f_\mathrm{m}$ for the uniform scenario case. Since the envelope is purely real, if samples are taken more often that every $4f_\mathrm{m}$, i.e. every quarter wavelength (or two per average fade length), then, in principle, the envelope can be correctly reconstructed by interpolation. In practice, it is convenient to strongly over-sample, say several tens of samples per fade, and then interpolation is either not required or is trivial. Such dense sampling is needed for correlation analysis of the short-term fading envelope, where the correlation displacement resolution is governed by the sample density.

The spectrum of the envelope signal is not always bandlimited as in the uniform, or omnidirectional, case. This is demonstrated using the two-path model with the effective scatterers directly in front of and behind the receiver, i.e. $u_1 = k_\mathrm{C}, u_2 = -k_\mathrm{C}$. The envelope is

$$\begin{aligned} r^{(2)}(z) &= \left| a_1 \mathrm{e}^{\mathrm{j}\alpha_1} \mathrm{e}^{\mathrm{j}k_\mathrm{C} z} + a_2 \mathrm{e}^{\mathrm{j}\alpha_2} \mathrm{e}^{-\mathrm{j}k_\mathrm{C} z} \right| \\ &\propto \left| \cos k_\mathrm{C} z \right| \quad \text{for} \quad a_1 \mathrm{e}^{\mathrm{j}\alpha_1} = a_2 \mathrm{e}^{\mathrm{j}\alpha_2}; \end{aligned} \tag{7.6.13}$$

which has an infinite spectrum, but this is a rather unusual, special case.

7.6.2.1 Clipped envelopes

The Rayleigh-like envelopes, or the power signals, encountered in mobile communications occupy a wide dynamic range, typically 10 dB above the local mean to 30 dB below the local mean, with the range decreasing for increasing bandwidth. The local mean varies owing to shadow fading, by tens of dB and, on top of this, the inverse-distance loss can vary by some 50 dB within a cell. This means that the dynamic range of the RSSI signal may be over 100 dB, which is larger than that available from a low-cost receiver (cf. Figure 7.6.4). In characterising the Rayleigh-like fading of real-world channels, measuring equipment needs to be expensive to cover such a dynamic range or else signal limiting, or clipping, occurs. The clipping level varies within the Rayleigh envelope variation range because of the shadow fading and inverse-distance loss variation.

For evaluating the correlation coefficient between channels using low-cost receivers and signal acquisition equipment, this section treats the effect of clipped

envelopes on the correlation coefficient. It is shown that even hard-clipped envelopes are useful for diversity channel evaluation using the correlation coefficient.

Under the assumption of the signals being mean- and correlation-ergodic over the interval of a measurement, the envelope is sampled and the correlation coefficient estimate is computed in the usual way. Typically, several short-interval (several tens of wavelengths) measurements are made to approximate stationarity of the signal, and the local correlation coefficients are averaged for a given physical environment such as a building or a street. Larger records can be taken and the envelope demeaned before correlation calculation. This latter technique is also important for cross-correlations, where the non-stationary nature of the signal is common to both the signals being correlated.

It is known by experimenters that clipping the lower values of the envelopes has little impact on the correlation coefficient. This is because the signal energy at the signal fades is low, and the relative duration of the deep fades is small, so the contribution of this part of the signal to the correlation coefficient estimate is small. It is assumed that the noise level is well below the clipping level. Noise effects are discussed in Eggers (1993).

7.6.2.2 Formulation for clipping

A limiter has input signal $r(t)$ and produces an output signal $y(t)$. If the limiter mapping function $y(r)$ is known then the moments of the output signal are known in terms of the pdf of the input signal, for example the correlation is denoted

$$R_{y_1y_2} = \langle y_1 y_2 \rangle = \iint y_1(r_1) y_2(r_2) p_{r_1r_2}(r_1, r_2)\, \mathrm{d}r_1 \mathrm{d}r_2. \tag{7.6.14}$$

For two Rayleigh distributed envelopes, the joint pdf is (cf. equation (3.7–13) in Rice (1944)),

$$p_{r_1r_2}(r_1, r_2) = \frac{r_1 r_2}{\sigma_1^2 \sigma_2^2 \left(1 - |\rho|^2\right)} I_0\left(\frac{|\rho|}{\left(1 - |\rho|^2\right)} \frac{r_1 r_2}{\sigma_1 \sigma_2}\right) \cdot \exp\left[-\frac{1}{2\left(1 - |\rho|^2\right)} \left(\left(\frac{r_1}{\sigma_1}\right)^2 \times \left(\frac{r_2}{\sigma_2}\right)^2\right)\right], \tag{7.6.15}$$

where here $\rho = \rho_{I_1I_2} = \rho_{Q_1Q_2}$ is the correlation coefficient of the components of the complex Gaussian random variables $s_1 = I_1 + \mathrm{j}Q_1$ and $s_2 = I_2 + \mathrm{j}Q_2$; σ^2 is the variance of each of the in-phase and quadrature components (I_1, I_2 and Q_1, Q_2); and the envelopes are $r_1 = |s_1|$, $r_2 = |s_2|$. Note that in general, $\rho = \rho_{I_1I_2} - \mathrm{j}\rho_{I_1Q_2}$, and for complex Gaussians, $\rho_{I_1Q_2} = -\rho_{I_2Q_1} = 0$.

$|\rho|^2$ is equal to the correlation coefficient of the power signals, $\rho_{r_1^2 r_2^2}$ for Gaussian components, and from the data of Pierce and Stein (1960), this is approximately (within 10 per cent) equal to the correlation coefficient of the envelopes of the complex Gaussians, $\rho_{r_1r_2}$ (see Appendix B). Here the joint pdf is converted to power quantities since the correlation coefficient formulation is then exact. The power signal is given

in terms of signal-to-noise ratio, although the term 'power signal' is adhered to in this section. The power and its mean value are denoted

$$\gamma = \frac{r^2}{N_0}, \quad \Gamma = \frac{\langle r^2 \rangle}{N_0} = \frac{2\sigma^2}{N_0}; \tag{7.6.16}$$

where N_0 is the noise power. We can interpret γ/Γ as the ratio of the 'instantaneous' carrier-to-noise ratio to the mean carrier-to-noise ratio. As noted above the noise is not considered explicitly here, i.e. it is assumed to be negligible. The joint pdf of the powers follows as (Stein, 1966)

$$p_{\gamma_1\gamma_2}(\gamma_1, \gamma_2) = \frac{1}{\Gamma_1\Gamma_2\left(1-|\rho|^2\right)} I_0\left(\frac{2|\rho|}{\left(1-|\rho|^2\right)}\sqrt{\frac{\gamma_1\gamma_2}{\Gamma_1\Gamma_2}}\right) \cdot \exp\left[-\frac{1}{\left(1-|\rho|^2\right)}\left(\frac{\gamma_1}{\Gamma_1}+\frac{\gamma_2}{\Gamma_2}\right)\right]. \tag{7.6.17}$$

The pdf of the power signals is the exponential form

$$p_\gamma(\gamma) = \frac{1}{\Gamma} \mathrm{e}^{-\gamma/\Gamma}. \tag{7.6.18}$$

7.6.2.3 Soft clipping

In soft clipping, a portion of the signal is passed linearly, the top is clipped at $\gamma = c_\mathrm{U}$, and the bottom is clipped at $\gamma = c_\mathrm{L}$. This approximates the 's-shaped' transfer function of Figure 7.6.4. The mapping function which produces the soft-clipped signal y is

$$\begin{aligned} \gamma < c_\mathrm{L} \quad & y = 0 \\ c_\mathrm{L} \le \gamma < c_\mathrm{U} \quad & y = \gamma - c_\mathrm{L} \\ \gamma \ge c_\mathrm{U} \quad & y = A, \end{aligned} \tag{7.6.19}$$

which is shown in Figure 7.6.5.

The mean of y is then

$$\mu_y = \Gamma \mathrm{e}^{-c_\mathrm{L}/\Gamma} + (A - \Gamma + c_\mathrm{L})\,\mathrm{e}^{-c_\mathrm{U}/\Gamma} \tag{7.6.20}$$

and for bottom clipping only, the expression reduces using $c_\mathrm{U} = \infty$, and for no clipping, i.e. with $c_\mathrm{L} = 0$ as well, the mean becomes Γ as expected. The correlation of y_1 and y_2 is

$$R_{y_1y_2} = \int_{c_{\mathrm{L}_1}}^{c_{\mathrm{U}_1}} \int_{c_{\mathrm{L}_2}}^{c_{\mathrm{U}_2}} \left(\gamma_1 - c_{\mathrm{L}_1}\right)\left(\gamma_2 - c_{\mathrm{L}_2}\right) p_{\gamma_1,\gamma_2}(\gamma_1, \gamma_2)\, \mathrm{d}\gamma_1\, \mathrm{d}\gamma_2. \tag{7.6.21}$$

The correlation coefficient is given by

$$\rho_{y_1y_2} = \frac{R_{y_1y_2}(\rho(\chi)) - \mu_{y_1}\mu_{y_2}}{R_{y_1y_2}(\rho(\chi=0)) - \mu_{y_1}\mu_{y_2}}, \tag{7.6.22}$$

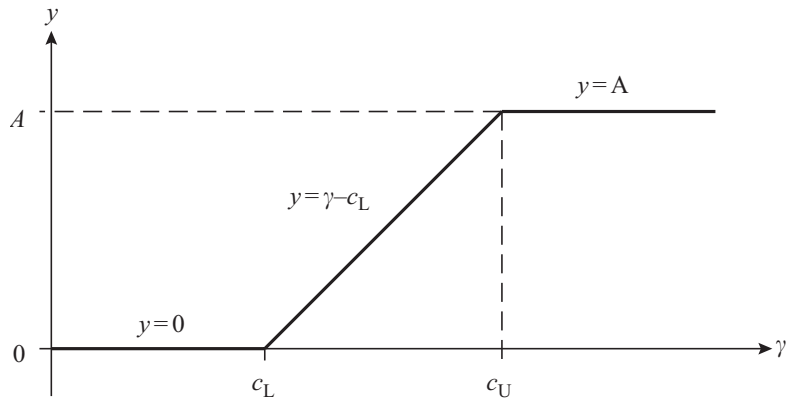


Figure 7.6.5 The soft limit transfer function depicted for the power signal γ. The lower and upper clipping levels are c_L and c_U and the output y varies between 0 and A

where χ is a correlation displacement for the correlation coefficient function appearing in p_{γ_1,γ_2}, and which is introduced for convenient formulation of the correlation of the input signals.

7.6.2.4 Hard clipping

The hard-clipping case can be viewed as a common clipping level of $c_L = c_U = c$ where the mapping is

$$\begin{aligned} \gamma < c \quad & y = 0 \\ \gamma \geq c \quad & y = A. \end{aligned} \tag{7.6.23}$$

The situation is shown in Figure 7.6.6 and the correlation integral of the hard limited signals is shown below to have a closed-form solution which can be solved iteratively.

The output signal is binary, which makes a practical correlation evaluation very low on computing requirements, and lends itself well to simple dedicated hardware solution for quasi-real time correlation processing. Measurement hardware is simplified as well, with just a floating reference level to be implemented to control the clipping level. The need for an ADC is removed, although the limiter can be viewed as a one-bit ADC. The question remains as to how to interpret the results.

The correlation coefficient for the hard-limiting case reduces to

$$\rho_{y_1 y_2} = \frac{(R_{y_1 y_2})/A^2 - \mathrm{e}^{-(c_1/\Gamma_1 + c_2/\Gamma_2)}}{\frac{1}{2}\left(\mathrm{e}^{-c_1/\Gamma_1} + \mathrm{e}^{-c_2/\Gamma_2}\right) - \mathrm{e}^{-(c_1/\Gamma_1 + c_2/\Gamma_2)}}. \tag{7.6.24}$$

Some care should be taken in interpreting the limiting values, because they depend on how c_1 and c_2 approach zero and $|\rho|^2$ approaches one. We have

$$R_{y_1 y_2}\left(|\rho|^2 = 0\right) = A^2\,\mathrm{e}^{-c_1/\Gamma_1}\mathrm{e}^{-c_2/\Gamma_2} \quad (c_1, c_2 > 0)\,. \tag{7.6.25}$$

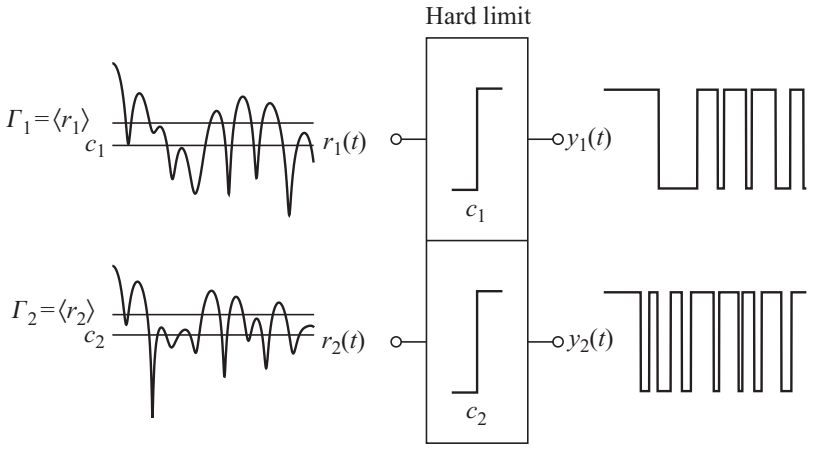


Figure 7.6.6 The hard limiting function for two Rayleigh distributed envelopes $r_1(t)$ and $r_2(t)$ with means Γ_1 and Γ_2, hard clipping levels c_1 and c_2, producing clipped signals $y_1(t)$ and $y_2(t)$. The correlation of y_1 and y_2 is the quantity of interest

Two limiting results are:

(i) for $|\rho|^2 = 1$, we get $\rho_{yy} = 1$; and
(ii) for $|\rho|^2 = 0$, we get $\rho_{yy} = 0$.

Both of these results are independent of the clipping levels. The first result can be used for establishing the variation of measurement results caused by the finite measurement interval. This is undertaken by having a common input signal, namely $r_1(t) = r_2(t)$, i.e. studying an autocorrelation. The second result cannot hold for a (finite interval) measurement since very large or very small clipping levels will produce d.c. signals which have a correlation coefficient of one instead of zero. Only for very long measurement intervals will the dissimilar signals momentarily reach a limiting threshold value, and only then will the correlation calculation (over the finite interval) result in a zero value. From a similar line of reasoning, for $c_1 = c_2$ approaching zero, we can get ρ_{yy} approaching zero independent of $|\rho|^2$, which also cannot hold in a measurement for the same reason, i.e. because of the finite measurement interval. Comparison of the graphical results below with any measured signals which are obtained while approaching these limiting values, should take the above points into account. In the graphs, we take $\rho_{yy} = 0$ for $|\rho|^2 = 0$.

For equal clipping levels, $c_1/\Gamma_1 = c_2/\Gamma_2$, a comparison is made in Figure 7.6.7 between the correlation coefficients of the unclipped power signal $|\rho|^2$, and the hard-clipped signal, with the clipping level as a parameter; ρ_{yy} drops from unity extremely quickly, especially for clipping values well away from the mean, and goes to zero for $|\rho|^2$ zero. However, for 'low' values of $|\rho|^2$ (less than about 0.8), ρ_{yy} is reasonably linear for clipping levels close to the mean. Also, the curve does not shift much for a few dB of variation of the clipping level, an encouraging practical point. For practical usage, the motivation of diversity evaluation is to experimentally establish

uncorrelated signals, which is interpreted here as signals for which the power signal or envelope correlation coefficient is below about 0.7, i.e. $|\rho|$ is less than about 0.85. Figure 7.6.7 includes the examples of a common clipping level 5 dB below the mean and 7 dB above the mean, and here the corresponding hard-clipped signal correlation coefficient varies between about 0.35 (at the extremes (i.e. +5 dB, −7 dB) of the clipping level) to about 0.5 when the clipping levels are near the mean value of the signal. For mapping back to the power correlation coefficient of unclipped signals, this gives a variation of 30 per cent, namely 0.6–0.84. (Similarly, for $|\rho|^2 = 0.9$, the variation is about 15 per cent, mapping to a range of about 0.8–0.95.) In practice, maintenance of similar clipping levels between branches is more important than holding the common clipping level relative to the mean level of the short-term fading. For practical applications, these 'errors' are on top of the statistical variation of a measurement set caused by the finite estimation interval.

There are large variations in the clipping level, however, reflecting a large available margin for variations caused by practical factors. From Section 6.5, estimation of the

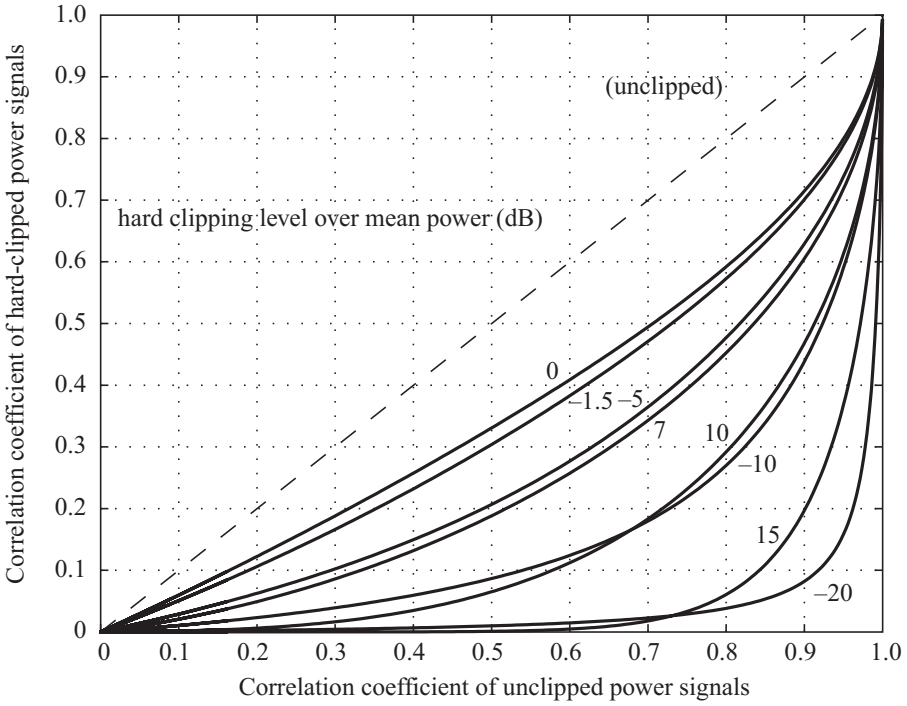


Figure 7.6.7 The correlation coefficient of the hard-clipped power signal against the correlation coefficient of the power signals (which is within 10 per cent of the envelope correlation coefficient) for equal clipping levels between diversity branches. The parameter is the common clipping level in dB relative to the unclipped signal mean. The dotted line is for one-to-one reference only

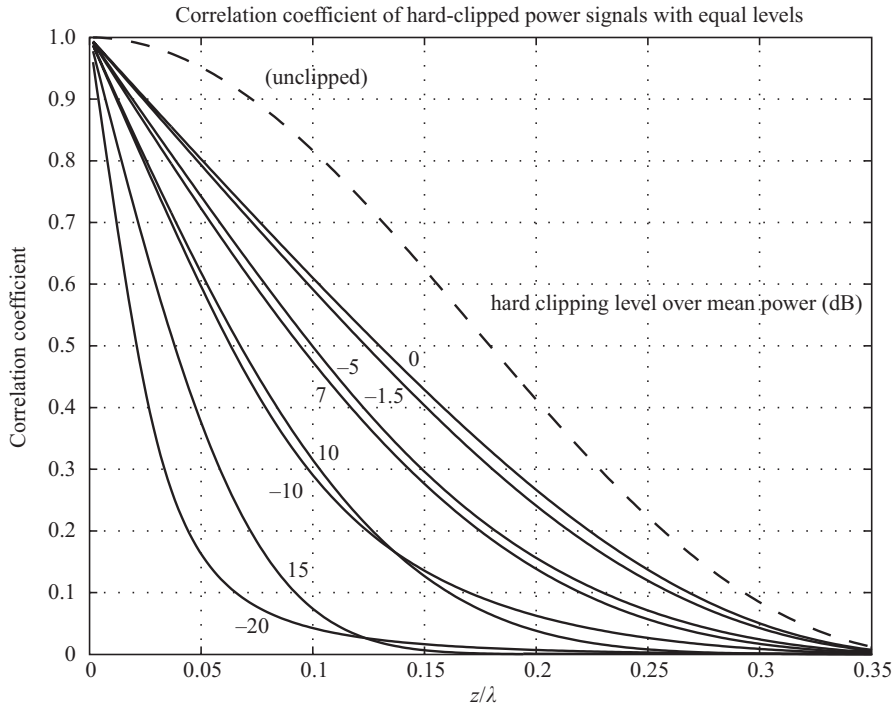


Figure 7.6.8 The correlation coefficients corresponding to ideal space diversity with omnidirectional antennas in a homogeneous uncorrelated source scenario. The parameter is the common hard-clipping level of the envelopes relative to the mean power level, e.g. 0 dB indicates that the clipping occurs at the mean power level of the signal. The dotted line is the unclipped envelope correlation coefficient

mean of a Rayleigh-distributed envelope over an interval of about 20 fades[6] gives a variation in the mean estimate of less than about ± 3 dB for a ± 2 standard deviation range (i.e. for 95 per cent of estimates if the distribution is taken as normal). So ± 3 dB seems a realistic goal to set for the practical clipping level variation. In practice, this means that the running mean should be calculated over about 20 fades and it is assumed that over this distance, the Rayleigh fading is stationary.

Finally, Figure 7.6.8 (Vaughan, 1996) is a plot of the correlation coefficient for idealised space diversity using omnidirectional antennas in a uniform azimuthal distribution of uncorrelated incident waves. The unclipped signals give $|\rho|^2 = J_0^2(\chi)$, and $\chi = kd$ where k is the wavenumber and d is the physical antenna spacing. This is the dotted line in Figure 7.6.8. The various common hard-clipping levels depicted offer a feel for the behaviour of the correlation coefficient of the hard-clipped signal as a function of the antenna spacing.

[6] For example, about ten wavelengths with omnidirectional incident power density and antenna, and a longer distance for directional antennas or directional incident power density.

7.6.2.5 Summary

For hard-clipped signals the correlation coefficient function can be similar to that with no clipping, although the similarity can be sensitive to the clipping parameters. With the clipping levels (relative to their unclipped signal means) the same in each channel, a variation of ± 3 dB in the common clipping level corresponds to only a 10 per cent change in the clipped signal correlation coefficient for power correlation coefficients of about 0.7. There is greater sensitivity to the difference in the clipping levels. For an optimum clipping level of about 1 dB above the mean in one channel, the clipping level in the other channel should not differ by more than about ± 1.5 dB to hold the accuracy to the 10 per cent figure. If these levels can be maintained in a measurement system, then the correlations of hard-clipped envelopes can be used to establish, with similar accuracy, the (unclipped) power correlation coefficients between diversity channels. Any system using bottom-only clipping or soft clipping will have better potential accuracy. For estimating correlation coefficients from measured envelopes, bottom clipping has little effect. If necessary, a measurement system should be set up to avoid top clipping at the expense of incurring bottom clipping.

7.6.3 Demeaning of envelopes

Measured envelopes contain both the short-term (Rayleigh-like) and long-term (shadow) fading components. The phrase 'short-term' can be somewhat misleading because the length, or duration, of the fade depends on the Doppler distribution of the effective scatterers. A small Doppler spread means that the Rayleigh-like fades are electrically far apart. In a limit of small Doppler spread which occurs in canyons, corridors, etc., the inverse fourth power decay (see Chapter 2) is actually an infinitely long 'short-term' fade. A measurement is often directed at characterising the short-term fading, normally for finding out about the spatial relations of the multipath and for evaluation of diversity systems which are for combating the short-term fading. The autocorrelation of the short-term component is produced from the measurement data. For compact diversity antennas, the long-term fading is essentially common to all the branches. A cross-correlation, needed for diversity evaluation, is boosted by this common signal. So for analysis of the cross-correlation of the short-term fading, the common shadow fading needs to be suppressed and this demeaning will decrease a typical cross-correlation function. In autocorrelations, where the function is a maximum at the origin and decreases away from the origin, the effect of the demeaning is more complicated and changes the form of the function. This is addressed in the remainder of the section which draws on Scott and Vaughan (1992).

The demeaning operation is to suppress the longer-term fading component so that only the short-term component remains. Figure 7.6.9 gives an example of a measured narrowband envelope as a function of position, with the local mean and the demeaned signal. The short-term fading envelope is Rayleigh-distributed and its power is exponentially distributed. The local mean of the envelope (or power) is lognormally distributed, from experimental results. Mathematically, the mean power of the Rayleigh envelope is proportional to the standard deviation of the envelope

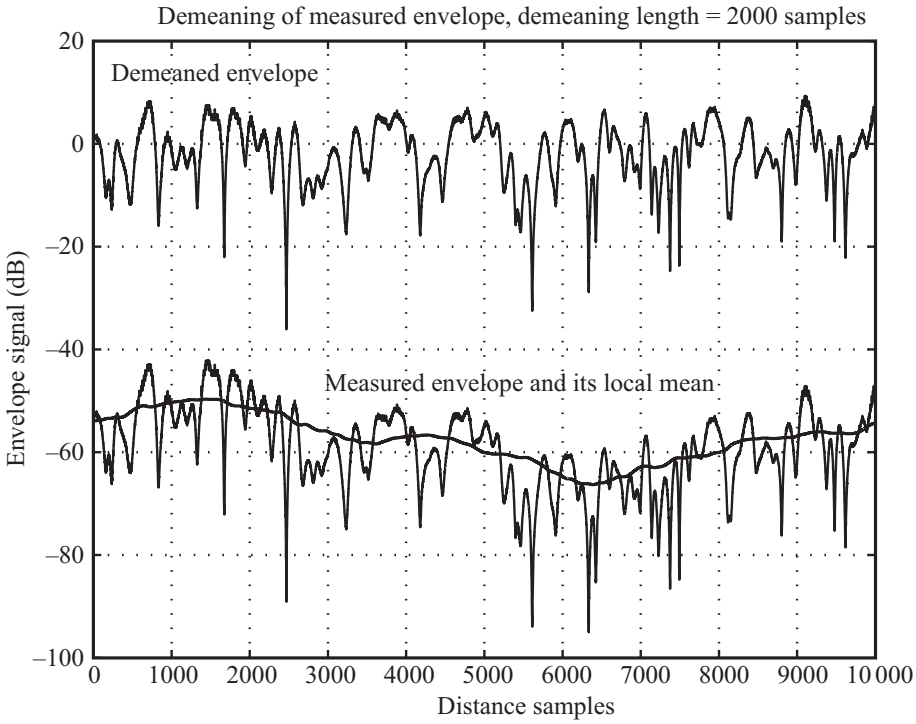


Figure 7.6.9 A section of demeaned fading signal with the measured narrowband envelope and local mean as a function of position

power,

$$\langle r^2 \rangle = \sigma_{r^2} = 2\sigma^2, \tag{7.6.26}$$

and so the lognormally distributed long-term (i.e. local mean) fading is multiplied by the Rayleigh distributed short-term envelope, to get the combined long- and short-term behaviour. This is the basis of the Suzuki distribution (see Appendix C).

In general, demeaning is complicated to implement perfectly, because in a measurement, the frequencies of both the short-term and long-term components are varying. For example, directional antennas moving (in particular, rotating) in a directional scenario, will receive signals with varying fading rates in the varying positional neighbourhoods. To simplify the situation, it is a convenient tradeoff to assume a constant, i.e. stationary, spectrum for both components. This is strictly valid for only distance-based sampling (or time-based with constant speed), and a constant set of effective sources, as will occur, for example, with omni antennas and a uniform scenario. The autocorrelation functions here are spatial, with variable Δz. The case of additive mean (see Section 7.6.3.1) is used to lead into the multiplicative mean case of the mobile channel.

Recall that for a real signal $r(z)$, the autocorrelation, $R_r(\Delta z)$, the autocovariance, $C_r(\Delta z)$, and the correlation coefficient are related by

$$\rho_r(\Delta z) = \frac{C_r(\Delta z)}{C_r(0)} = \frac{R_r(\Delta z) - \mu_r^2}{\sigma_r^2}, \tag{7.6.27}$$

where Δz is the correlation displacement, or lag, μ_r is the mean of $r(z)$, and σ_r^2 is the variance of $r(z)$. If the signal is ergodic, then the correlation is

$$R_r(\Delta z) = \int r(z)\, r(z + \Delta z)\, \mathrm{d}z, \tag{7.6.28}$$

where the integral is over $\pm\infty$. For finite record lengths, this integration interval reduces to be equal to or less than the signal record length, depending on the displacement. A finite interval gives only an estimate of a correlation, because the signals are of greater duration than the interval.

7.6.3.1 Additive mean

For the short-term signal of interest (one which is demeaned), $r_\mathrm{d}(z)$, with an additive long-term component (the local mean), $m(z)$, i.e.

$$r(z) = r_\mathrm{d}(z) + m(z), \tag{7.6.29}$$

demeaning involves the calculation of $m(z)$ and its subtraction from the measured signal, which then gives the wanted short-term signal,

$$r_\mathrm{d}(z) = r(z) - m(z). \tag{7.6.30}$$

The local mean is

$$m(z) = \frac{1}{2Z} \int_{-Z}^{Z} r(z + \chi)\, \mathrm{d}\chi, \tag{7.6.31}$$

where $2Z$ is the window length.

An approach to finding the effect of the window size on the autocorrelation is to consider the demeaning operation as a lowpass filter, with impulse response $w_\mathrm{LP}(z)$, i.e.

$$m(z) = r(z) * w_\mathrm{LP}(z) \tag{7.6.32}$$

where $*$ is convolution and

$$\begin{aligned} w_\mathrm{LP}(z) &= 1 \quad \text{for } -Z \le z \le Z \\ &= 0 \quad \text{elsewhere.} \end{aligned} \tag{7.6.33}$$

It follows that $\hat{r}_\mathrm{d}(z)$ may be written as a highpass filtering of $r(z)$,

$$r_\mathrm{d}(z) = r(z) * w_\mathrm{HP}(z), \tag{7.6.34}$$

where the equivalent highpass filter is found from Fourier transformation as

$$w_\mathrm{HP}(z) = \delta(z) - w_\mathrm{LP}(z). \tag{7.6.35}$$

The autocorrelation of the demeaned signal and the measured signal can be related by (see Appendix B)

$$\begin{aligned} R_{r_\mathrm{d}}(\Delta z) &= R_r(\Delta z) * R_{w_\mathrm{HP}}(\Delta z) \\ &= R_r(\Delta z) * \left(\delta(\Delta z) + R_{w_\mathrm{LP}}(\Delta z) - 2w_\mathrm{LP}(\Delta z)\right) \\ &= R_r(\Delta z) + R_r(\Delta z) * \left(R_{w_\mathrm{LP}}(\Delta z) - 2w_\mathrm{LP}(\Delta z)\right) \end{aligned} \tag{7.6.36}$$

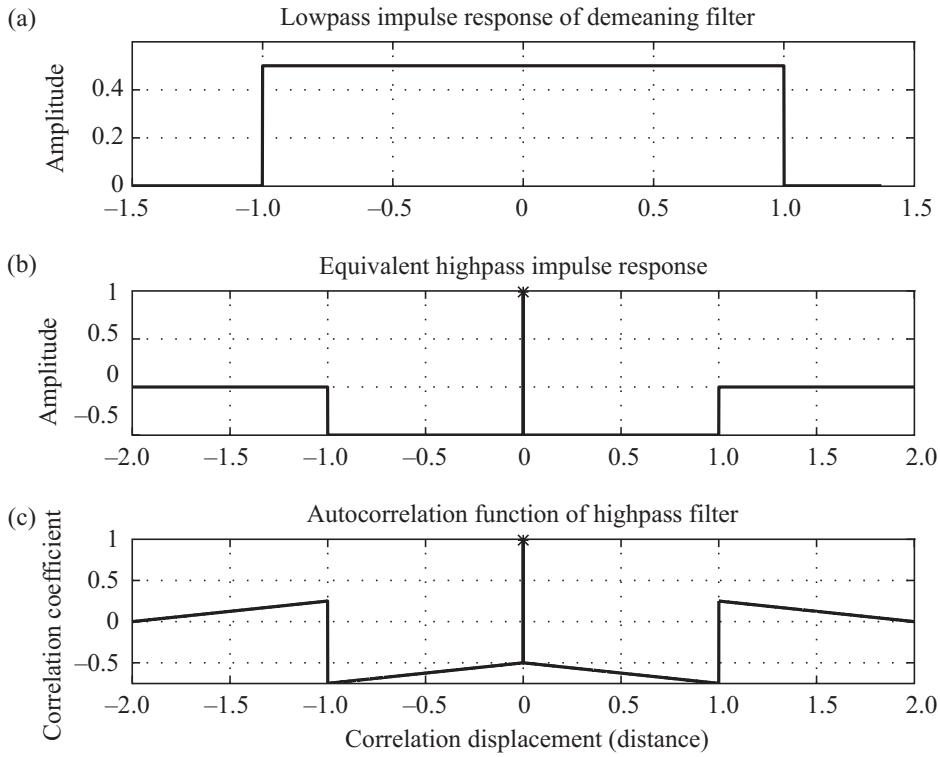


Figure 7.6.10 *(a) The lowpass filter of the demeaning operation for the additive case. (b) The equivalent highpass filter. (c) The autocorrelation of the highpass filter which in the demeaning operation convolves with the autocorrelation of the measured signal to produce the autocorrelation of the demeaned signal. The line with an asterisk on top indicates a delta function*

and if the demeaned signal is taken as zero-mean, then the autocorrelation coefficient function is just

$$\rho_{r_\mathrm{d}}(\Delta z) = \frac{R_{r_\mathrm{d}}(\Delta z)}{R_{r_\mathrm{d}}(0)}. \tag{7.6.37}$$

So the autocorrelation coefficient function of the demeaned signal can be calculated directly from the autocorrelation function of the measured, undemeaned signal. Figure 7.6.10 (Scott and Vaughan, 1992) shows a lowpass filter response, $w_{\mathrm{LP}}(z)$, the equivalent highpass response $w_{\mathrm{HP}}(z)$, and its autocorrelation function $R_{w_{\mathrm{HP}}}(\Delta z)$. All demeaning filters have this, or a similar, highpass form.

7.6.3.2 Multiplicative mean

In the mobile channel, the short-term signal has a multiplicative long-term variation as described above. This situation is more complicated than the additive case but we can call on the additive results to show that the multiplicative demeaning has the same effect. The measured signal is a product,

$$r(z) = r_\mathrm{d}(z)\, m(z), \tag{7.6.38}$$

so the demeaning operation is

$$r_{\mathrm{d}}(z) = \frac{r(z)}{m(z)}. \tag{7.6.39}$$

This can be converted to the additive case by taking the log (normalisation is understood),

$$\begin{aligned} \log r_{\mathrm{d}}(z) &= \log r(z) - \frac{1}{2Z}\int_{-Z}^{Z} \log r(z+\chi)\,\mathrm{d}\chi \\ &= w_{\mathrm{HP}}(z) * \log r(z), \end{aligned} \tag{7.6.40}$$

which shows the similarity to the additive case. For Rayleigh-distributed signals, the envelope autocorrelation is very similar to that of the log envelope (De Neumann, 1989). Because of this, and the fact that the mean $m(z)$ is uncorrelated with the envelope, the use of the log makes essentially no difference to the autocorrelation behaviour.

The effect of the demeaning on the measured correlation is to decrease the width of the main lobe, and in so doing decrease the first minima and increase the first maxima. This is demonstrated in Figure 7.6.11, in which a typical autocorrelation function of

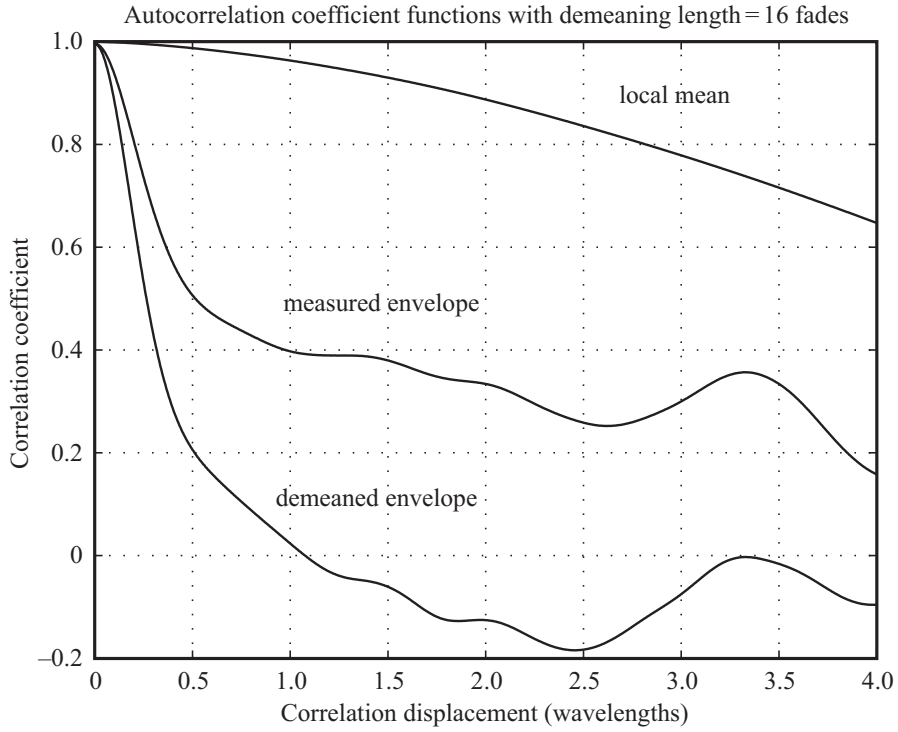


Figure 7.6.11 Autocorrelation coefficient functions of the envelopes before and after demeaning, and of the demeaning signal, for the measured signals of Figure 7.6.9. For the envelopes, the main lobe is narrower and the near maxima and minima are altered as a result of the demeaning

a measured envelope and the correlation after demeaning are shown. Near-in maxima and minima have been created in this example. The correlation distance to (say) $\rho_r = 0.7$ decreases from $\Delta z = 0.27\lambda$ to 0.17λ. This difference is significant when seeking or evaluating compact space diversity designs.

In practice, the demeaning window should be between five and twenty average fade lengths. This is a tradeoff between affecting the short-term fading (at five fades) and not removing sufficient long-term effects. For any given measurement, the demeaning details should accompany all presented data, since the demeaning operation contributes to the important derived quantities such as the decorrelation distance.

7.6.3.3 Summary

The demeaning operation is to remove the shadow fading component of an experimentally measured envelope but the operation also affects the short-term envelope as well. The choice of demeaning distance has an impact on the autocorrelation function, so if the demeaning is used in the processing of data, the demeaning distance should be documented along with other important parameters of the measurement.

7.7 Wideband channels: channel sounding

For wideband signals suffering multipath impairment, the total power in the band varies less than for the narrowband case, and the instantaneous frequency selective fading, or its 'instantaneous statistics' such as the delay spread at a point in space, becomes the focus. Therefore a different measurement technique, called channel sounding, after radar and sonar developments, is used for wideband signals. This involves applying a pulse to the channel in an efficient way and processing the response to find the transfer function. The pulse can take many forms, ranging from a basic rectangular shape to more sophisticated forms such as a maximum length digital sequence or a swept frequency chirp. The advantage of these more sophisticated forms over the rectangular pulse is that they can sound the spectrum channel more fully and the peak powers can be reduced, resulting in a better channel sounding efficiency and better resolution of multipath components.

For frequency selective channels, the channel characterisation is possible by either a continuous wave (cw) frequency sweep, which gives directly the transfer function, or by sounding the channel with pulses which results in a measure of the impulse response. When the sounding pulse becomes a chirp-type waveform, the two approaches are distinguished more by implementation rather than by concept. Simple rectangular pulses can be used for partial, but usually sufficient, channel characterisation. The cases for chirp pulse and rectangular pulse sounding are discussed.

In digital communications, the modulated signal itself is often used to sound the channel. For simple modulations, a $\sin x/x$ type of spectrum results, normally with the first spectral nulls at the band edges. Such signals do not fully occupy the available spectrum. Using steeper roll-off filters or the more sophisticated multicarrier

modulations such as OFDM creates signals with more efficient usage of the available spectrum. The use of single pulses to sound the channel fully offers some interesting options for quick and efficient channel sounding. The pulse shapes are complicated but these are normally calculated once and stored for real-time usage. An example of such a sophisticated pulse, used for television delay profile estimation and ghosting correction (Koo, 1993), is given in Section 7.7.1. In Section 7.7.2, signal processing to super-resolve simpler pulses is used for estimating the delay profile. This section draws from Vaughan and Scott (1999).

7.7.1 Full channel sounding

For a given channel bandwidth, full channel sounding can only be achieved by using a pulse which essentially occupies the whole of the channel spectrum. A maximum length sequence sounder is often used and this typically has a pseudo-random series of pulses (chirps) whose central spectral lobe approximates the desired flat spectrum. This type of sounding is well documented in the literature (for example, Parsons and Gardiner, 1989; Parsons, 1992; Steele, 1992). A similar, but more direct and efficient (in the signal theory sense) approach, is to use the classical frequency modulated chirp. This can be viewed as a single frequency sounder which sweeps its frequency though the channel band. Here we follow the method of Koo (1993, 1995) (see also USATSC, 1992, 1993; Gardiner, 1993, 1994): a basic linear FM chirp $x(t) \Leftrightarrow X(\omega)$ ($\Leftrightarrow$ indicates the Fourier transform) is modified so that its power spectrum is essentially flat and confined to the channel bandwidth. The modification is by applying a tradeoff between the requirements in the time (finite duration chirp) and frequency (flat and compact support) domains. This tradeoff is in the form of a window, denoted $w(t) \Leftrightarrow W(\omega)$, which is typically a raised cosine or Tukey window. The chirp can be made real by constraining its spectrum to be Hermitian (see Equation 7.7.3). The Hermitian spectrum of a chirp denoted $x(t) \Leftrightarrow X(\omega)$ has the form (by definition – see below) of $X(\omega) = e^{-jsgn(\omega)b\omega^2}$, and so the spectrum of the modified, or windowed, chirp is

$$X(\omega)W(\omega) = \mathrm{e}^{-\mathrm{j}\,\mathrm{sgn}(\omega)b\omega^2} W(\omega) \tag{7.7.1}$$

in which b is a constant. Fourier transformation then gives the highly efficient sounding pulse and this is stored digitally within the communications system.

A basic linear FM chirp in time t is complex. The basic chirp and its spectrum are denoted (Papoulis, 1977)

$$\tilde{x}(t) = \mathrm{e}^{\mathrm{j}\pi kt^2} \quad \Leftrightarrow \quad \tilde{X}(\omega) = \frac{1}{\sqrt{k}} \mathrm{e}^{\mathrm{j}\pi/4} \mathrm{e}^{-\mathrm{j}(\pi/k)(\omega/(2\pi))^2}. \tag{7.7.2}$$

This signal is seen to have an instantaneous frequency of kt Hz, and the frequency sweep rate is k Hz s^{-1}. Its spectrum is seen to be also a complex chirp which has a sweep rate of $-1/k$ s Hz^{-1}.

This basic chirp is modified to be purely real, denoted $x(t)$, by defining it in the frequency domain and transforming. Following Koo, the positive and negative frequencies of the spectral chirp are separated and the one part is conjugated to produce a Hermitian spectrum, which means that the time chirp becomes purely

real. The spectrum is also filtered by using the real, symmetric window $W(\omega)$. The modified chirp can now be written in the form (cf. Koo, 1993)

$$x(t) = \frac{1}{2\pi}\int_0^\infty X(\omega)W(\omega)e^{j\omega t}d\omega \tag{7.7.3}$$

$$= \frac{1}{2\pi}\int_0^\infty \mathrm{e}^{-\mathrm{j}\frac{\pi}{k}\left(\frac{\omega}{2\pi}\right)^2} W(\omega)\mathrm{e}^{\mathrm{j}\omega t}\,\mathrm{d}\omega + \frac{1}{2\pi}\int_{-\infty}^0 \mathrm{e}^{\mathrm{j}\frac{\pi}{k}\left(\frac{\omega}{2\pi}\right)^2} W(\omega)\mathrm{e}^{\mathrm{j}\omega t}\,\mathrm{d}\omega.$$

In a computer calculation of the chirp samples, the infinite limits are replaced with a finite spectral support value. The support is made sufficiently greater than the filter cut-off for the energy outside the support to be negligible. This chirp sweeps from d.c. towards the cutoff frequency of the filter $W(\omega)$.

The filter, or window, is defined in the time-domain and transformed to get the frequency domain expression. Using a raised cosine of frequency c which is limited to a half 'cycle' by a rectangle function $p_{\pi/c}(t)$, the window, in the time and frequency domains, is (Koo, 1993; Papoulis, 1977)

$$w(t) = p_{\pi/c}(t)\left(\frac{1}{2} + \frac{1}{2}\cos ct\right) \quad \Leftrightarrow \quad W(\omega) = \frac{\pi^2 \sin\left(\frac{\pi}{c}\omega\right)}{\omega\left(\pi^2 - \left(\frac{\pi}{c}\right)^2\omega^2\right)} \tag{7.7.4}$$

The process of calculating the chirp can be viewed as an iterative process which trades off the finite duration and finite bandwidth with a flat response across the channel bandwidth. Here, only one iteration is used and this provides an acceptable solution for most applications. Many iterations may improve the pulse, depending on the complexity required, and this type of algorithm originated for solving the Fourier phase problem (Gerchberg and Saxton, 1972).

As an example, the ghost cancellation pulse used for PAL television is real and has a spectrum which is flat to about 4.8 MHz, and is more than 25 dB down outside about 5.2 MHz. This has an extra term in the window to cater for the finite support of the Fourier integral (see Koo, 1993). It is shown in Figure 7.7.1 (cf. Koo, 1993; Gardner, 1993; USATSC, 1993). I. Goodwin is acknowledged for providing information regarding the ghost cancellation pulse for television, and M. Poletti is acknowledged for discussions and assistance with the pulse presented here and its calculation.

Figure 7.7.2 depicts the channel sounder operation. The output from the static channel is the convolution of the channel impulse response $h(t)$ and the chirp $x(t)$. To process the received chirp, we first multiply by the conjugate of the chirp (i.e. just the chirp if it is real) and then convolve with the chirp. The convolution peaks at the delays of the multipath components since the chirp has the required property $x(t) * x^*(t) = \delta(t)$. (The $*$ denotes convolution.) The measured impulse response is proportional to

$$h_\mathrm{m}(t) = \big((h(t) * x(t)) \cdot x^*(t)\big) * x(t). \tag{7.7.5}$$

The first convolution operation is the channel being sounded and the second two operations represent the receiver processing. The order of this processing – the multiplication before the convolution – may be important because it can remove any constraint on the sweep rate otherwise imposed by the bandwidth of the channel (Bold *et al.*, 1981; Poletti, 1988).

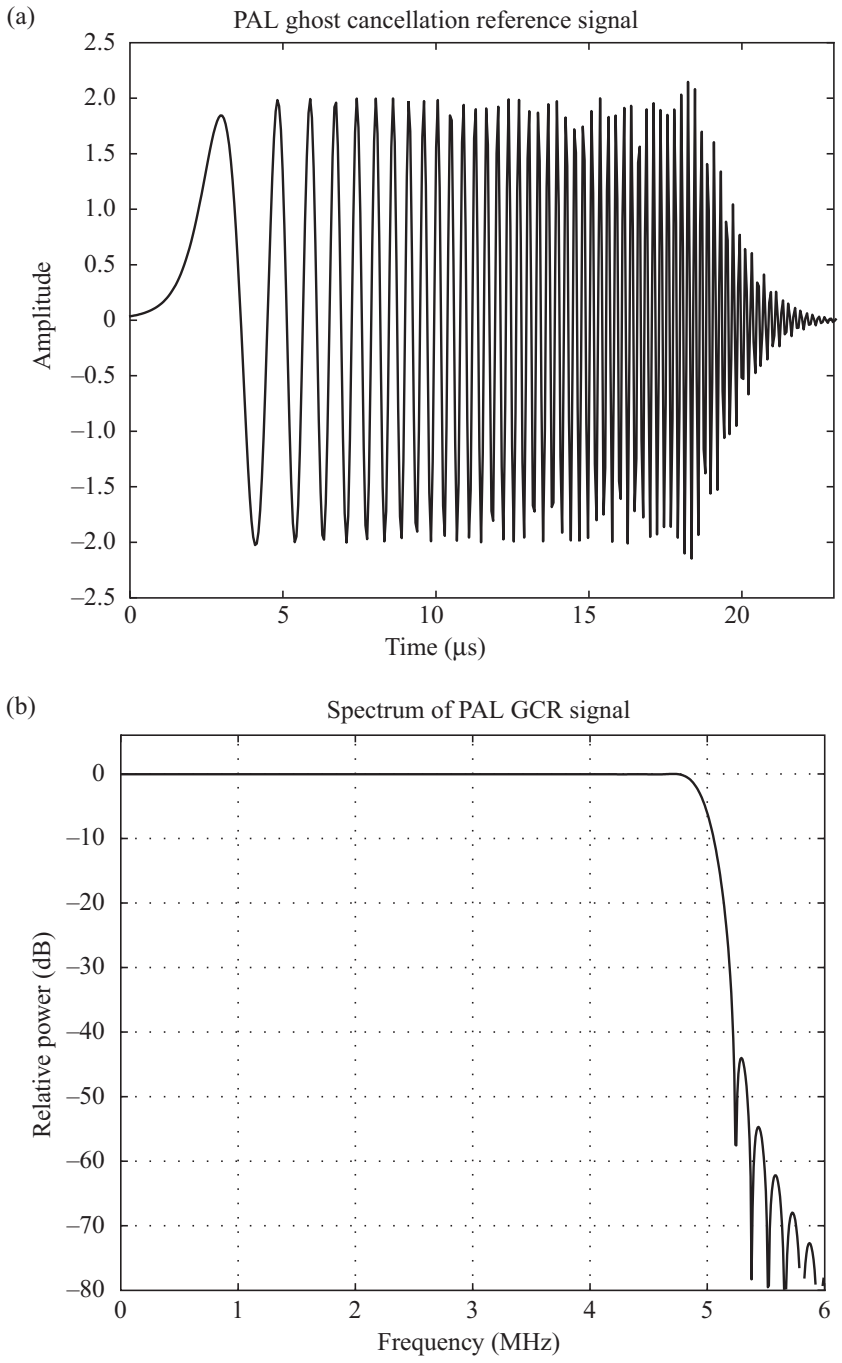


Figure 7.7.1 *(a) A chirp and (b) its spectrum which is used for ghost cancellation in PAL broadcast television. The sweep rate is $k = 2.813(10)^{11}\ Hz\,s^{-1}$, the duration is about 24 μs and the band is flat from 0 to about 4.8 MHz*

7.7.2 Partial channel sounding by rectangular pulse

Partial channel sounding using a rectangular pulse is simpler because the waveform is simpler to generate and process. If the channel behaviour can be considered wide-sense stationary in frequency (uncorrelated in delay time), then partial sounding can

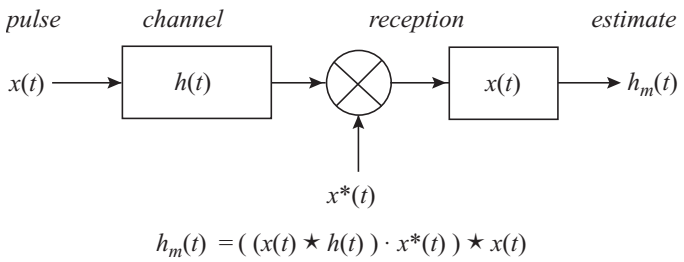


Figure 7.7.2 Channel sounding by linear FM chirp pulse. Mathematically, the sounding pulse is convolved with the channel and the reception operation is multiplying by the conjugate of the pulse and then convolving with the pulse

provide statistical quantities representative of the full channel. Furthermore, if it can be assumed that the real part and imaginary part of the varying impulse response have the same statistics, then there is need to measure only one of these parts, which will then provide statistical characterisation representative of the complex channel.

The channel is sounded by the real rectangular pulse $p_{\mathrm{T}}(t)$ which may have a duration greater than the Fourier bandwidth-limited resolution. One motivation for this longer duration is to keep the peak amplitude, relative to the mean amplitude, at a low as possible level. This minimises the distortion from the nonlinear effects of the electronics of the transmitter and receiver. A shorter pulse, with the same average energy, requires a higher amplitude to maintain the SNR. A measure of this potential distortion is the crest factor, or peak-to-average voltage of the pulse (Bold *et al.*, 1981). In this context, the rectangular pulse has a disadvantage relative to the modified chirp pulse, or to a maximum length sequence pulse. The spectrum of the rectangular pulse has a sinc form, and so the channel is only partially sounded, in general. This case is the same as for classical radar.

7.7.2.1 Spectral occupation efficiency

To discuss 'how partially' the channel is being sounded, we require a measure of the 'spectral occupation efficiency', which is effectively how similar the spectra of the pulse and the channel are, over the bandwidth of interest. For a channel with spectral support given by $g_1(\omega)$ and a pulse power spectrum given by $g_2(\omega)$, the spectral occupation efficiency be defined by (cf. a matched filter)

$$\eta_{\mathrm{soe}} = \frac{2\int g_1(\omega)g_2(\omega)\,\mathrm{d}\omega}{\int \left(g_1^2(\omega) + g_2^2(\omega)\right)\,\mathrm{d}\omega}, \tag{7.7.6}$$

where g_1 and g_2 are real. The integrals are over the channel band. Note that this efficiency is independent of the energy efficiency of the pulse (see Section 7.7.2.2), so good spectral occupation efficiency does not imply efficient energy utilisation of the pulse. In fact, these efficiencies have opposing behaviour. For a channel of bandwidth

B rad s^{-1} and unity power spectral density,

$$\begin{aligned} g_1 &= 1 \quad -B/2 < \omega < B/2, \\ &= 0 \quad \text{elsewhere}, \end{aligned} \tag{7.7.7}$$

and for an isolated rectangular pulse of width T,

$$g_2 = T^2 \left(\sin(\omega T/2)/(\omega T/2)\right)^2; \tag{7.7.8}$$

then

$$\eta_{\text{soe}} = \frac{2\left(\text{Si}(BT/2) - (BT/4)\left((\sin BT/4)/(BT/4)\right)^2\right)}{\sqrt{BT\int_0^{BT/4}(\sin x/x)^4\ \mathrm{d}x}}, \tag{7.7.9}$$

where the sine integral is

$$\text{Si}(z) = \int_0^z \frac{\sin t}{t}\,\mathrm{d}t. \tag{7.7.10}$$

This efficiency measure is plotted in Figure 7.7.3 in which B is expressed in Hz (Vaughan and Scott, 1999) BT is dimensionless) instead of radians, for simple interpretation.

7.7.2.2 Energy efficiency

The energy efficiency of the pulse is the fraction of its power in the band,

$$\eta_{\text{e}} = \frac{\int_{-B/2}^{B/2} g_2^2\,\mathrm{d}x}{\int_{-\infty}^{\infty} g_2^2\,\mathrm{d}x} = \frac{3}{\pi}\int_0^{BT/4}\left(\frac{\sin x}{x}\right)^4\,\mathrm{d}x, \tag{7.7.11}$$

which is also plotted in Figure 7.7.3, and the product of the two efficiencies is included.

7.7.2.3 Product of efficiencies

The product, $\eta_{\text{soe}}\eta_{\text{e}}$, has a maximum at $B_{\text{Hz}}T \sim 1$ (B_{Hz} is the bandwidth in Hz). For pulses $T < \sim 1/B_{\text{Hz}}$, the efficiency is dominated by the energy term and for $T > \sim 1/B_{\text{Hz}}$, the spectral occupancy term dominates. If signal-to-noise requirements call for an energy efficiency of 0.5, for example, then the pulse duration should be $T > \sim 0.4/B_{\text{Hz}}$ (in seconds). The resolution in time is restricted to this value unless extra processing, in particular deconvolution, is applied to the data.

7.7.3 Super-resolution by deconvolution

The impulse response of the propagation channel can be modelled as a summation of effective scatterers,

$$h(\tau) = \sum_i^I a_i \delta(t - \tau_i), \tag{7.7.12}$$

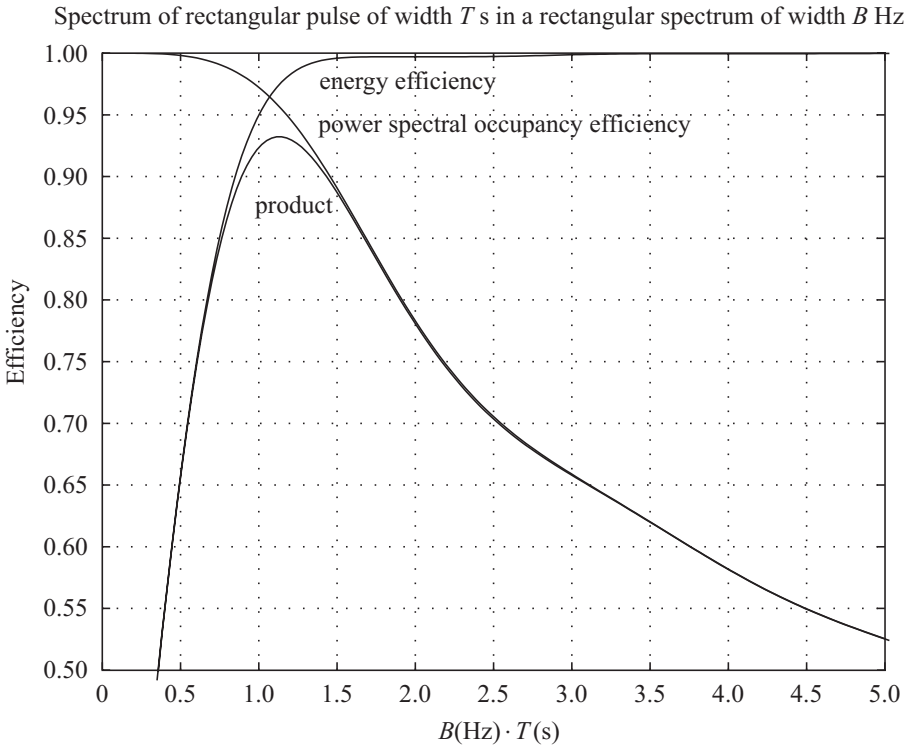


Figure 7.7.3 The efficiencies of the sounding for a rectangular channel of bandwidth B Hz and a rectangular pulse of duration T. The spectral occupation efficiency is how well the sounding pulse spectrum is matched to the channel spectrum. The energy efficiency is the proportion of pulse energy in the channel, and the product gives the overall channel sounding efficiency

where a_i are the complex magnitudes of I effective scatterers. (They are referred to as effective scatterers because the weighting of the antenna pattern is included.) This modelling step must be interpreted carefully. The liberties taken by using the delta-function give the impression of an infinite bandwidth. The propagation channel (antennas *not* included, the channel output quantities are fields) gives a reasonable approximation to this for practical, narrow relative bandwidth, channel modelling purposes, although there is little physical foundation from general scattering for such a model over a wide relative bandwidth. Any realisable antenna, which explicitly appears in the signal model already, must impose severe relative bandwidth restrictions, and also its pattern is probably changing over a wide bandwidth. Implicit bandlimiting also occurs in other components of the transmitting and receiving system. For a modelled impulse response, the delta functions become sinc functions for an ideal linear-phase rectangular bandwidth imposed by the antenna and receiving electronics, for example. This bandlimiting thus combines the propagation channel multipath components so that resolving them is not possible in the general case (see Section 6.1.3). However, if the excitation waveform is known, and there is negligible dispersion over the bandwidth from the individual scatterers, then deconvolution

techniques can resolve the a_i so that the sum of delta functions can fairly represent the channel for most communications purposes. The power delay profile of the propagation channel is then given by the discretised model, with components potentially resolved more finely than both the pulse duration and the reciprocal of the channel bandwidth. The delay spread of the propagation channel follows. The deconvolution needs to use a pulse which is the convolution of the ideal sounding pulse with the impulse responses of the transmit and receive antennas (and front ends). This deconvolving pulse is measured in experiments using a propagation channel which is well-behaved, i.e. has negligible multipath ($I = 1$).

7.7.3.1 Deconvolution

Two methods for the deconvolution are subtractive deconvolution and a modified inverse filter.

The estimate of the impulse response, $\hat{h}(\tau)$, is sought. The terminology follows that of the original literature (Högbom, 1974). The clean pulse, $\tilde{p}(\tau)$, is a bandlimited and distorted (by the receiver and the channel) form of the sounding pulse $p(\tau)$, and is the same for all the responses for a given channel. By defining the clean pulse this way, the distortion introduced by the receiver is removed during the deconvolution.

7.7.3.2 Subtractive deconvolution

The subtractive deconvolution algorithm used here is (cf. Högbom, 1974; Bates and McDonnell, 1986):

(i) initialise the dirty map $d(\tau) = r(\tau)$ and the clean map $c(\tau) = 0$;
(ii) form the correlation coefficient function $\rho(\tau) = \tilde{p}(\tau) \odot \mathrm{d}(\tau)$ (normalisation omitted and $\odot$ means correlation);
(iii) find peaks ρ_i and their positions τ_i in $\rho(\tau)$;
(iv) if all $\rho_i <$ threshold, stop;
(v) clean the dirty map by $\mathrm{d}(\tau) = \mathrm{d}(\tau) - \rho_i\, \tilde{p}(\tau - \tau_i)$;
(vi) update the clean map by $c(\tau) = c(\tau) + \rho_i \delta(\tau - \tau_i)$;
(vii) continue until stop;
(viii) the impulse response estimate is then $\hat{h}(\tau) = c(\tau)$.

Note in this version that the loop gain of the algorithm is unity and that this cannot cause instability because the correlation function is always kept normalised.

7.7.3.3 Inverse filter

In the spectral domain, the algorithm is simpler. The incomplete spectral occupancy of the pulse and any very deep fades in the channel are the complication. An algorithm which gets around this complication, is ($\Leftrightarrow$ denotes Fourier transform):

(i) form $R(\omega) \Leftrightarrow r(\tau)$ and $\tilde{P}(\omega) \Leftrightarrow \tilde{p}(\tau)$;
(ii) numerically tag the frequencies where $\tilde{P}(\omega)$ is small;
(iii) form $\hat{H}(\omega) = R(\omega)/\tilde{P}(\omega)$, except at the tagged frequencies, at which $\hat{H}(\omega)$ is set to zero;
(iv) form $\hat{h}(\tau) \Leftrightarrow \hat{H}(\omega)$.

This algorithm works reasonably well because the neglected (tagged) parts of $H(\omega)$ comprise relatively low energy.

The above algorithms can be checked during development using synthetic, noiseless data. Such a check verifies the algorithms and also offers a feel for their behaviour. In this absence of noise, the inverse filter performs better than subtractive deconvolution. But with experimental data, the inverse filter is susceptible to high noise levels. Specifically, several near-coincidental and opposing pulses can be generated to model what is simply a single impulse plus noise. The subtractive deconvolution identifies impulses directly from the experimental data, so this technique is not prone to the pitfall of inventing such near-coincidental impulses. The finite amplitude threshold of the reduced set of impulses (resulting from the presence of noise) is typically the limiting criterion here to deconvolution accuracy.

7.7.3.4 Instantaneous delay spread

The delay spread is calculated from the measured power delay profile,

$$P(\tau) = \langle |h(\tau)|^2 \rangle, \tag{7.7.13}$$

where the expectation is over the phase-mixing process of the addition of the incident wave contributions in the antenna. For the propagation channel, $h(t)$ is the discretised, or clean, form, produced by the deconvolution. Implementing the expectation can be in the form of spatial averaging, where the delay profile is averaged from its measurement at different positions, z, written $P_z(\tau) = \langle P(\tau, z) \rangle_z$. Alternatives include averaging over frequency for wideband systems, i.e. $P_\omega(\tau) = \langle P(\tau, \omega) \rangle_\omega$, etc. In the spatial averaging there are other alternatives, namely using the antenna pattern. In practice, the two techniques should give compatible results.

The first technique is spatial averaging which uses an omnidirectional antenna. The antenna can be moved in a circle, for example, about the spatial position of the measurement. The radius of the circle is $r_{\text{cir}} = n\beta\lambda/(2\pi)$, where n is the number of measurement positions, and β is the distance in wavelengths between the measurements. For example, in a Clarke scenario which has spatial autocorrelation $\rho_r \approx J_0(k_C d)$, $\beta \approx 0.2$ gives a spatial envelope correlation of about 0.7 between the positions. At a frequency of 1.8 GHz, a radius of 0.53 m gives 50 samples, each spaced by 0.2 wavelengths and having an envelope correlation of 0.7 between adjacent positions. The n measured power profiles are averaged and then the calculated delay spread represents the properly averaged value over the locus of the measurement (in this case a circle). The second technique uses a directional antenna which is rotated about a point in space which is the single spatial position of the measurement. In a Clarke uniform scenario, the beam displacement should be about half of the half power beamwidth (HPBW) for an envelope correlation of 0.7 between adjacent sample directions (see Chapter 8). The number of samples then gives the required directivity. For example, an HPBW of 0.126 radians gives about $2\pi/\text{HPBW}/2 \approx 50$ sample directions which will have envelope correlations of 0.7 between adjacent directions, in a Clarke scenario. The omnidirectional antenna technique offers the advantage of a lower measurement uncertainty because the wave contributions from all directions are

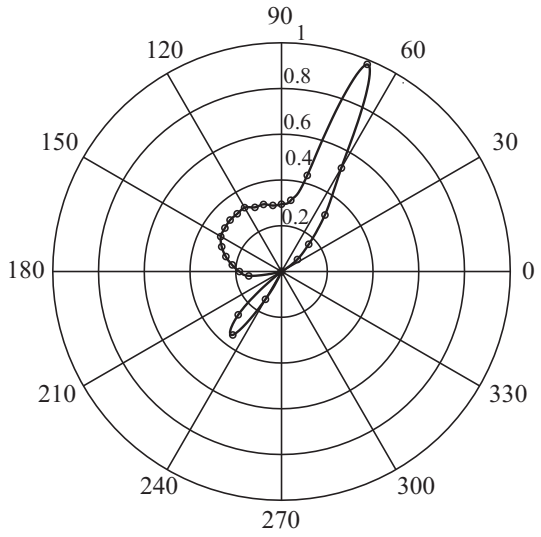


Figure 7.7.4 An example of the instantaneous delay spread as a function of the changing azimuth direction of the antenna

sampled at each spatial position. The directional antenna approach has the advantage of producing a delay spread which occurs at a well-defined point in space. However, there is still some spatial averaging at work in the sense that the directional antenna has a non-zero effective cross-section. If the results of the directional antenna and the omni antenna approaches are to be compared, then the two antenna polarisations must be the same. Simulations are easily produced to confirm the similarity (they are not identical) of the delay spread estimates in given scenarios. In general multipath conditions, the delay spread will be different for different polarisations. The directional antenna approach has a practical advantage in this sense because its polarisation can be changed conveniently, for example by rotating the antenna about its directional axis by 90° in the case of linearly polarised antennas.

An example of instantaneous delay spread measurement results (Vaughan and Scott, 1996) is given in Figure 7.7.4. This shows the instantaneous (unaveraged) delay spread variation as the receiving antenna pattern is changed, here by antenna rotation in azimuth. In the example, the instantaneous delay spread peaks to about 1 μs at about 70°. It is worth emphasising here that the delay spread cannot be calculated by averaging these instantaneous delay spreads. The instantaneous power delay profiles must be first averaged, and then the delay spread calculated.

7.8 References

ABDI, A., WILLS, K., BARGER, H.W., ALOUINI, M.-S., and KAVEH, M.: 'Comparison of the level crossing rate and average fade duration of Rayleigh, Rice, and Nakagami fading models with mobile channel data', *Proceedings of the IEEE Vehicular Technology Conference*, (VTC2000/Fall), Boston, September 2000.

AKAIKE, H.: 'A new look at the statistical model identification', *IEEE Transactions*, 1974, **AC-19** (6), pp. 716–723.

AL-JALILI, Y.O.: 'Analysis and detection algorithms for complex time zeros of bandlimited functions', *Proceedings of the IEE*, Part I, June 1991, **138**, pp. 189–200.

ANDERSEN, J. BACH, JENSEN, J., JENSEN, S.H., and FREDERIKSEN, F.: 'Prediction of future fading based on past measurements', *Proceedings of the Fall Vehicular Technology Conference*, Delft, The Netherlands, September 1999.

ANDERSON, T.W.: *An Introduction to Multivariate Statistical Analysis*, Wiley, 1958.

BALABANIAN, N. and LE PAGE, W.R.: 'What is a minimum-phase network?', *IEEE Trans. Comm. Electronics*, January 1956, pp. 785–787.

BATES, R.H.T. and McDONNELL, M.J.: *Image Restoration and Reconstruction*, Oxford Engineering Science Series, Clarendon Press, Oxford, 1986.

BOLD, G.E.J., JOHNS, M.D., and BIRDSALL, T.G.: 'Signals for optimum bandlimited system interrogation with application to underwater acoustics', *Journal of the Acoustic Society of America*, 1981, **81** (4), pp. 991–999.

BOX, G.E.P. and MULLER, M.E.: 'A note on the generation of random normal derivates', *American Mathematical Statistics*, 1958, **29**, pp. 610–611.

CLARK, M.V.: *Diversity and Equalization*, PhD Thesis, University of Canterbury, New Zealand, 1992.

CLARKE, R.H.: 'A statistical theory for mobile radio communication', *Bell System Technical Journal*, July 1968, **47**, pp. 957–1000.

DAVENPORT, W.B. and ROOT, W.L.: *An Introduction to the Theory of Random Signals and Noise*, McGraw-Hill, 1958; also reprinted by the IEEE, 1987.

DE NEUMANN, B.: 'Envelope correlation coefficient for logarithmic diversity receivers', *Electronic Letters*, 30th March 1989, **25** (7), pp. 459–460.

DONALDSON, B.P., FATTOUCHE, M., and DONALDSON, R.W.: 'Characterization of in-building UHF wireless radio channels using spectral energy measurements', *IEEE Transactions*, January 1996, **AP-44** (1), pp. 80–86.

DUEL-HALLEN, A., HALLEN, H., HU, S., and HALLEN, H.: 'Long range prediction of fading signals', *IEEE Signal Processing Magazine*, May 2000, **17** (3), pp. 62–75.

EGGERS, P.: 'Note on the sensitivity of envelope correlations', COST231, *Technical Document* (93)-107, Limerick, July 1993.

EKMAN, T.: *Prediction of Mobile Radio Channels*, PhD Thesis, Uppsala University, Sweden, November 2000.

EYCEOZ, T., DUEL-HALLEN, A., and HALLEN, H.: 'Deterministic channel modelling and long range prediction of fast-fading mobile radio channels', *IEEE Communications Letters*, September 1998, **2**, pp. 254–256.

FISHER, R.A.: 'On the probable error of a coefficient of correlation deduced from a small sample', *Metron.*, 1921, **1** (4), pp. 3–32.

FLEURY, B., TSCHUDIN, M., HEDDERGOTT, R., DALHAUS, D., and PEDERSEN, K.I.: 'Channel parameter estimation in mobile environments using the SAGE algorithm', *IEEE Journal on Selected Areas in Communications*, March 1999, **17**, pp. 434 and 450.

GANS, M.J.: 'A power spectral theory of propagation in the mobile-radio environment', *IEEE Transactions*, February 1972, **VT-21** (1), pp. 27–38.

GARDINER, P.: 'Ghost cancellation studies in Europe', *Proceedings of ITU/BR Workshop on Enhanced Television*, Auckland, New Zealand, 1993, pp.110–115.

GARDINER, P., BROCKHURST, J.S., and CURTIS, J.S.: 'Ghost cancellation for 625-line systems', *Proceedings of International Broadcasting Convention*, GT V/EPS 242, 1994, pp. 673–678.

GERCHBERG, R.W. and SAXTON, W.O.: 'A practical algorithm for the determination of phase from image and diffraction plane pictures', *Optik*, April 1972, **35** (2), pp. 237–246.

HASHEMI, H.: 'Simulation of the urban radio propagation channel', *IEEE Transactions*, August 1979, **VT-28** (3), pp. 213–225.

HÖGBOM, J.A.: 'Aperture synthesis with a non-rectangular distribution of interferometer baselines', *Astronomy and Astrophysics*, Supplement Series, 1974, **15** (417–427).

HOTELLING, H.: 'New light on the correlation coefficient and its transforms', *Journal of the Royal Statistical Society*, 1953, Part B, **15**, pp. 193–225.

HWANG, J.K. and WINTERS, J.H.: 'Sinusoidal modelling and prediction of fast fading processes', *Proceedings GLOBECOM* '98, Sydney, November 1998, pp. 892–897.

JAKES, W.C. (Ed.): *Microwave Mobile Communications*, Wiley, 1974. Also reprinted IEEE 1989.

JERUCHIM, M.C., BALABAN, P., and SHANMUGAN, K.S.: *Simulation of Communication Systems*, Plenum, New York, 1992.

KOO, D.: 'Ghost cancellation for television in the United States: properties and applications of the new ghost cancellation reference signal', (Philips Laboratories, Briarcliff Manor, New York), *CCIR Task Group 11*, 1993; also in *Proceedings of ITU/BR Workshop on Enhanced Television*, Auckland, New Zealand, 1993, pp. 105–109.

KOO, D.: 'Ghost cancellation with ITU system-C standard ghost cancellation reference signal', *SMPTE Journal*, June 1995, pp. 370–376.

KUMARESAN, R. and TUFTS, D.W.: 'Estimating the angles of arrival of multiple plane waves', *IEEE Transactions*, January 1983, **AES-19**, pp. 134–139.

LEE, W.C.Y.: *Mobile Communications Engineering*, McGraw-Hill, 1982.

LEONARD-TAYLOR, M.: 'Delay spread simulator', Industrial Research Limited, Report 758, July 1997.

LEONARD-TAYLOR, M.: 'Delay spread for non-minimum, non-maximum phase two-path model', The New Zealand Institute of Industrial Research Limited Report 820, April 1998.

LOEVE, M.: *Probability Theory*, Van Nostrand, Princetown, New Jersey, 1955.

MIDDLETON, D.: *An Introduction to Statistical Communications Theory*, McGraw-Hill, 1960. Also reprinted IEEE Press, 1997.

NØRKLIT, O. and VAUGHAN, R.G.: 'Angular partitioning to yield equal Doppler contributions', *IEEE Transactions*, September 1999, **VT-48** (5), pp. 1347–1442.

OGILVY, J.A.: *Theory of Wave Scattering from Rough Surfaces*, IOP Publishing, 1991.

OPPENHEIM, A.V. and SCHAFER: *Digital Signal Processing*, Prentice-Hall, 1975.

PAPOULIS, A.: *Signal Analysis*, McGraw-Hill, 1977.

PAPOULIS, A.: *The Fourier Integral and its Applications*, McGraw-Hill,1987

PARSONS, J.D.: *The Mobile Radio Propagation Channel*, Pentech Press, London, 1992.

PARSONS, J.D. and GARDINER, J.G.: *Mobile Communications Systems*, Blackie, London, 1989.

PÄTZOLD, M., KILLAT, U., and LAUE, F.: 'An extended Suzuki model for land mobile satellite channels and its statistical properties', *IEEE Transactions*, May 1998, **VT-47** (2), pp. 617–630.

PIERCE, J.R. and STEIN, S.: 'Multiple diversity with non-independent fading', *Proceedings of the IRE*, January 1960, **48**, pp. 89–104.

POLETTI, M.A.: 'Linearly swept frequency measurements, time-delay spectrometry, and the Wigner distribution', *Journal of the Audio Engineering Society*, 1988, **36** (6), pp. 457–468.

POLETTI, M.A.: 'The homomorphic analytic signal', *IEEE Transactions*, August 1997, **SP-45** (8), pp. 1943–1953.

POLETTI, M.A. and VAUGHAN, R.G.: 'Reduction of multipath fading effects in single variable modulations', *Proceedings of the International Symposium on Signal Processing and Applications* (ISSPA 90), Gold Coast, Australia, August 1990, pp. 672–676.

RAICH, R. and VAUGHAN, R.G.: 'Source power distribution for multipath environment using near field circular array beamformer', *Electronics Letters*, October 1999, **35** (22), pp. 1893–1894.

RICE, S.O.: 'Mathematical analysis of random noise', *Bell System Technical Journal*, 1944, **23**, pp. 282–332.

RICE, S.O.: 'Mathematical analysis of random noise', *Bell System Technical Journal*, 1945, **24**, pp. 46–156.

RICE, S.O.: 'Statistical properties of a sine wave plus random noise', *Bell System Technical Journal*, 1948, **27**, pp. 109–157.

ROSENBLATT, M.: 'The multidimensional prediction problem', *in* HOFFMAN, W. C. (Ed.): 'Statistical methods in radiowave propagation' (Pergamon Press, 1960).

ROY, R. and KAILAITH, T.: 'ESPRIT-estimation of signal parameters via rotational invariance techniques', *IEEE Transactions*, July 1989, **ASSP-37** (7), pp. 984–995.

SCHMIDT, R.O.: 'Multiple emitter location and signal parameter estimations', *IEEE Transactions*, March 1986, **AP-34**, pp. 276–280.

SCOTT, N.L. and VAUGHAN, R.G.: 'The effect of signal demeaning on its autocorrelation', *Proceedings of the International Symposium on Signal Processing and its Applications* (ISSPA '92), Gold Coast, Australia, August 1992, pp.130–133.

STEELE, R. (Ed.): *Mobile Radio Communications*, Pentech Press, London,1992.

STEIN, S.: Part III of Schwartz, M., Bennett, W.R., and Stein, S.: *Communications Systems and Techniques*, McGraw-Hill, 1966.

SUZUKI, H.: 'A statistical model for urban radio propagation', *IEEE Transactions*, July 1977, **Com-25** (7), pp. 673–680.

TEAL, P.: *Real Time Characterisation of the Mobile Multipath Channel*, PhD Thesis, Australian National University, 2002.

TEAL, P. and VAUGHAN, R.G.: 'Simulation and performance bounds for real time prediction of the multipath mobile channel', *Proceedings of the 2001 IEEE Workshop on Statistical Signal Processing*, Singapore, August 2001, pp. 548–551.

THERRIEN, C.W.: *Discrete Random Signals and Statistical Signal Processing*, Englewood Cliffs, New Jersey, Prentice Hall, 1992.

TONG, L., XU, G., and KAILAITH, T.: 'Blind identification and equalisation based on second order statistics: a time domain approach', *IEEE Transactions*, March 1994, **IT-40** (2).

TREICHLER, J.R. and AGEE, B.A.: 'A new approach to multipath correction of constant modulus signals', *IEEE Transactions*, April 1983, **ASSP-31** (2).

TUFTS, D.W. and KUMARESAN, R.: 'Estimation of frequencies of multiple sinusoids: making linear prediction behave like maximum likelihood', *Proceedings of the IEEE*, September 1982a, **70**, pp. 975–989.

TUFTS, D.W. and KUMARESAN, R.: 'Singular value decomposition and improved frequency estimation using linear prediction', *IEEE Transactions*, August 1982b, **ASSP-30**, pp. 671–675.

TURIN, G.L.: 'Introduction to spread-spectrum antimultipath techniques and their application to the urban digital radio', *Proceedings of the IEEE*, March 1980, **68**, pp. 328–353.

TURIN, G.L., CLAPP, F.D., JOHNSTON, T.L., FINE, S.B. and LAVRY, D.: 'A statistical model of urban multipath propagation', *IEEE Transactions*, February 1972, **VT-21** (1), pp. 153–161.

UHLENBECK, G.E.: 'Theory of the random process', *MIT Radiation Laboratory Report*, October 15, 1943 , No. 454.

US-ATSC: United States advanced television steering committee standard: ghost cancelling reference signal, Doc. A/49, August 1992/ May 1993.

VAUGHAN, R.G.: 'Signal limiting in envelope correlations for diversity channel evaluation: a statistical approach', *Proceedings of the International Symposium on Personal, Indoor, and Mobile Communications*, (PIMRC 96), Taipei, October 1996, pp. 781–785.

VAUGHAN, R.G.: 'Finite sample estimates for mobile channels', *Proceedings of the IEEE Vehicular Technology Conference* (VTC 2000), Boston, September 2000, **2**, pp. 797–804.

VAUGHAN, R.G. and ANDERSEN, J. BACH.: 'Antenna diversity in mobile communications', *IEEE Transactions*, November 1987, **36** (4), pp. 149–172.

VAUGHAN, R.G. and ANDERSEN, J. BACH.: 'Prediction of signals in mobile communications and systems that use the prediction for signal reception and decoding', New Zealand patent application 280348 (November 1995), 286131 (March 1996), 329173, (November 1997); WO99/26423.

VAUGHAN, R.G. and SCOTT, N.L.: 'Correlation of delay spread and CCIR impairment', *Electronic Letters*, 6th July, 1995, **31** (14), pp. 1123–1124. (Also Erratum, *EL*, 3 August, 1995, **31** (16), p. 1396.)

VAUGHAN, R.G. and SCOTT, N.L.: 'Super-resolution of pulsed multipath channels for delay spread characterization', *IEEE Transactions*, March 1999, **Comms 47** (3), pp. 343–347.

VAUGHAN, R.G., SCOTT, N.L., and SOUTHON, J.P.T.T.K.: 'Characterization of standing wave movement by in-line antennas', *Proceedings of the Antennas and Propagation Society Symposium IEEE 91CH3036-1*, London, Ontario, June 1991, **2**, pp. 974–977.

VAUGHAN, R.G., TEAL, P.D., and RAICH, R.: 'Short-term mobile prediction using discrete scatterer propagation model and subspace signal processing algorithms', *Proceedings of the IEEE Vehicular Technology Conference* (VTC2000/Fall), Boston, September 2000, **2**, pp. 751–758.

VOELCKER, H.B.: 'Towards a unified theory of modulation'. Part I: Phase-Envelope Relationships, *Proc IEEE*, March 1966, **54** (3), pp. 340–353; Part II: Zero Manipulation, *Proc IEEE*, May 1966, **54** (5), pp. 735–755.

WAX, M. and KAILAITH, T.: 'Detection of signals by information theoretic criteria', *IEEE Transactions*, April 1985, **ASSP-33**, pp. 387–392.

WITHERS, C.S.: 'A non-parametric confidence interval for the correlation coefficient', *Internal DSIR Report*, New Zealand, 1977.

WU, J. and AGHVAMI, A.H.: 'A new adaptive equaliser with channel estimator for mobile radio communications', *IEEE Transactions*, August 1996, **45** (3), pp. 467–474.

YONG, Y.L.: *The Multivariate Normal Distribution*, Springer Series in Statistics, Springer Verlag, 1990.

Chapter 8

Antenna principles

8.1 Basic antenna parameters and elements

This section introduces the fundamental antenna parameters for the context of mobile communications. Some of the terms have been used already in previous chapters, but their formal introduction is given context in this chapter. The basic antenna elements – the dipole and loop – are discussed both to fix ideas and to demonstrate design considerations for compact antennas.

Antenna elements are often categorised from the basis of the antenna's construction: hence the terms wire antennas, slot antennas, patch antennas, aperture (e.g. horn) antennas, etc. Other categorisations reflect a physics approach: electric source antennas, magnetic source antennas, etc., or some engineering parameter type; for example: low-, medium- or high-gain antennas. Here, we are mostly interested in compact and relatively simple elements, although the mounting, often on a complicated platform such as a handheld terminal, can inadvertently make the antenna appear complicated.

8.1.1 Directivity and gain from patterns

The directivity of an antenna is classically defined in a single direction corresponding to the direction of maximum power radiation:

$$D(\theta, \phi) = \frac{\text{Total power radiated per solid angle in a direction } (\theta, \phi)}{\text{Average power radiated per unit solid angle}}. \qquad (8.1.1)$$

The directivity is thus a function only of the 3D shape of the radiation pattern. It is a measure of how directional the antenna beam is, referenced to an isotropic (perfectly spherical) pattern. The principle of reciprocity for antennas states that the receiver and transmitter can be interchanged between the two antennas of a link, and this implies that the transmit pattern and the receive patterns are the same.

In classical antenna treatments, it is conventional to denote both the incident and radiated (i.e. antenna pattern) electric fields with a vector $\boldsymbol{E}$. In mobile communications, $\boldsymbol{E}$ is often used to denote the incident propagating fields and the far-field antenna radiation complex pattern, for the electric fields, is denoted using the complex height (cf. Chapter 2),

$$\boldsymbol{h}(\theta, \phi) = \hat{\theta} h_\theta(\theta, \phi) + \hat{\phi}\, h_\phi(\theta, \phi) \tag{8.1.2}$$

and the units are recalled as m. The complex height has the special proportionality relationship $V_\mathrm{o} = \boldsymbol{E}_\mathrm{inc} \cdot \boldsymbol{h}$ in which a plane wave given by its incident field $\boldsymbol{E}_\mathrm{inc}$ results in the open circuit antenna voltage V_o.

A pattern function, proportional to the radiation pattern, $|h(\theta, \phi)|$, is used for finding pattern-related parameters. Here we denote such a radiation pattern function as $g(\theta, \phi) = |\boldsymbol{g}(\theta, \phi)|$ where $\boldsymbol{g}(\theta, \phi)$ is directly proportional to $\boldsymbol{h}(\theta, \phi)$.

8.1.1.1 Far field and directivity relationship

A normalised pattern has

$$\frac{1}{4\pi} \int_{4\pi} |\boldsymbol{g}(\Omega)|^2 \, \mathrm{d}\Omega = 1, \tag{8.1.3}$$

where the elemental solid angle is $\mathrm{d}\Omega = \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi$. With this normalisation, the electric field strength radiated at a large distance r (i.e. the far field, see below) is

$$|\boldsymbol{E}(\theta, \phi, r)| = \sqrt{\frac{P_\mathrm{rad} Z_0}{2\pi}} \left(\frac{1}{r}\right) |\boldsymbol{g}(\theta, \phi)| \tag{8.1.4}$$

where P_rad is the total radiated power, k_0 is the free-space wavenumber and $Z_0 \approx 377\,\Omega$ is the free-space impedance. The inverse distance term (in brackets) gives the spherical wave dependence, and the phase term for the outgoing wave, $\mathrm{e}^{-\mathrm{j}kr}$, is omitted in these magnitude expressions. The proportionality constant (the square root term) is from the radiated field power density, $\frac{1}{2}|\boldsymbol{E}|^2/Z_0$, being integrated over the sphere at radius r. The field quantities are discussed further below. For this section, only the radiation pattern, $|\boldsymbol{g}(\theta, \phi)|$, is of interest. The spherical elemental area on Cartesian coordinates is described in Figure 8.1.1.

The radiated power density per unit solid angle (in W m^{-2} steradian^{-1}) is

$$\Phi(\theta, \phi) = \frac{1}{4\pi} P_\mathrm{rad} |\boldsymbol{g}(\theta, \phi)|^2. \tag{8.1.5}$$

The directional gain is defined only from the pattern, as the ratio of powers

$$\begin{aligned} D(\theta_0, \phi_0) &= \frac{|\boldsymbol{g}(\theta_0, \phi_0)|^2}{(1/4\pi) \int_0^{2\pi} \int_0^{\pi} |\boldsymbol{g}(\theta, \phi)|^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi} \\ &= \frac{|g_\theta(\theta_0, \phi_0)|^2 + |g_\phi(\theta_0, \phi_0)|^2}{\frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} \left(|g_\theta(\theta, \phi)|^2 + |g_\phi(\theta, \phi)|^2\right) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi} \end{aligned} \tag{8.1.6}$$

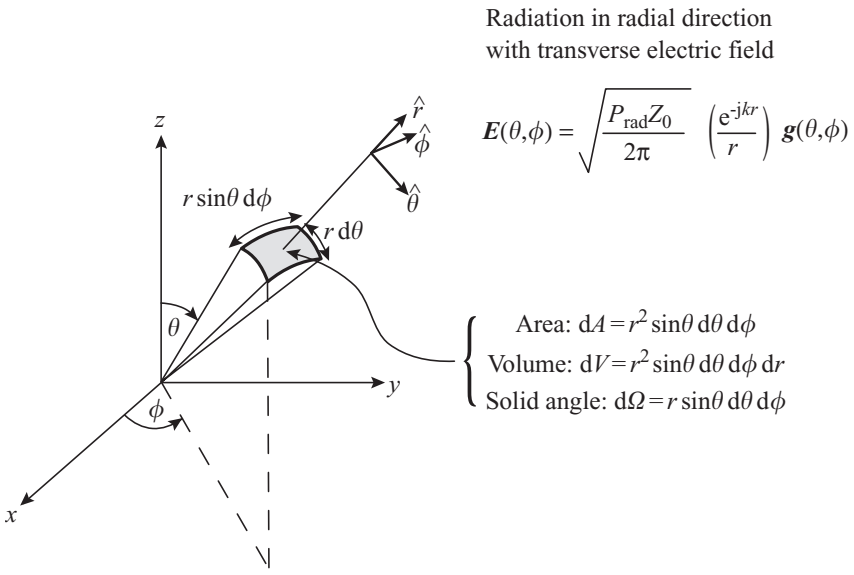


Figure 8.1.1 The elemental area and solid angle in spherical coordinates with a Cartesian datum coordinate system. The radiated electric far field is given in terms of the normalised pattern function $\boldsymbol{g}(\theta, \phi)$. In the far field, the electric and magnetic fields have $\hat{\theta}$- and $\hat{\phi}$-directions, and the power flow is in the $\hat{r}$-direction

and if (θ_0, ϕ_0) is a direction of the maximum radiated power density, then D is the maximum directivity, which is often called simply the directivity. Since the directivity is a ratio of pattern functions, the complex height could also be used in the directivity formula. Note that both polarisations are required for defining the directivity.

The directivity of an array is given in Section 8.4, where the diversity gain and array gain are treated.

8.1.1.2 Distributed directivity

In mobile communications, the directional gain over a set of directions, or the *distributed directivity*, is often a more useful measure, because it is common to illuminate (or receive from) the surrounding scatterers which are in turn illuminated by the transmitting antenna or other scatterers. Even for a line-of-sight situation, a single beam illuminating many users needs to be shaped and the distributed directivity is a measure of such coverage. This is the same situation as in satellite communications where a beam must cover a distribution of terrestrial regions. The distributed directional gain is defined here as the proportion of total power illuminating the wanted area

$$D_{\text{d}}^{(S)} = \frac{\int_0^{2\pi}\int_0^{\pi} S(\theta,\phi)|\boldsymbol{g}(\theta,\phi)|^2 \sin\theta\,d\theta\,d\phi}{(1/4\pi)\int_0^{2\pi}\int_0^{\pi} |\boldsymbol{g}(\theta,\phi)|^2 \sin\theta\,d\theta\,d\phi} \tag{8.1.7}$$

and $S(\theta, \phi)$ is a function describing the wanted area, normalised so that the distributed directivity reduces to the directional directivity for a single direction. For

a single direction, the delta function for the above normalisation is $S(\theta_0, \phi_0) = 1/(\sin\theta_0)\delta(\theta - \theta_0)\delta(\phi - \phi_0)$.

For a given distribution of directions, say uniform over $(\Delta\theta, \Delta\phi)$ and centred at (θ_0, ϕ_0), a measure for the distributed directivity is given by the amount of power radiated to the target area over the total radiated power:

$$D_{\mathrm{d}}(\theta_0, \phi_0; \Delta\theta, \Delta\phi) = K \frac{\int_{\phi_0-\Delta\phi/2}^{\phi_0+\Delta\phi/2} \int_{\theta_0-\Delta\theta/2}^{\theta_0+\Delta\theta/2} |\boldsymbol{g}(\theta, \phi)|^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi}{\int_0^{2\pi} \int_0^{\pi} |\boldsymbol{g}(\theta, \phi)|^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi}, \quad (8.1.8)$$

where the factor $K = 4\pi/(2\Delta\phi \sin\theta_0 \sin\Delta\theta)$ is to allow the same sized target area to be anywhere on the sphere. In practice, this location on the sphere is fixed, and when comparing patterns, the factor is common and can be omitted. The distributed quantity is a figure-of-merit for guiding the design of mobile antennas, analogous to the directivity for point-to-point applications. The better the average match of the pattern to the target area, the better the average gain of the path(s). An example using the distributed directivity to estimate an excess average path loss from beam misalignment is given below for a base station antenna. It is also used in Section 8.4 for guiding the evaluation of various patch antennas for roof-mounted vehicular elements.

8.1.1.3 Directive gain

The antenna gain is the same as the directivity except that the gain measure includes the antenna losses. These losses are couched with the *antenna efficiency*, η_{ant}, which is unity for zero losses, so the relationship between gain and directivity is written

$$G(\theta, \phi) = \eta_{\mathrm{ant}} D(\theta, \phi) \quad (8.1.9)$$

and similarly for the distributed gain. The losses can be in the imperfect conductors of the antenna and also in any imperfect dielectric components in the antenna. A clear definition of where the antenna ports are is important here. The normal definition of the antenna efficiency (see below) specifically excludes losses in any matching circuit, for example.

Note that the gain and directivity are defined using the power in both polarisations. If *co-polar gain* is required, then the definitions need a change. For example, the vertical polarisation (defined here as the pattern component whose electric field is always in the θ direction) directivity is

$$G_\theta(\theta_0, \phi_0) = \eta_{\mathrm{ant}} \frac{|g_\theta(\theta_0, \phi_0)|^2}{(1/4\pi) \int_0^{2\pi} \int_0^{\pi} \left(|g_\theta(\theta, \phi)|^2 + |g_\phi(\theta, \phi)|^2\right) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi}. \quad (8.1.10)$$

So in a gain or directivity measurement or calculation, the patterns of both polarisations are still required to find the co-polar gain. Only for purely polarised antennas can the cross-polar component be omitted from a measurement.

8.1.1.4 Pattern models

Mathematical models of patterns are useful for getting design and system performance guidelines, in particular for directive beams. Generally we take only one polarisation, and use a real, positive function, for the pattern, $g(\theta, \phi)$. A beam-type pattern (neglecting any sidelobes) can be modelled with a Gaussian function or a circular function. Models using the circular functions are well suited to the circular support of real space in spherical coordinates ($\theta \leq \theta \leq \pi,\ 0 \leq \phi \leq 2\pi$). The Gaussian form is very convenient for integrations, but is complicated by the fact that it requires truncation within the circular support of spherical coordinates.

8.1.1.5 Gaussian beam

The 1D (in spherical coordinates) Gaussian beam pattern, pointing to the direction $\phi = \phi_0$ is

$$g(\phi) = \exp\left\{-\frac{(\phi - \phi_0)^2}{2\sigma_g^2}\right\}, \quad -\pi \leq (\phi - \phi_0) \leq \pi \tag{8.1.11}$$

so the function is purely real. This is taken as the amplitude pattern, and is not normalised. The Gaussian function must be truncated at the limits and from now the pattern direction is kept along the x-axis, i.e. to $\phi_0 = 0$. If the above pattern is taken as an amplitude pattern, the half power beamwidth (HPBW, which is here the one-dimensional angle between the directions where the power pattern is half of its maximum value) is related to the Gaussian spread parameter (standard deviation), σ_g, as

$$HPBW = \sigma_g^2\sqrt{2\ln\sqrt{2}}, \quad \sigma_g \approx 0.6HPBW. \tag{8.1.12}$$

The power pattern, normalised here to be unity in the maximum direction, follows as

$$\left(\frac{g(\phi)}{g(0)}\right)^2 = \exp\left\{-2(2\ln\sqrt{2})^2\left(\frac{\phi}{HPBW}\right)^2\right\} \tag{8.1.13}$$

which gives, for example, a level of about -0.8 dB at an angle of a quarter the *HPBW* from the direction of maximum. The 2D pattern is given in Figure 8.1.2. To relate the beamwidth to the directivity, a 2D beam is required.

A circularly symmetric Gaussian beam results when the Gaussian beam is denoted using the zenith coordinate, θ. The directivity of the two-dimensional $g(\theta)$ is then

$$D_g(\sigma_g) = \frac{2}{\int_0^\pi \exp\left\{-\theta^2/\sigma_g^2\right\}\sin\theta\,\mathrm{d}\theta}. \tag{8.1.14}$$

The pencil beam formula (Kraus, 1988) for directivity,

$$D_p(\text{degrees}) = \frac{41\,000}{(HPBW \text{ in degrees})^2} \tag{8.1.15}$$

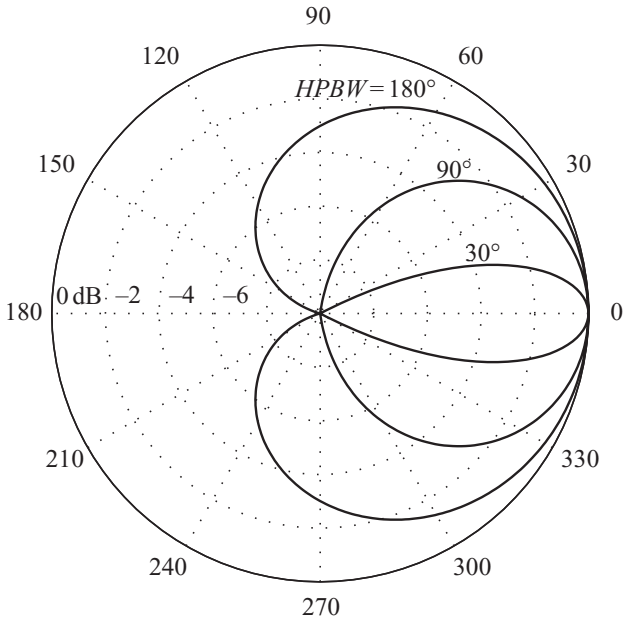


Figure 8.1.2 Polar plots of Gaussian beams for half power beamwidths of 30°, 90° and 180°

is a good estimate, being within 0.5 dB as long as σ_g is less than about 1 rad at which value the directivity formulas give $D_g = 6.75$ dB and $D_p = 6.61$ dB. For asymmetric beams, the product of the *HPBW*s in each plane can be used instead of the *HPBW* squared term. Kraus' approximation for Gaussian beams ($D_g \approx 0.88 D_p$) holds accurately for directivities larger than about 8 dB, as seen in Figure 8.1.3. So the formula is good for higher-gain base-station antennas, at least those with symmetric vertical patterns but not for compact base stations and mobile terminals, which are low gain. Low-gain 'beams' are more conveniently modelled with circular functions.

8.1.1.6 Pattern models from circular functions

The simplest beam model using circular functions is, for the 1D pattern,

$$g(\phi) = \cos^n\left(\frac{\phi}{2}\right) \tag{8.1.16}$$

which is zero at $|\phi| = \pi$ and is unity at $\phi = 0$. It is straightforward to have any direction for the maximum, with circular functions. If $g^2(\phi)$ again represents the power pattern, the HPBW and the beamwidth parameter n are related by

$$HPBW = 4\cos^{-1}\left(2^{-1/2n}\right), \quad n = -\frac{\log\sqrt{2}}{\log\cos\left(\frac{HPBW}{4}\right)}. \tag{8.1.17}$$

These beams are used in Section 6.4 for investigating spatial correlations for antennas with directive patterns. For highly directive beams, n can become excessive (hundreds) and more reliable numerical computations may result if the model is changed

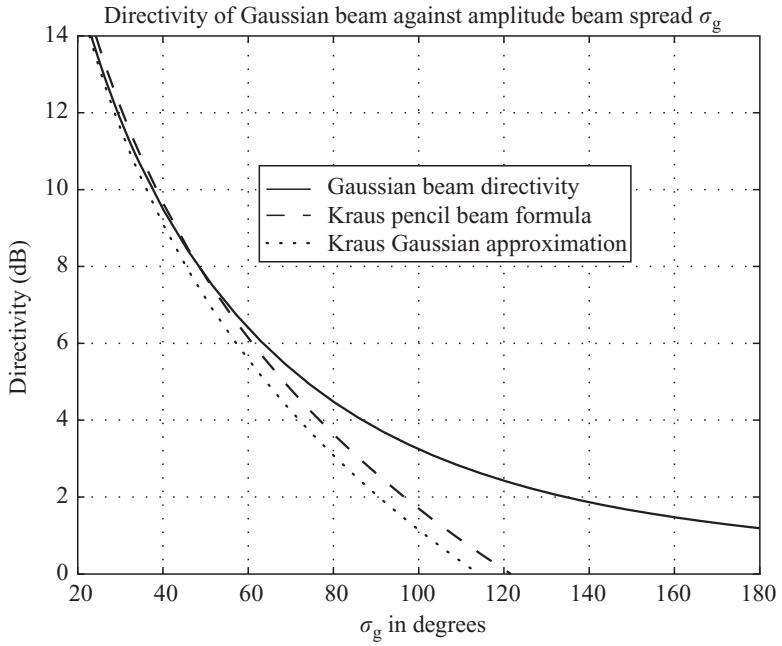


Figure 8.1.3 The directivity of a Gaussian beam against the standard deviation parameter of the beam amplitude function. The pencil beam formula (Kraus, 1988) is accurate for directivities over about 8 dB

by scaling the angle, namely $g(\phi) = \cos^n(m\phi)$, and reducing the integration limits appropriately to maintain a single beam, and modifying the HPBW formula.

The simplest expression for a 2D, or solid beam, is one directed along the x-axis and written

$$g(\theta, \phi) = \sin^{n_\theta}\theta \cos^{n_\phi}\left(\frac{\phi}{2}\right) \tag{8.1.18}$$

and n_θ, n_ϕ are separate parameters (with separate HPBW formulas) for the 2D pattern widths.

8.1.1.7 Effect of beam misalignment using the distributed directivity

Recall that the distributed directivity is a measure of how well an area is illuminated by an antenna pattern. An example of using this is for evaluating the effect of down-tilt of a fixed base-station antenna mounted above the target area containing the mobile terminals. In a cellular system, the down-tilted beam illuminates the wanted area and so radiates less towards the horizon where distant cells reuse the frequency. The down-tilt is also known to decrease the dispersion, or delay spread, causing less intersymbol interference degradation to the BER (Kozono *et al.*, 1991). The amount of down-tilt may need to be optimised for each base station antenna height and environment.

A typical situation is to design the antenna such that its fixed beamwidth covers a typical field-of-coverage target area. The beam is then directed downwards so that it gives the best system performance. The beam tilt is depicted in Figure 8.1.4. Here, for simplicity, the target area of wanted directions is taken to be fixed in the horizontal

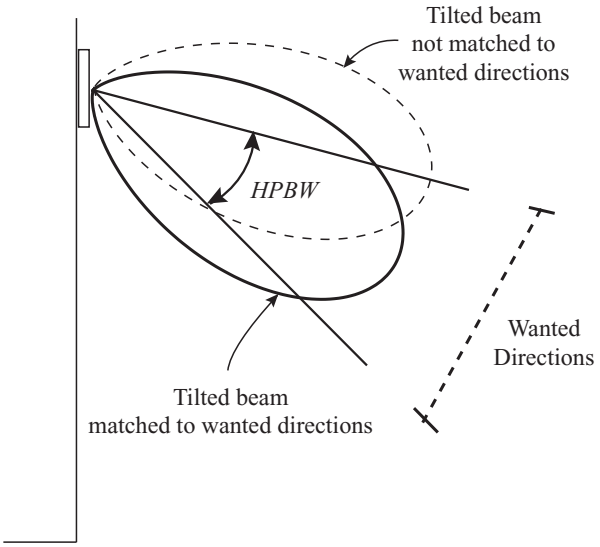


Figure 8.1.4 The down-tilt of an antenna beam towards the wanted directions. In this example, the half power beamwidth is configured to cover the extent of the wanted directions and the tilt angle is chosen for maximum distributed directional gain

plane so that the horizontal beamwidth and the horizontal beam direction are fixed and do not play a role in the vertical tilting calculations.

The effect of varying the antenna direction is estimated by modelling with a symmetric real beam in elevation angle only. The half power beamwidth (HPBW) gives the angles to the target area so that when the beam is centred on the wanted area, the distributed directivity is a maximum. The distributed directivity is (omitting the scale factor, K)

$$D_{\mathrm{d,max}} = \frac{\int_{-HPBW/2}^{HPBW/2} g(\theta)\,\mathrm{d}\theta}{\int_{-\pi}^{\pi} g(\theta)\,\mathrm{d}\theta} \tag{8.1.19}$$

where only the forward hemisphere is considered. When the beam is offset from the centre of the wanted directions, or misaligned, by an angle ϕ, then this distributed directional gain is

$$D_{\mathrm{d}}(\phi) = \frac{\int_{-HPBW/2}^{HPBW/2} g(\theta+\phi)\,\mathrm{d}\theta}{\int_{-\pi}^{\pi} g(\theta)\,\mathrm{d}\theta} \tag{8.1.20}$$

and so a normalised directional gain to a target area of fixed bearing is, here with $HPBW/2 = \alpha$,

$$\begin{aligned} D_{\mathrm{d,norm}}(\phi) &= \frac{\int_{-\alpha}^{\alpha} g(\theta+\phi)\,\mathrm{d}\theta}{\int_{-\alpha}^{\alpha} g(\theta)\,\mathrm{d}\theta} \\ &\approx \frac{1}{2\,\mathrm{erf}(\sqrt{\ln 2})}\left(\mathrm{erf}\left(\sqrt{\ln 2}\left(1-\frac{\phi}{\alpha}\right)\right) - \mathrm{erf}\left(\sqrt{\ln 2}\left(-1-\frac{\phi}{\alpha}\right)\right)\right) \end{aligned} \tag{8.1.21}$$

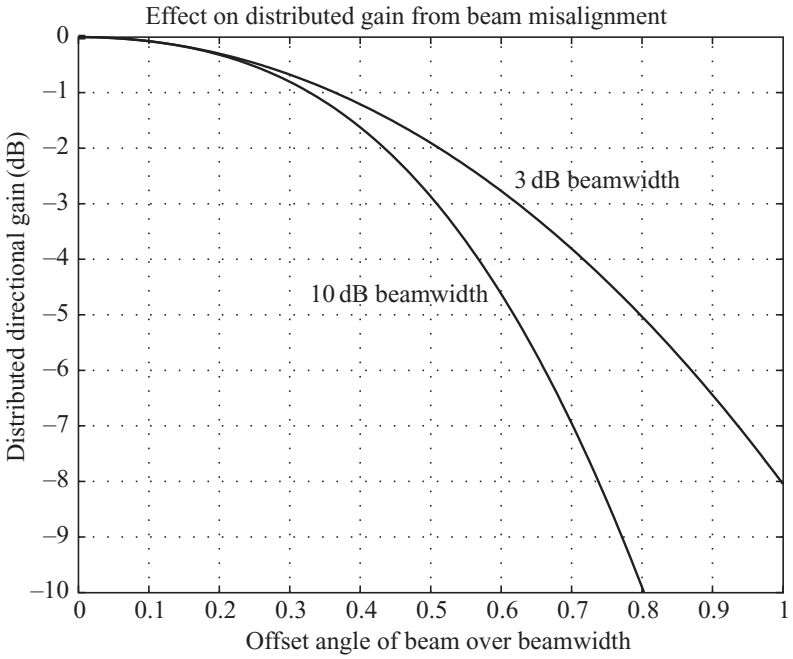


Figure 8.1.5 The change in the distributed directional gain with the misalignment of the tilt angle, where the wanted area of illumination in the cell is given by the extent of the beam to the 3 db and 10 dB beamwidths, respectively

where a Gaussian beam model has been used from above (see also Appendix C). This is plotted in Figure 8.1.5, in which the 10 dB beamwidth (instead of the HPBW) has also been used. It makes no difference in this model whether the misalignment is upwards or downwards, because of the symmetry of the model. If the beam misalignment increases to about half of the HPBW, for example, then the distributed directional gain decreases by about 1.5 dB.

8.1.2 Antenna efficiency factors

The antenna losses are separated from other losses in the Friis transmission equation,

$$P_{\mathrm{R}} = P_{\mathrm{T}} \frac{G_{\mathrm{T}} G_{\mathrm{R}}}{L} \eta, \tag{8.1.22}$$

where L is the path loss (see Appendix A) and η is a total efficiency factor that can be broken down to the product of many other efficiency terms. Of these, the polarisation efficiency, η_{pol}, and the impedance match efficiency, η_{imp}, are discussed below. Also, the antenna efficiency used above, η_{ant}, which accounts for the ohmic losses in each antenna, is given. Note that in a scattering environment with distributions of directions being illuminated, the gain takes on a different interpretation from the classical point-to-point form, called the mean effective gain; see Section 8.2.

8.1.2.1 Polarisation efficiency

The classical polarisation mismatch factor between two line-of-sight antennas is

$$\eta_{\text{pol}} = |\hat{\rho}_{\text{T}} \cdot \hat{\rho}_{\text{R}}|^2 \tag{8.1.23}$$

where $\hat{\rho}_{\text{T}}$ is the unit vector describing the electric field orientation of the single plane wave transmitted and incident on the receiving antenna; and similarly, $\hat{\rho}_{\text{R}*}$ is for the receiving antenna receiving pattern in the direction of the incident wave. The transmit pattern of the receiving antenna has polarisation $\hat{\rho}_{\text{R}*}$, and this conjugation does not affect linear polarisations, but changes the sense of circular polarisation.

The Poincare sphere (see, for example, Collin and Zucker, 1969), originally introduced for antennas by G.A. Deschamps in 1951, is a visualisation of polarisation and the match between polarisations. Each point on the surface of the sphere represents a polarisation. The north pole is left-hand circular, the south pole is right-hand circular, and on the equator are the various orientations of purely linear polarisation. All other points on the sphere are the elliptic polarisations. Polarisations on opposite points of the sphere are orthogonal, for example horizontal and vertical polarisations are on opposing points of the equator. The mismatch factor is given by $\cos^2 \zeta$ where 2ζ is the angular distance between the points on the sphere representing the transmit and receive polarisations. For a good polarisation match, the transmit and receive polarisations should be close on the sphere. Polarisation relations are discussed more in Section 8.2.

Considering a single, mobile antenna, the multiple incident waves are randomly polarised (Chapter 2) and so there are a multitude of polarisation mismatches which are implicitly accounted for in the basic receiving antenna equation (2.1.2). For completely random polarisation, there is an average 3 dB loss owing explicitly to the polarisation mismatch. In line-of-sight situations, such as some terrestrial mobile services and most satellite mobile systems, the polarisation becomes more important because there is an obvious advantage in trying to realise a mobile antenna to be matched to the line-of-sight polarisation. Circular polarisation is used for satellite transmission in order to avert the depolarising effect on linear polarisations caused by the Faraday rotation in the ionosphere. For a linearly polarised mobile antenna, the orientation of an antenna is important for the polarisation match in any line-of-sight path. With circularly polarised antennas, this orientation is no longer important. Circularly polarised antennas are discussed in Section 8.3.

8.1.2.2 Load match parameters

Another efficiency factor which often appears in the Friis equation is the antenna impedance mismatch, here written for both receiving and transmitting antennas,

$$\eta_{\text{imp}} = \left(1 - |\Gamma_{\text{T}}|^2\right)\left(1 - |\Gamma_{\text{R}}|^2\right) \tag{8.1.24}$$

where Γ is the voltage reflection coefficient at the antenna (or antenna plus matching circuit). The voltage reflection coefficient on the transmission line to the

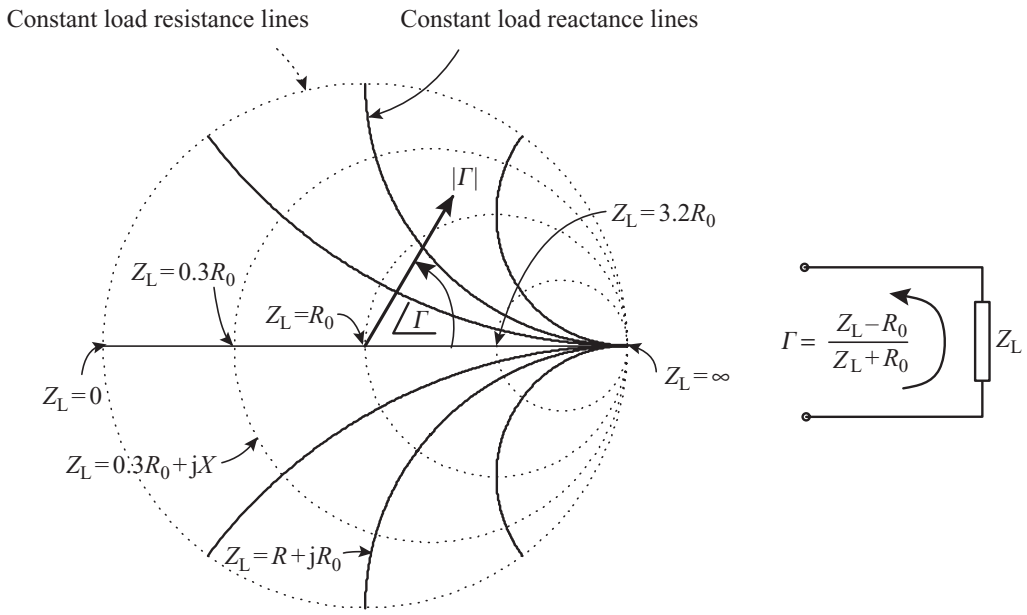


Figure 8.1.6 The Smith chart is the polar plot of the voltage reflection coefficient with lines of constant load resistance and constant load reactance. The left end is the short circuit and the right end is the open circuit

receiver/transmitter is given by the impedance match of the characteristic impedance of the transmission line, Z_0, and the impedance of the antenna (load), Z_L,

$$\Gamma = \frac{Z_L - Z_0}{Z_L + Z_0}. \tag{8.1.25}$$

The Smith chart relates these quantities graphically, comprising a polar plot of Γ, shown in Figure 8.1.6. In terms of the impedance, the centre of the polar plot corresponds to $\Gamma = 0$, i.e. the load is $Z_L = Z_0$, which is typically 50 Ω.

The reflection of a wave from a mismatched load mixes with the incident wave to form a standing wave, in the same way as with the two-path propagation model discussed in Chapter 5. The maximum voltage is $V_{max} = 1 + |\Gamma|$ and similarly the minimum voltage is $V_{min} = 1 - |\Gamma|$. The voltage standing wave ratio (VSWR) is the ratio of the maximum standing wave voltage and the minimum standing wave voltage, so Γ is related to the VSWR by

$$VSWR = \frac{1 + |\Gamma|}{1 - |\Gamma|}, \quad |\Gamma| = \frac{VSWR - 1}{VSWR + 1}. \tag{8.1.26}$$

If the antenna is at resonance, i.e. the transmission line load is purely resistive, $Z_L = R$, then the VSWR values are given by

$$\begin{aligned} VSWR &= \frac{R}{Z_0}, \quad R > Z_0 \quad \text{(at voltage maximum)}, \\ VSWR &= \frac{Z_0}{R}, \quad R < Z_0 \quad \text{(at voltage minimum)}. \end{aligned} \tag{8.1.27}$$

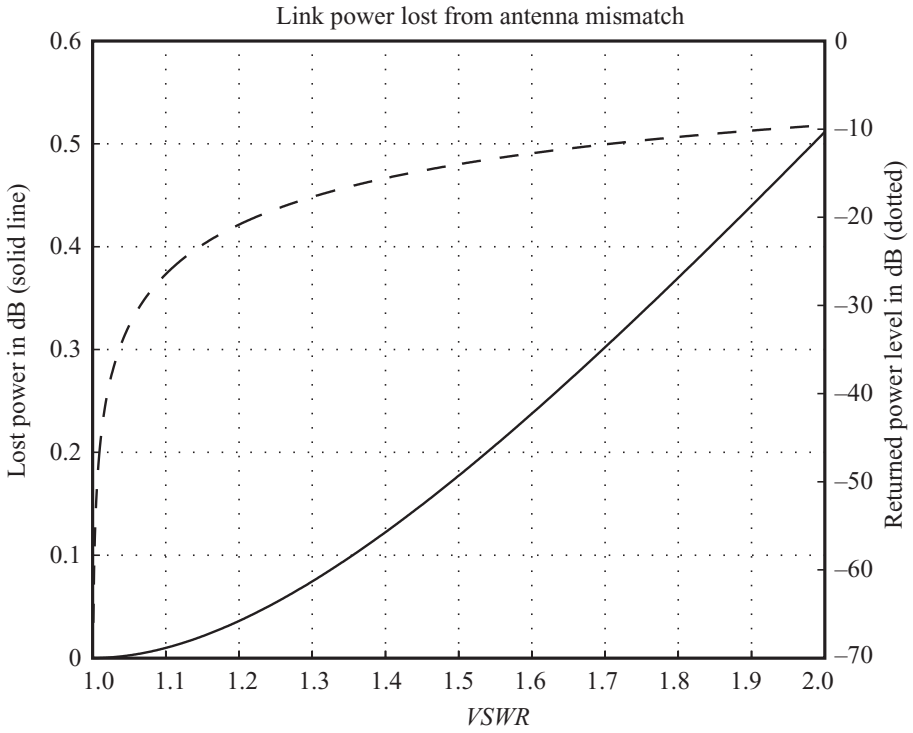


Figure 8.1.7 The reflected power, or lost power (left axis) caused by a mismatch with a given VSWR, and the return loss, or level of the returned power (right axis)

The reflected power normalised by the incident power is $|\Gamma|^2$, and this is also known as the *return loss*, and similarly the normalised transmitted power is

$$|T|^2 = 1 - |\Gamma|^2 = \frac{4\,VSWR}{(VSWR+1)^2}. \tag{8.1.28}$$

The reflected power and return loss, both relative to the incident power, are plotted in Figure 8.1.7 to provide a convenient conversion between the match parameters. $VSWR \leq 2$ is a ubiquitous criterion for defining an impedance match, although for high-power broadcast applications, a return loss of 50 dB is sometimes used and this is particularly demanding to realise across a band.

8.1.2.3 Radiation efficiency (antenna efficiency)

If an antenna has a matching circuit, then this is typically considered part of the antenna, and the ohmic losses of the matching circuit are considered as contributing to losses of the antenna.

In an equivalent series network, the radiated power can be represented as the power dissipated in a resistor, R_{rad}, and the ohmic losses in the antenna as R_{ohmic}, and the losses in any matching circuit as R_{match}. Figure 8.1.8 includes the network model. The *radiation efficiency* is the proportion of power radiated to the total power

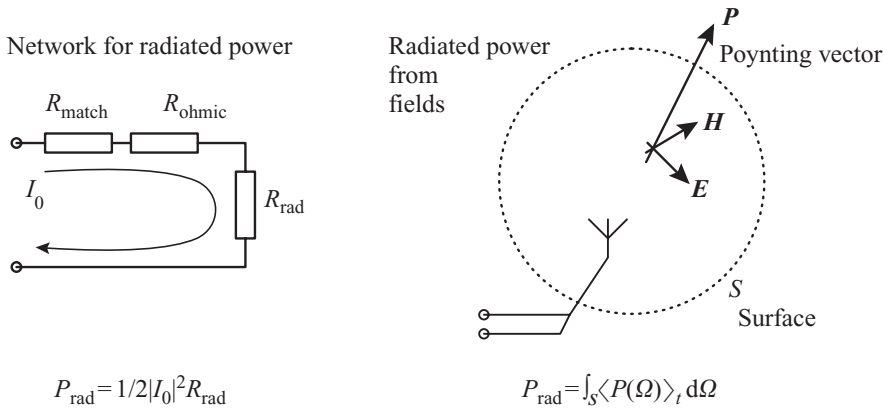


Figure 8.1.8 The radiated power from the tuned network terminal quantities is equated to the radiated power derived from field quantities for defining parameters such as the radiation resistance

entering the matching network plus antenna, and this proportion reduces to

$$\eta_{rad} = \frac{R_{rad}}{R_{rad} + R_{ohmic} + R_{match}}. \tag{8.1.29}$$

This states that if the radiation resistance is small, then the losses in any matching circuit become important for maintaining the antenna efficiency, and this is often the case with compact antennas.

It is obviously a practical advantage if a matching network can be avoided altogether by arranging the antenna feed structure, or by making the matching network as an integrated part of the antenna structure (such as adding wires, stubs, etc.)

In the context of the Friis transmission equation, there must be a radiation efficiency term for both the transmit and receive antennas.

8.1.2.4 Antenna radiation efficiency

Finally, the *antenna efficiency*, sometimes called the *antenna radiation efficiency*, used to relate the directive gain to the directivity above, is just

$$\eta_{ant} = \frac{R_{rad}}{R_{rad} + R_{ohmic}}, \tag{8.1.30}$$

and so this is the same as the radiation efficiency if there is no loss in any matching circuit. The antenna efficiency accounts for the ohmic losses only in the antenna itself. In the Friis equation, this term is not to be included if the radiation efficiency is already included.

The similarity of these terms is confusing, but the accepted definition (*IEEE Standard Test Procedures for Antennas*, p. 113) is that the antenna efficiency does not include losses of the matching circuit. These are often significant, in particular for compact antennas. For an antenna measurement, the matching losses are not included

in the gain, but for writing out the Friis transmission equation (below), the matching losses need to be included, for example using η_{rad} instead of η_{ant}.

In handset antennas for mobile communications, patterns are often presented which include the effects (including the losses) of the human presence. It is tempting to incorporate these losses into the antenna efficiency, but strictly speaking, such losses are strictly part of the path loss of the mobile communications channel. Only if the human is explicitly taken as an integral part of the antenna, can the antenna efficiency include the losses of the human body. The mean effective gain, discussed in Section 8.2, is an appropriate antenna gain quantity for such situations.

8.1.3 Radiation resistance from the Poynting vector

The radiation resistance is defined from the radiating fields by integrating the outgoing power over space. The Poynting vector, in W m^{-2}, is used for this, defined from the fields at a point in space, and is in the direction which is orthogonal to both fields,

$$\boldsymbol{P} = \boldsymbol{E} \times \boldsymbol{H}^* \quad \text{(instantaneous, complex Poynting vector).} \tag{8.1.31}$$

Normally, only the harmonic time dependent, time averaged (over several rf cycles) form is of interest. The real part of the time averaged Poynting vector leaving a volume gives the radiated power from the volume and in the far-field region, the radiation power pattern is proportional to $|\boldsymbol{E}(\theta, \phi)|^2$.

To show this, it is convenient to express the vector spatial fields in terms of their vector complex envelope, for example: $\boldsymbol{E}(x, y, z, t) = \text{Re}\{\boldsymbol{E}_{\text{CE}}(x, y, z) \exp(\text{j}\omega t)\}$, and the subscript 'CE' means (vector) complex envelope. The complex, instantaneous, Poynting vector is now written as

$$\begin{aligned}
\boldsymbol{E} \times \boldsymbol{H}^* &= \tfrac{1}{4}(\boldsymbol{E}_{\text{CE}}\, \text{e}^{\text{j}\omega t} + \boldsymbol{E}^*_{\text{CE}}\, \text{e}^{-\text{j}\omega t}) \times \tfrac{1}{4}\left(\boldsymbol{H}_{\text{CE}}\, \text{e}^{\text{j}\omega t} + \boldsymbol{H}^*_{\text{CE}}\, \text{e}^{-\text{j}\omega t}\right) \\
&= \tfrac{1}{4}(\boldsymbol{E}_{\text{CE}} \times \boldsymbol{H}^*_{\text{CE}} + \boldsymbol{E}^*_{CE} \times \boldsymbol{H}_{\text{CE}}) \\
&\quad + \tfrac{1}{4}(\boldsymbol{E}_{\text{CE}} \times \boldsymbol{H}_{\text{CE}}\, \text{e}^{\text{j}2\omega t} + \boldsymbol{E}^*_{\text{CE}} \times \boldsymbol{H}^*_{\text{CE}}\, \text{e}^{\text{j}2\omega t}) \\
&= \tfrac{1}{2}\text{Re}\{\boldsymbol{E} \times \boldsymbol{H}^*\} + \tfrac{1}{2}\text{Re}\{\boldsymbol{E} \times \boldsymbol{H}\}.
\end{aligned} \tag{8.1.32}$$

The first term $1/2\ \text{Re}\left\{\boldsymbol{E} \times \boldsymbol{H}^*\right\}$, is independent of time. (The factor of 1/2 is from $\boldsymbol{E}$ and $\boldsymbol{H}$ denoting peak values, and is omitted when using effective value notation.) This is the *average power density* and gives rise to the radiated power. The second term, $1/2\ \text{Re}\{\boldsymbol{E} \times \boldsymbol{H}^*\}$, is time dependent, as indicated in the second line of the equation. It is changing at a rate of 2ω, and represents a flow of energy density in the stored fields back and forth through space which does not contribute to the radiation.

The time averaged total power leaving the volume with surface S with outward-pointing normal, $\hat{n}$, is

$$\begin{aligned} P_{\text{rad}} &= \left\langle \int_S \boldsymbol{P} \cdot \hat{n}\, \mathrm{d}S \right\rangle_t \\ &= \int_S \left\langle \left(\frac{1}{2}\mathrm{Re}\{\boldsymbol{E} \times \boldsymbol{H}^*\} + \frac{1}{2}\mathrm{Re}\{\boldsymbol{E} \times \boldsymbol{H}\} \right) \cdot \hat{n} \right\rangle_t \mathrm{d}S \\ &\approx \frac{1}{2Z_0} \int_0^{2\pi} \int_0^{\pi} \left(|E_\theta(\theta, \phi)|^2 + |E_\phi(\theta, \phi)|^2 \right) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \quad \text{(radiated far field).} \end{aligned} \tag{8.1.33}$$

In the second line of this equation, the second term, representing the stored fields, drops out in the averaging process to give the radiated power. The approximation in the third line is for including only the far fields, where a plane wave has formed and has the field relations $E/H = Z_0$.

The radiated power is used for defining the radiation resistance for current excited systems;

$$\tfrac{1}{2}|I_0|^2 R_{\text{rad}} = P_{\text{rad}} \tag{8.1.34}$$

where I_0 is the peak input current at the antenna terminals. If effective values are used, then the half drops out and the equation is of course identical to the d.c. version of Ohm's law. Similarly, for voltage excited systems, the input conductance is well defined by $1/2|V_0|^2 G_{\text{rad}} = P_{\text{rad}}$ where for a matched, lossless antenna, V_0 is the peak voltage at the feed. The technique of equating the power at the terminals with that of the radiated fields is also used in classical arrays to find impedance quantities (Collin and Zucker, 1969). The situation is depicted in Figure 8.1.8.

8.1.4 Relation between directivity, effective area, and complex height

The gain of an antenna is defined from its transmit mode of operation in a line-of-sight, single-path, far-field situation. The principle of reciprocity allows a link between such transmit parameters and receiving parameters such as the complex height (see Chapter 2) and the effective receiving area (see Appendix A).

The Friis transmission equation for the received power in the receiver load, P_{R}, can be written $P_{\text{R}} = P_{\text{T}} G_{\text{T}} G_{\text{R}} G_{\text{Path}}$, where P_{T} is the transmitted power from the transmit antenna, G_{R} and G_{T} are the receive antenna and transmit antenna gains respectively, and G_{Path} is the path gain which is the reciprocal of the path loss, $G_{\text{Path}} = 1/L_{\text{Path}}$. In this form, it is assumed that the antennas are polarisation matched and that their loads are matched. If this is not the case, then the Friis equation is written out, using the load match and polarisation efficiency factors above, as

$$\begin{aligned} P_{\text{R}} = P_{\text{T}} \eta_{\text{pol}} \left(1 - |\Gamma_{\text{T}}|^2\right) \left(1 - |\Gamma_{\text{R}}|^2\right) \eta_{\text{antT}} D_{\text{T}}(\theta_{\text{T}}, \phi_{\text{T}}) \\ \times \eta_{\text{antR}} D_{\text{R}}(\theta_{\text{R}}, \phi_{\text{R}}) \left(\frac{1}{4\pi(d/\lambda)} \right)^2, \end{aligned} \tag{8.1.35}$$

where the directivity, $D(\theta, \phi) = G(\theta, \phi)/\eta_{\text{ant}}$ has been used and the free-space path loss is given for a distance d between antennas operating with a wavelength λ.

Note that by using only η_{ant} in this equation, instead of η_{rad}, there is no provision here for losses in the matching circuit. A factor can be included on the right-hand side,

$$\eta_{\text{C}} = \frac{\eta_{\text{rad}}}{\eta_{\text{ant}}} = \frac{1}{1 + R_{\text{match}}/(R_{\text{rad}} + R_{\text{ohmic}})} \tag{8.1.36}$$

for each antenna, which then accounts for any matching circuit losses. The result is the replacement of η_{ant} with η_{rad} for each antenna in the above form of the Friis equation, or alternatively stated, the ohmic losses, R_{ohmic}, now include the losses of the matching circuit, R_{match}.

The transmit antenna sets up a flow of power density at the receive antenna, of $S = P_{\text{T}} G_{\text{T}}/(4\pi d^2)$ W m^{-2}. The receive antenna can be represented as an area in space which captures this flow of power density. The power intersected by the area is the power delivered to the load at the receiver. Denoting the area A_{R}, the received power is $P_{\text{R}} = SA_{\text{R}}$ and the effective area follows as

$$A_{\text{R}}(\theta, \phi; \eta_{\text{pol}}; \Gamma_{\text{R}}; R_{\text{ohmic}}) = \frac{\lambda^2}{4\pi}(1 - |\Gamma_{\text{R}}|^2)\eta_{\text{pol}}\eta_{\text{radR}} D(\theta, \phi) \tag{8.1.37}$$

where the dependencies on the receiving antenna are made explicit: direction, polarisation match, load match and ohmic losses in both the antenna and any matching circuit.

The complex height (Chapter 2), or length, of an antenna is defined from the open circuit voltage of a receiving antenna,

$$V_{\text{O}} = \boldsymbol{E} \cdot \boldsymbol{h} = |\boldsymbol{E}||\boldsymbol{h}|\sqrt{\eta_{\text{pol}}} \tag{8.1.38}$$

and the incident electric field given as the power density, $S = 1/2|\boldsymbol{E}|^2/Z_0$, where here the peak-value notation is used for E. The equivalent circuits for the transmitting and receiving antenna modes are given in Figure 8.1.9. For a matched receiver load, $R_{\text{L}} = R_{\text{rad}} + R_{\text{ohmic}}$ (again, R_{ohmic} includes losses in any matching circuit) and $X_{\text{L}} = -X_{\text{A}}$.

The current in the receiving antenna is

$$I_{\text{R}} = \frac{V_{\text{O}}}{(R_{\text{rad}} + R_{\text{ohmic}} + \text{j}X_{\text{A}}) + (R_{\text{L}} + \text{j}X_{\text{L}})} \tag{8.1.39}$$

and the received power, which is that dissipated in the load resistance, is

$$\begin{aligned} P_{\text{R}} &= \tfrac{1}{2}|I_R|^2 R_{\text{L}} \\ &= \frac{1}{4}\frac{|V_{\text{O}}|^2}{R_{\text{rad}}} \quad \text{(lossless, matched)} \end{aligned} \tag{8.1.40}$$

again using peak value notation.

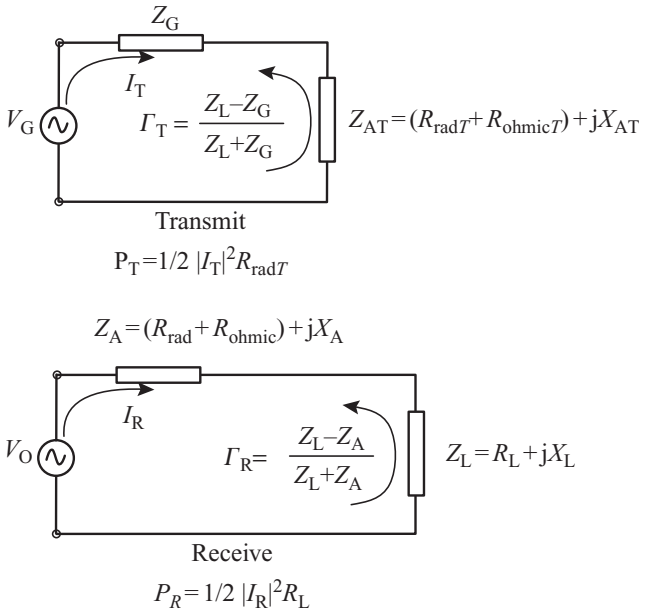


Figure 8.1.9 *The equivalent circuits for transmitting and receiving antennas. The Friis transmission equation, linking the transmit and receive quantities, is for a single line-of-sight far-field path between antennas. The transmit power is the power from the generator that is dissipated in the radiation resistance,* R_{radT}. *The effective receiving area intercepts the power density flow and captures the power which is that dissipated in the load resistance,* R_{L}

Equating the received powers from the circuit representation and from the power density approach, i.e. $|E|^2/Z_0 A_{\text{R}} = |I_{\text{L}}|^2 R_{\text{L}}$, results in

$$A_{\text{R}}(\theta, \phi; \eta_{\text{pol}}; \Gamma_{\text{R}}; R_{\text{ohmic}}) = \frac{R_{\text{L}} Z_0}{(R_{\text{rad}} + R_{\text{ohmic}} + R_{\text{L}})^2 + (X_{\text{A}} + X_{\text{L}})^2} \eta_{\text{pol}} |\boldsymbol{h}(\theta, \phi)|^2$$

$$A_{\text{R}}(\theta, \phi; \eta_{\text{pol}} = 1; \Gamma_{\text{R}} = 0; R_{\text{ohmic}} = 0) = \frac{1}{4} \frac{Z_0}{R_{\text{rad}}} |\boldsymbol{h}(\theta, \phi)|^2 \quad \text{(lossless, matched).} \tag{8.1.41}$$

Using this, the effective height can be related to the directivity and gain, here for matched and lossless conditions, as

$$|\boldsymbol{h}(\theta, \phi)|^2 = 4 \frac{R_{\text{rad}}}{Z_0} \frac{\lambda^2}{4\pi} D(\theta, \phi) \quad \text{(lossless, matched).} \tag{8.1.42}$$

The effective area does not represent the total power captured by the receiving antenna. Even if the antenna is lossless and matched in polarisation and load, half of the power captured is dissipated in the receiving antenna resistance of the equivalent circuit, $R_{\text{A}} = R_{\text{rad}}$. Physically, this power is irretrievably scattered by the receiving antenna. So the capture area representing the total power gathered by the receiving antenna is always at least twice the effective area.

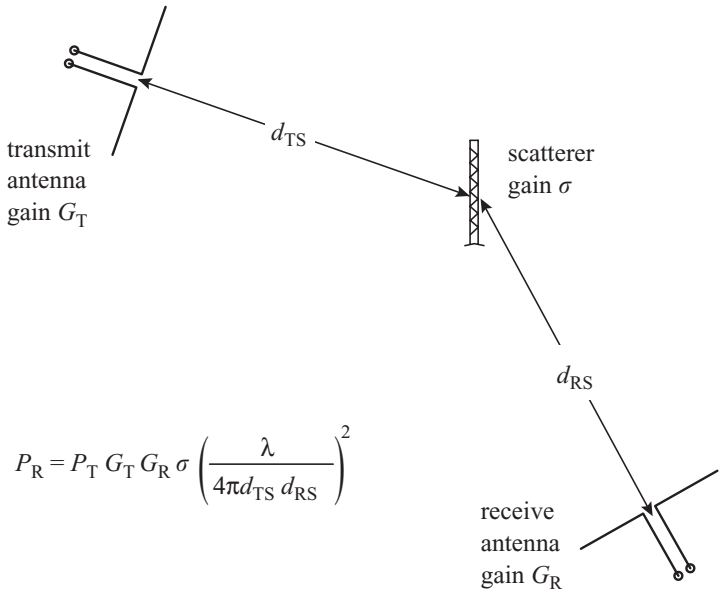


Figure 8.1.10 The equivalent gain of a single, far-field scatterer is the radar cross section

8.1.4.1 Effective gain of scatterer (radar cross section)

A single scatterer defining the only path between a transmitting and a receiving antenna can be incorporated into the Friis equation. The scatterer has an effective gain, often referred to as the radar cross section, denoted σ. This appears in a two-stage form: one from the transmitting antenna to the scatterer which is distance d_{TS}; the other from the scatterer to the receiving antenna, which is distance d_{RS}; see Figure 8.1.10. The Friis transmission equation becomes, here for matched, lossless antennas,

$$P_R = P_T \eta_{pol} \sigma G_T(\theta_T, \phi_T)\, G_R(\theta_R, \phi_R) \left(\frac{1}{4\pi (d_{TS}/\lambda)(d_{RS}/\lambda)}\right)^2, \qquad (8.1.43)$$

and this form is also known as the radar range equation. The polarisation efficiency is defined from the receiving antenna polarisation, and the polarisation of the wave incident on the receiving antenna, and so the polarisation of the transmitter does not appear explicitly.

8.1.5 Antenna temperature

Apart from the wanted signal, many random sources illuminate receiving antennas and contribute noise power to the received signal; this is called the antenna noise. The radiation sources include extra-terrestrial (galactic and solar, etc.), as well as from the atmosphere, the ground and the environment in general. The contribution from the transmission line connecting an antenna to the receiver is also included.

For most mobile communications systems, the antenna noise is negligible since it is often below the thermal noise power contribution of the receiver electronics, and certainly well below interference in mass communications systems. However, in satellite mobile communications, for example, where the signal strength can be very weak despite a line-of-sight situation, it should be accounted for.

At radio frequencies, the power emitted by the black body towards the antenna, or the *brightness*, is (Rayleigh–Jeans law) $B = 2k_{\mathrm{B}}Tf^2/c^2$ W m^{-2} steradian^{-1} cycle^{-1} (Collin and Zucker, 1969), where k_{B} is Boltzmann's constant $= 1.38(10)^{-23}$ J degree^{-1}, $c \approx 3(10)^8$ m s^{-1}, and f is the radio frequency in Hz. The law can be expressed as the one-sided power spectral density of the thermal voltage across a resistor R at temperature T, over a frequency band Δf over which the resistor is taken as constant, as $V_{\mathrm{eff}}^2 = P = 4k_{\mathrm{B}}TR\Delta f$ (Nyquist's formula). In practice the frequency band is often calculated as an effective bandwidth, which passes the equivalent amount of flat spectral power as the more complicated bandpass shape or variable spectral densities of real-world systems.

The brightness is treated as a distribution, $B(\theta, \phi)$, and the received noise power, over a bandwidth Δf, at a perfectly matched antenna with gain $G(\theta, \phi)$, is (Collin and Zucker, 1969)

$$N_{\mathrm{rec}} = \frac{\Delta f}{2} \iint B(\theta, \phi) G(\theta, \phi) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \tag{8.1.44}$$

where the factor of two arises from the polarisation being considered as random, giving a polarisation efficiency of $\eta_{\mathrm{pol}} = 1/2$.

If the brightness distribution can be observed at the antenna, then the brightness can be related to the time averaged incident fields from the black body, $\boldsymbol{E}_{\mathrm{inc}}^{(\mathrm{b})}(\theta, \phi)$ as

$$B(\theta, \phi)\Delta f = \frac{1}{2\eta_0} \left\langle \left| \boldsymbol{E}_{\mathrm{inc}}^{(\mathrm{b})}(\theta, \phi) \right|^2 \right\rangle_t . \tag{8.1.45}$$

An effective temperature distribution, $T_{\mathrm{b}}(\theta, \phi)$, as observed at the antenna, can be defined and the *antenna temperature* is then the mean of this temperature distribution,

$$T_{\mathrm{A}} = \frac{\int_0^{2\pi} \int_0^{\pi} T_{\mathrm{b}}(\theta, \phi) G(\theta, \phi) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi}{\int_0^{2\pi} \int_0^{\pi} G(\theta, \phi) \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi}. \tag{8.1.46}$$

The antenna temperature is the temperature at which the radiation resistance of the antenna produces thermal noise power equivalent to the power received by the antenna from the black body radiation. The power available from the antenna due to an antenna temperature T_{A} is $k_{\mathrm{B}}T_{\mathrm{A}}\Delta f$.

The temperature distribution arises from two basic components, the sky noise and the ground noise, and this breakdown is of most interest for satellite communications. For landmobile communications, the antenna may be operating indoors, for example, and here the temperature distribution is essentially constant at the ambient ground temperature, $T_{\mathrm{b}}(\theta, \phi) = T_0 = 290$ K, and so the antenna temperature is about $T_{\mathrm{A}} = T_0$, independent of the pattern. This is analogous to the mean effective gain of an

antenna illuminated by distributed scatterers which radiate the wanted signal instead of noise signals of the temperature distribution. As with the antenna temperature, the mean effective gain becomes independent of the pattern when the pattern is confined to a uniform, continuous distribution of incoming waves; see Section 8.2. With the brightness, however, the temperature of the surface is not necessarily the whole story; for example, a highly reflective surface such as a groundplane provides an image to the receiving antenna of areas with different brightness to that of the groundplane.

8.1.5.1 Sources of noise

The sky noise comprises many different contributing mechanisms. The result is that the sky noise is strongly dependent on frequency over the spectral bands for communications. However, over a given communications channel band, which comprises a narrow relative bandwidth, the temperature is constant with frequency, as seen above.

The sky noise and typical man-made noise is given in Figures 8.1.11 (Chapter 34 of Jordan, 1985). At lower frequencies, the dominant natural source is atmospheric noise (waves from distant lightning propagating via ionospheric reflection), but the man-made noise, comprising interference from electrical machines and harmonics of the a.c. power system and ignition, etc., dominates over much of the microwave communication band. At the very low frequencies, the man-made noise is typically in the near field of an antenna and both the magnetic and electric fields may need to be measured separately.

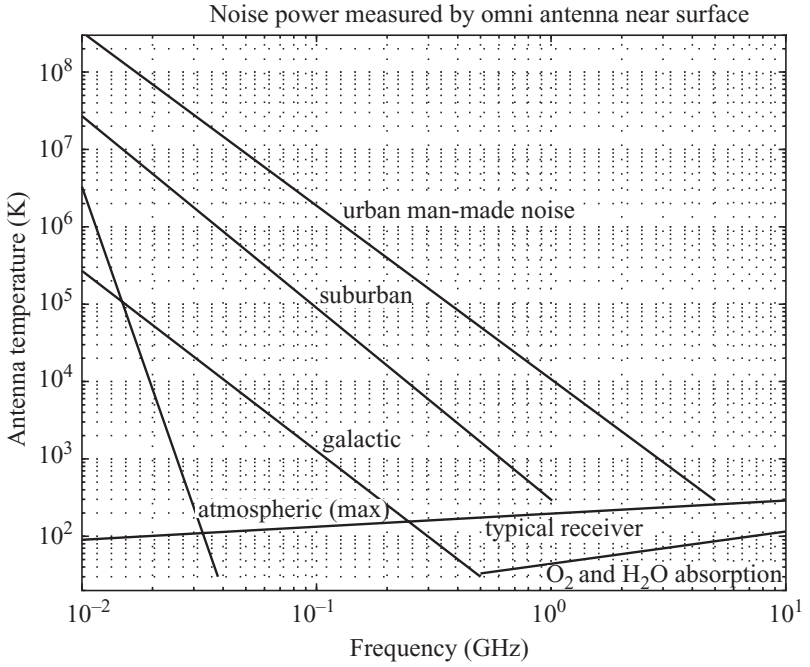


Figure 8.1.11 The noise power as antenna temperature, measured using an omni antenna near the ground and taken from the median of the average values (after Jordan, 1985)

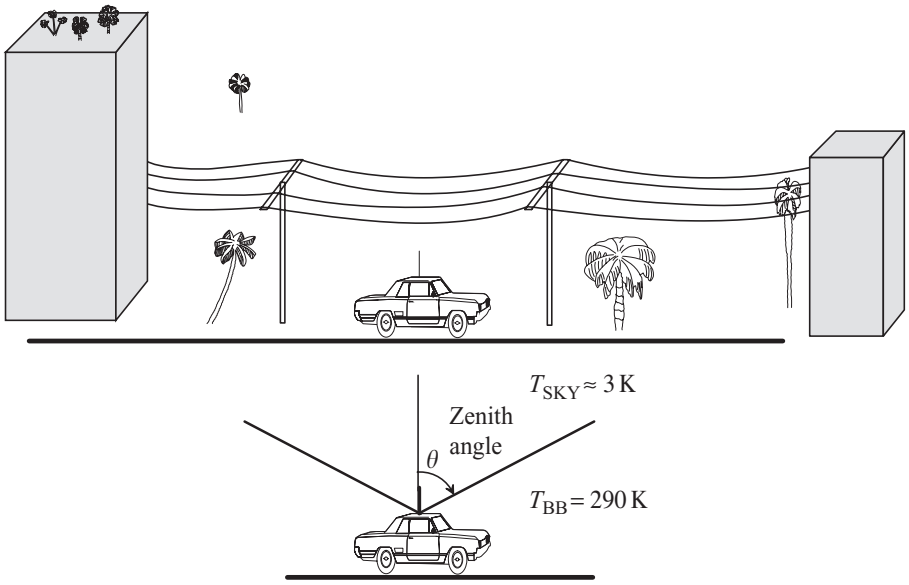


Figure 8.1.12 A rotationally symmetric model for the average temperature distribution 'seen' by a rooftop-mounted satellite receiving antenna

8.1.5.2 Gain-to-temperature and modelled temperature distribution for mobile satellite antennas

A receiver's performance figure-of-merit is the gain-to-temperature ratio, G/T. This ratio is of particular interest for satellite antennas, where the signal level, although probably comprising a line-of-sight path, is typically low owing to the large space loss and the power constraints of satellite systems. The antenna temperature can be estimated from the receiving power pattern and a model of the temperature distribution, as can be seen from the defining equation above. As an example, a simple averaged temperature distribution model for rooftop-mounted vehicular mobile satellite terminals is described (Vaughan and Scott, 1993). The model distribution is rotationally symmetric, with the vehicle antenna sitting in a conical 'landscape' horizon, and the angle to the horizon is estimated empirically. The landscape has a temperature of 290 K and above the horizon, the temperature is 3 K. The situation is depicted in Figure 8.1.12.

Using experimental results for the G/T from Bell *et al*. (1988), and their antenna pattern, empirical estimates of the horizon angles for different environments can be found. Table 8.1 gives these, along with estimates from photographs of the environments given by Bell *et al*. This set of horizon angles can be used for the simple temperature distribution model for estimating the average antenna temperature for other rotationally symmetric antenna patterns operating in the various environments.

8.1.6 Noise figure contributions from connections to a receiver

8.1.6.1 Antenna with transmission line

A lossy transmission line connecting a lossless antenna to the receiver (load) also contributes to the noise at the receiver. For a line of length L and line physical

Table 8.1 Table of zenith-to-horizon angles for the temperature distribution model. The measurements are from Bell et al. (1988), taken at 1.5 GHz

Environment used for measurement of G/T	Zenith angle of artificial horizon from photos of environment (degrees)	Measured temperature (Kelvin)	Computed temperature using angle estimates from photos of environment and receiving pattern (Kelvin)
Hill top	100	25 ± 6	30
Rural field	90	25 ± 6	40
Main street, small town	85	45 ± 7	50
Tree-lined residential street	75	25 ± 6	90
Open land with trees and power lines	70	80 ± 10	120
Between three-storey buildings	60	175 ± 30	175

temperature T_{L}, attenuation α Np m^{-1}, the power loss through the line is $\exp(-2\alpha L)$, and the antenna (plus line) temperature seen at the input to the receiver is

$$T_{\mathrm{AL}} = T_{\mathrm{A}}\,\mathrm{e}^{-2\alpha L} + T_{\mathrm{L}}\left(1 - \mathrm{e}^{-2\alpha L}\right). \tag{8.1.47}$$

The noise power available to the receiver, under matched conditions, is now $N_{\mathrm{R}} = k_{\mathrm{B}} T_{\mathrm{AL}} \Delta f$, and T_{AL} is the temperature of a thermal noise-producing resistor matched to the input of the receiver; see Figure 8.1.14 below. The antenna temperature contribution, as seen by the receiver, is reduced by the line attenuation but the noise contribution from the line can dominate for satellite antennas. As an example, for an antenna temperature of 80 K, a line temperature of $T_{\mathrm{L}} = T_0 = 290$ K, and line attenuation 0.1 dB, (i.e. line gain $\exp(-2\alpha L) = 0.9772$), then $T_{\mathrm{AL}} = 78.2\,\mathrm{K} + 6.6\,\mathrm{K}$; but for a 10 dB line attenuation, $T_{\mathrm{AL}} = 8\,\mathrm{K} + 261\,\mathrm{K}$.

8.1.6.2 System temperature

The *system temperature*, which includes the noise contributions from the receiver, should be referred to the antenna terminals rather than to the input to the receiver as given above, and this follows as

$$T_{\mathrm{sys}} = T_{\mathrm{A}} + T_{\mathrm{L}}\left(\frac{1}{\mathrm{e}^{-2\alpha L}} - 1\right) + \frac{T_{\mathrm{R}}}{\mathrm{e}^{-2\alpha L}} \tag{8.1.48}$$

where the noise of the receiver, which has temperature T_{R}, is now included (Kraus, 1988). The calculation of T_{R} is given below.

If the line is infinitely lossy, i.e. the line gain is zero, or $\exp(-2\alpha L) = 0$, then the system temperature becomes infinite and the power incident at the receiver terminals of course contains no signal from the antenna. If the antenna and line are the same temperature, $T_A = T_L = T_0$, which can be the case for terrestrial mobile communications, then the system temperature reduces to $T_{sys} = L_L(T_0 + T_R)$, where $L_L = \exp(2\alpha L)$ is the line loss.

For a lossy antenna, the system temperature is found the same way. At the antenna terminals, the system temperature has an additive factor, $T_{ant}(1/\eta_{ant} - 1)$, where $\eta_{ant} = 1/L_A$ is the antenna efficiency and T_{ant} is the physical temperature of the antenna. The gain-to-temperature ratio follows as

$$\frac{G}{T} = \frac{D\eta_{ant}}{T_A + (L_A - 1) + L_A T_{rec}}. \tag{8.1.49}$$

For a G/T of multiple antennas with a beam former network, see Kraft (2000).

8.1.6.3 Noise figure with circuits

Referring the noise output of a component to be referred back to the input allows the noise contributions from the various stages to be compared and optimised. This 'back-referencing' is conveniently handled using the *noise factor, F*, often called the *noise figure* when expressed in dB, $F_{dB} = 10\log_{10} F$. F is defined as the ratio of noise powers at the output of a circuit, and it can be instructively written as a ratio of SNRs for any circuit element,

$$\begin{aligned} F &= \frac{\text{total output noise power of circuit when source is at } T_0^0}{\text{output noise power of circuit due to only the source at } T_0^0} \\ &= \frac{\text{SNR available from source at input of circuit}}{\text{SNR available at output of circuit}}. \end{aligned} \tag{8.1.50}$$

F is always greater than unity in practice (in the case of an idealised circuit element which contributes no noise, $F = 1$) and it represents a power loss term that can be cascaded corresponding directly to cascaded link components or circuits as shown below.

If a circuit element with power gain G_C contributes noise power N_C, then for input signal power P_{IN}, the noise factor of the circuit is

$$F_C = \frac{P_{IN}/(k_B T_0 \Delta f)}{G_C P_{IN}/(G_C k_B T_0 \Delta f + N_C)} = 1 + \frac{N_C}{G_C k_B T_0 \Delta f} \tag{8.1.51}$$

so the noise power contributed by the circuit alone is

$$N_C = G_C(F_C - 1)k_B T_0 \Delta F. \tag{8.1.52}$$

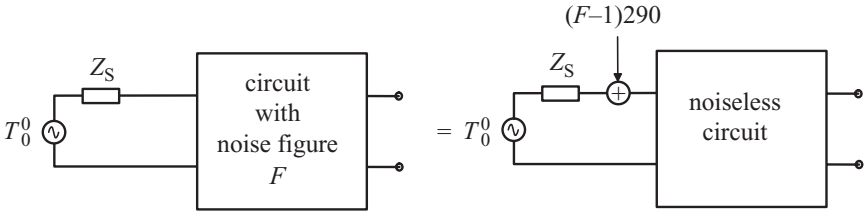


Figure 8.1.13 The use of the noise figure to refer the noise contribution of a circuit to the source, here with $T_0 = 290$ K

This noise power appears at the output of the circuit, but can be referred to the input of the circuit as the available power from a resistor, matched to the input, which is at temperature

$$T_C = (F_C - 1)T_0. \tag{8.1.53}$$

This is the relation between the temperature of a device and its noise figure. The contribution of the circuit-generated noise to the source is shown in Figure 8.1.13

Applying the principle to the transmission line,

$$F_L = 1 + \left(\frac{1 - L_L}{L_L}\right)\frac{T_L}{T_0} \tag{8.1.54}$$

and if the line temperature is the same as T_0, then the noise figure reduces to the gain of the line, $F_L = 1/L_L$.

8.1.6.4 Cascade circuits

The relations for cascade of receiver circuit elements are depicted in Figure 8.1.14. Here a cascade of two networks (e.g. a two-stage amplifier) is attached to the antenna and transmission line.

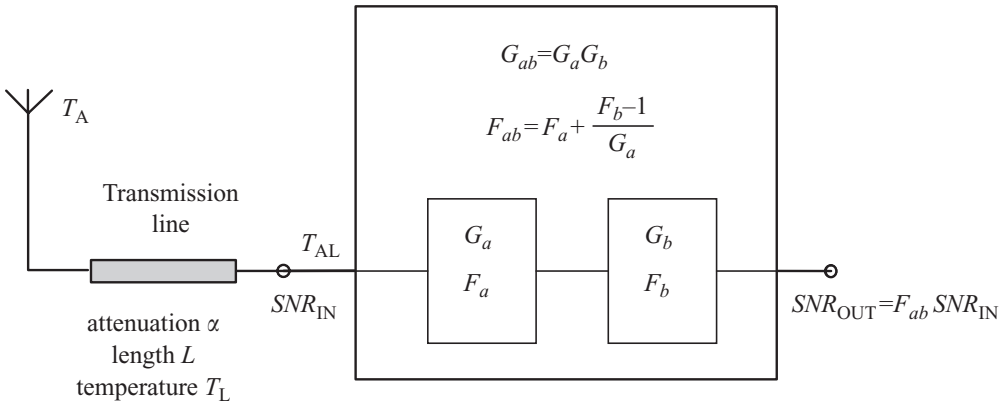


Figure 8.1.14 Antenna with antenna temperature T_A attached to a transmission line and then to a cascade of two networks. Further cascade networks are conveniently included using the noise figure

If the input noise power to the cascade is just $k_B T_0 \Delta f$, then the total noise power at the output of b is

$$N_{ab} = F_{ab} G_{ab} (k_B T_0 \Delta f) = F_{ab} G_a G_b (k_B T_0 \Delta f). \tag{8.1.55}$$

The output noise from network a appearing at the output of b is

$$N_{b|a} = N_a G_b = F_a G_a G_b (k_B T_0 \Delta f). \tag{8.1.56}$$

The noise added by b is

$$N_{b|b} = (F_b - 1) G_b (k_B T_0 \Delta f) \tag{8.1.57}$$

and from the total noise power being $N_{ab} = N_{a|a} + N_{b|b}$, the noise figure for the cascade is

$$F_{ab} = F_a + \frac{F_b - 1}{G_a}. \tag{8.1.58}$$

For a cascade of circuits with gains $G_1, G_2, \ldots$, so the overall power gain is $G = G_1 G_2 G_3 \ldots$, and with noise figures $F_1, F_2, \ldots$, then the overall noise figure for the cascade system is

$$F_{\text{sys}} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + \frac{F_4 - 1}{G_1 G_2 G_3} + \cdots, \tag{8.1.59}$$

or in terms of the temperature, the system temperature of the cascade receiver is

$$T_{\text{sys}} = T_{\text{R}} = T_1 + \frac{T_2}{G_1} + \frac{T_3}{G_1 G_2} + \frac{T_4}{G_1 G_2 G_3} + \cdots. \tag{8.1.60}$$

The use of a low noise amplifier (LNA) close to the antenna can reduce the system noise. The configuration is shown in Figure 8.1.15. As long as the gain is well over unity and the noise figure F is small (typically $F_{\text{dB}} = 2$–$10\,\text{dB}$), then the LNA performance dominates the overall noise performance of the receiver system. The problem of the noise introduced by the cable, which can otherwise dominate system performance, is mitigated.

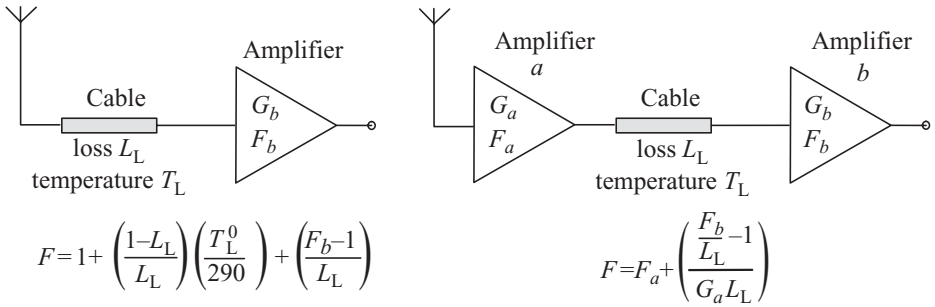


Figure 8.1.15 The noise figure for an antenna connected by cable to an amplifier, and for the configuration of using a first, 'masthead', amplifier. The first amplifier, with a large G_a and low F_a, dominates the overall noise figure

8.1.7 Groundplane principles for antennas

The infinite, perfectly conducting groundplane concept is an important design tool for antennas. The groundplane acts to short circuit the tangential electric fields, which in turn produces an image of an antenna as seen from the upper hemisphere. So an antenna to be realised on or above a groundplane, operating in the upper hemisphere only, can be designed or modelled as the antenna plus its image operating in the whole sphere, and vice versa. In practice, the groundplane can be the platform for the antenna, such as in a vehicular rooftop mounting, or wall-mounted base station, or even on a facet of a handheld terminal, if the facet is large enough electrically. An artificial groundplane can be realised with wires, rather than a solid surface, and for sparse wires, the wires must be aligned with the currents that would flow on a solid surface. (For example, see the wire groundplane below in Figure 8.1.22.) The effect on impedance of a finite-sized groundplane is discussed further below.

8.1.7.1 Source elements

The groundplane acts to short circuit the tangential electric fields. It follows that any electric current source, such as a current on a wire, will have an image but with a sign change for the horizontal components (cf. a reflection coefficient of $R = -1$: see Section 3.1) and no sign change for the vertical components ($R = +1$). Examples are depicted in Figure 8.1.16. It is clear that tangential electric current images cancel when the current is on the groundplane, and that the magnetic currents add (i.e. double) when they are on the groundplane.

8.1.7.2 Antenna impedances

The impedance of antennas in a hemisphere relative to their image arrangements in the full sphere, is halved. This can be seen, for example, in a monopole above

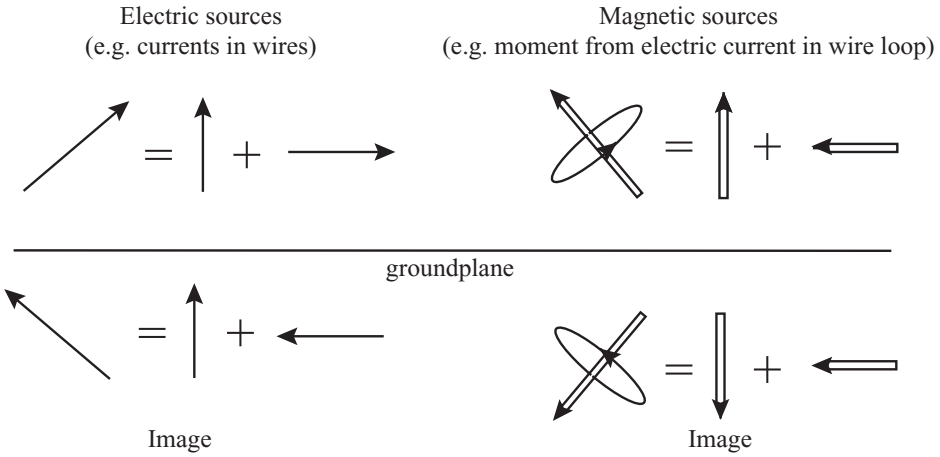


Figure 8.1.16 The perfectly electrically conducting, infinite groundplane gives images of electrical sources such that the tangential electric fields on the surface of the groundplane are zero

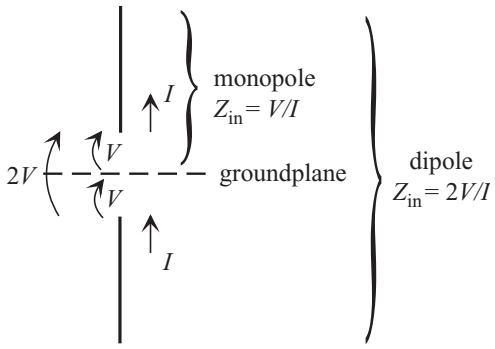


Figure 8.1.17 The antenna over the groundplane operating in a hemisphere has half the voltage excitation but the same current as its full-sphere counterpart, so its impedance is half of that of the full-sphere antenna

a groundplane which has half the feed voltage relative to the monopole plus its image, i.e. a dipole; see Figure 8.1.17. If the monopole voltage is V, and the current I, both measured at the groundplane, the input impedance is $Z_{\text{mono}} = V/I$. The dipole input impedance is $Z_{\text{dipole}} = 2V/I$.

An important difference between half-space and full-space antenna equivalent antennas is that the feeding can be implemented in a different way. For example, dipoles lend themselves naturally to balanced feed lines whereas monopoles can be fed directly from a coaxial from under the groundplane.

Because the radiated power is confined to the hemisphere in the presence of an infinite groundplane, the directivity of the hemispheric antenna is doubled relative to its full sphere counterpart. The bandwidth is the same for the hemispherical antenna and the full sphere counterpart.

8.1.8 Generic elements: dipoles and monopoles

The dipole element is generic in a theoretical sense because it is perhaps the simplest antenna and the dipole mode is the simplest radiation mode. The dipole must be fed from a balanced line. The hemispherical counterpart is the monopole on a groundplane which can be fed directly from a coaxial line. Formulas for the impedance of a straight wire in free space are available (see below), and moment methods are established for analysing more complicated configurations. For example, the thin wire, quarter wavelength long monopole has an input impedance of about $36 + \text{j}21\ \Omega$, whereas the thin dipole is about $72 + \text{j}42\ \Omega$. Details of the feeding act to alter the impedance, in particular the reactance. This also applies to mutual impedances. For example, for spaced quarter wave monopoles on a groundplane, the phase of the mutual impedance is some 13° different (SMA connectors at microwave frequencies) from that of spaced dipoles, and the magnitude of the mutual impedance is the same.

If it is important to attain resonance (for wire antennas, the best impedance match, defined by a maximum allowable *VSWR*, to a $50\ \Omega$ line, often occurs away from resonance), then the monopole can be shortened by an amount that depends on the

thickness of the wire. Some values give a feel for the effect (Stutzman and Thiele, 1981). For a very thin wire (length to diameter ratio of $L/D = 5000$), the resonant length is about $L/\lambda = 0.49$. For a thin wire ($L/D = 50$), $L/\lambda = 0.475$, and for a thick wire ($L/D = 10$), $L/\lambda = 0.455$. The bandwidth (discussed in Section 8.2) is also affected by the wire thickness, increasing the thicker the wire. The $VSWR = 2$ relative bandwidth is about 8 per cent for $L/D = 2500$ and about 16 per cent for $L/D = 50$.

8.1.8.1 Folded dipoles

The folded dipole comprises a pair of half-wavelength wires which are joined at the non-feed end. The configuration can be analysed by considering a transmission line mode which is non-radiating, and a dipole (cf. single conductor of the same thickness of wire as one of the folded dipole conductors) mode, or radiating mode; see Figure 8.1.18. The radiation pattern is the same as for the single conductor dipole.

If the transmission line formed by the spaced conductor pairs of each arm of the dipole has characteristic impedance Z_C, then the input impedance of one arm of the shorted transmission line of length $L/2$, seen from the feed, is

$$Z_{TL} = jZ_C \tan\left(k\frac{L}{2}\right) = jZ_C \tan\left(\pi\frac{L}{\lambda}\right). \tag{8.1.61}$$

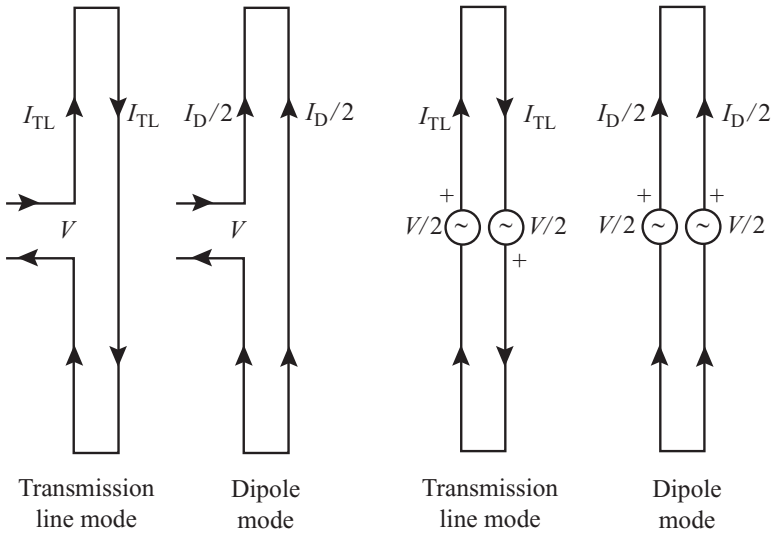


Figure 8.1.18 The folded dipole antenna currents are modelled as the sum of a transmission line mode and a dipole mode. The feed line currents and voltages are the left pair and the circuit model is given by the right pair. The transmission line mode current, I_{TL}, is a travelling wave around the loop and is non-radiating, whereas the radiating dipole mode current, I_D, is divided equally between the two conductors

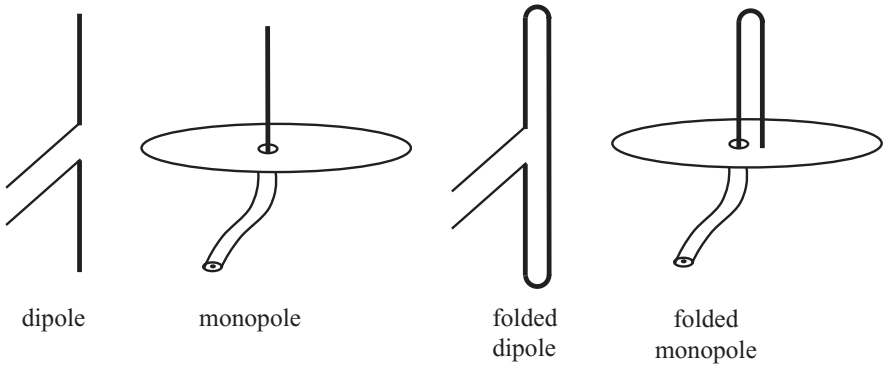


Figure 8.1.19 The dipole and equivalent monopole on a finite groundplane, and the folded dipole and folded monopole

The input current in the feeding conductor is

$$\begin{aligned} I_{\text{in}} &= I_{\text{TL}} + \frac{I_{\text{D}}}{2} \\ &= \frac{V}{2Z_{\text{TL}}} + \frac{V}{4Z_{\text{D}}}, \end{aligned} \tag{8.1.62}$$

where Z_{D} is the input impedance of the dipole and the voltage is divided between the conductors as shown in Figure 8.1.18. The input impedance for the folded dipole follows as

$$Z_{\text{in}} = \frac{4\ Z_{\text{TL}} Z_{\text{D}}}{Z_{\text{TL}} + 2Z_{\text{D}}} = 4Z_{\text{D}} \left(1 + \frac{Z_{\text{D}}}{Z_{\text{TL}}}\right) \tag{8.1.63}$$

so for the folded dipole of total length $L = \lambda/2$, the input impedance is $Z_{\text{in}} \approx 4Z_{\text{D}}$.

This demonstrates an advantage of the configuration: a four-fold increase in the impedance with no appreciable increase in antenna size. The input resistance is almost 288 Ω (four times the dipole input resistance of 72 Ω) and this antenna is often fed with 300 Ω balanced transmission line. For the equivalent folded monopole, the input resistance is about 144 Ω. Furthermore, the different frequency behaviour of the reactance of the two parts also helps in improving the bandwidth. The basic configurations are depicted in Figure 8.1.19.

8.1.8.2 Folded *N*-pole

The folded dipole can be extended to many conductors. Figure 8.1.20 shows a circuit model of the excitation.

The conductors are taken as all the same dimension and closely spaced so that all the currents are the same, $I_n = I_1$, and for simplified analysis, all the self- and mutual-impedances are taken as the same, $Z_{1n} = Z_{11}$. The relation between the currents and the common voltage is

$$\frac{V}{N} = \sum_{n=1}^{N} I_n Z_{\text{in}} \approx N I_1 Z_{11} \tag{8.1.64}$$

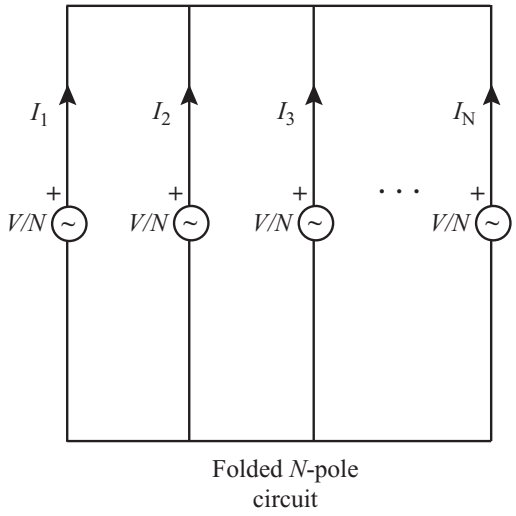


Figure 8.1.20 The folded N-pole antenna has N conductors. For simplified analysis, the conductors are closely spaced such that they have the same current and the self-impedance and mutual impedances are the same

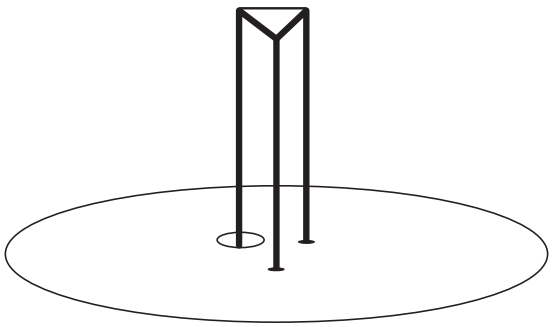


Figure 8.1.21 A monopole realised as a folded tripole on a finite groundplane

and the input impedance follows as

$$Z_{\text{in}} = \frac{V}{I_1} \approx N^2 Z_{11} \approx N^2 Z_{\text{D}}, \tag{8.1.65}$$

i.e. the input impedance is N^2 times the dipole impedance. For the half-wavelength folded tripole, for example, the input impedance at resonance is about $72^*3^2 \approx 650\,\Omega$. The groundplane version has an input resistance of about 325 Ω, and this concept is depicted in Figure 8.1.21. In the tripole configuration, the assumption that the mutual impedances are the same is good, and for the quarter-wavelength, the input impedance is not particularly sensitive to the conductor thickness or spacing.

The above simplified analysis is to show the principles of operation. For an accurate solution, and for different conductor sizes and spacings which give extra degrees of freedom in the design, the moment method should be used. The Goubau antenna

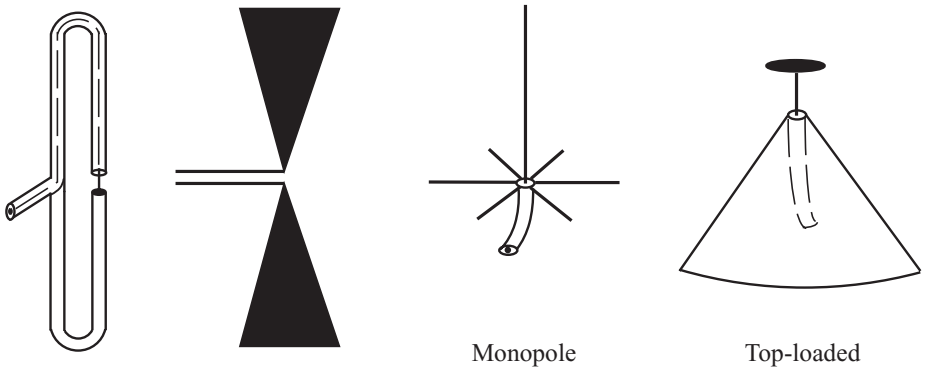


Figure 8.1.22 Variations on the dipole antenna. The top-loaded conical dipole has very wide bandwidths of several octaves and is often used for channel testing

analysis (Section 8.2.4) extends the concept to produce shorter dipoles for a 50 Ω match, in order to produce more compact designs which are also very wideband.

Well-known variations on the dipole include techniques for broadbanding, which usually means some thickening of the generic wire elements, such as in the 'bow-tie' shape, with flat, wide conductors, and its 3D counterpart, the bi-conical dipole. A practical, although electrically large, ultra-wideband antenna is the top-loaded conical form, which comprises a conical 'groundplane' with a top-loaded monopole. Some such variations are given in Figure 8.1.22, including a coax-fed folded dipole.

8.1.8.3 Approximations for the currents on wire dipoles

Accurate analyses for the currents on thin wires can be obtained from moment method formulations. However, simplified approximations allow reasonably accurate calculations for the impedance and the radiation parameters and are insightful for other antennas. For a straight wire in free space, the current distribution is approximated by a sinusoidal form,

$$I(z) = I_{\max} \sin\left(k(L - |z|)\right) \tag{8.1.66}$$

and k is the propagation constant along the wire. The subscript 'max' is used to indicate the maximum value of the current distribution model and this may not be the same as the feed current of the wire antenna. For a conducting wire with no dielectric coating, this is about the same as the free space propagation constant, elaborated below. For finding input impedances $Z_{\text{IN}} = V/I$ of wavelength-long dipoles, more accurate formulas, where the feedpoint current does not go to zero, are required (see below).

If the wire is short ($L \ll \lambda/4$), then the sinusoidal current distribution taper at the end of the wire can be simplified to a linear taper,

$$I(z) = I_{\max}\left(1 - \frac{|z|}{L}\right). \tag{8.1.67}$$

For very short, top-loaded monopoles (see below) the distribution can be simplified to linear, $I(z) = I_{\max}$. The far-field pattern is the same for both cases, but the radiation resistance is different, as shown below.

8.1.9 Fields of a current distribution

The fields of a general 3D current distribution $\boldsymbol{J}$ can be found by integral relations. It is conventional to use (x, y, z) to denote the Cartesian observation space where the fields are being calculated, and (x', y', z') for the space of the current source. Similarly, the polar position of the observation space is $\boldsymbol{r} = (r, \theta, \phi)$ and the sources are at $\boldsymbol{r}' = (r', \theta', \phi')$, and

$$\boldsymbol{R} = \boldsymbol{r} - \boldsymbol{r}' \tag{8.1.68}$$

is the vector between the source position and the observation point, with

$$R = |\boldsymbol{R}| = \sqrt{(x - x')^2 + (y - y')^2 + (z - z')^2}, \tag{8.1.69}$$

and the dependency of R on r' and r is implicit, and discussed further below. The geometry is given in Figure 8.1.23.

8.1.9.1 Basic relations for radiation (far fields) from moving charges

The changing currents on a conductor in space set up fields which radiate. Expressing the changing current as a charge q (in the volume V'), which has an acceleration $\boldsymbol{a}$, the electric field at a point in space at a large distance $R(r)$ away in free space is, from Maxwell's equations,

$$\boldsymbol{E}(t, r) = -\mu_0 \left(\frac{1}{4\pi R}\right) q\, \boldsymbol{a}_{\mathrm{T}} \left(t - \frac{R}{c}\right) \tag{8.1.70}$$

where the subscript 'T' means the projection transverse to the direction of R and μ_0 is the permittivity in free space, $\mu_0 = 4\pi(10)^{-7}$ H m^{-1}. The time is expressed with

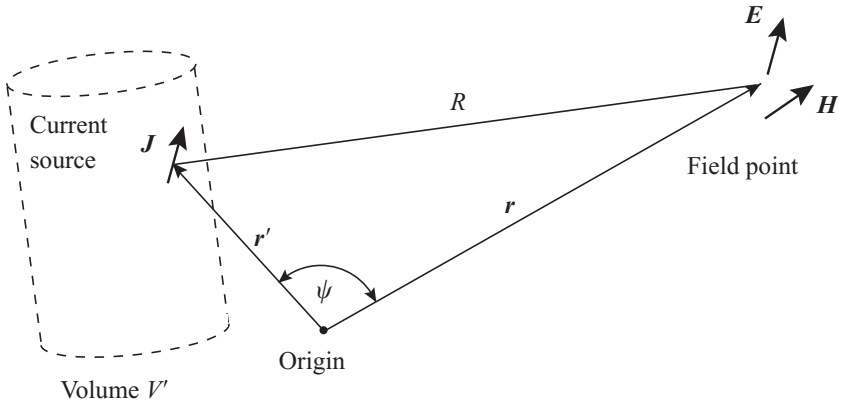


Figure 8.1.23 The coordinates for the field calculations from a current distribution inside the volume V'

a delay, R/c, according to the distance between the accelerating charge position and the observation point. This time delay is often dropped and the charge, or its equivalent current, becomes referred to as the *retarded* charge or current, and similarly for further derived quantities. An assumption has been made in this formula that the distance R is essentially constant, i.e. the extent of the accelerating charge is small relative to the distance R. This is the equivalent of the far-field approximation, given below.

For a sinusoidal time-variation of the current with angular frequency ω, the speed of the charge along a direction $\hat{\boldsymbol{w}}$ is $\boldsymbol{v}(t) = \mathrm{e}^{\mathrm{j}\omega t}\hat{\boldsymbol{w}}$ and the acceleration is $\boldsymbol{a}(t) = \mathrm{j}\omega\,\mathrm{e}^{\mathrm{j}\omega t}\hat{\boldsymbol{w}}$. The electric field is, now expressed as a function of position with the time dependence understood,

$$\boldsymbol{E}(\boldsymbol{r}) = -\mathrm{j}\omega\mu_0 \frac{\mathrm{e}^{-\mathrm{j}kR}}{4\pi R} q(r')\boldsymbol{v}_{\mathrm{T}}(r') \tag{8.1.71}$$

and the wavenumber, or propagation constant, is

$$k = k_0 = 2\pi/\lambda_0 = \omega/c = \omega\sqrt{\mu_0\epsilon_0} \tag{8.1.72}$$

and the permittivity of free space is $\epsilon_0 = 8.854(10)^{-12} \approx 1/(36\pi)(10)^{-9}$ farad m^{-1}.

Superimposing many elemental contributions,

$$\boldsymbol{E}(\boldsymbol{r}) = -\mathrm{j}\omega\mu_0 \sum_i \frac{\mathrm{e}^{-\mathrm{j}kR_i}}{4\pi R_i} q_i(r')\boldsymbol{v}_{\mathrm{T}i}(r'). \tag{8.1.73}$$

Now the $q\boldsymbol{v}_{\mathrm{T}}$ terms are replaced by current density $\boldsymbol{J}$ A m^{-2} in the volume integral form,

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{r}) &= -\mathrm{j}\omega\mu_0 \int_{V'} \frac{\mathrm{e}^{-\mathrm{j}kR}}{4\pi R}\boldsymbol{J}(r')\,\mathrm{d}V' \\ &= -\mathrm{j}\omega\mu_0 \int_{V'} G(R)\boldsymbol{J}(r')\,\mathrm{d}V',\end{aligned} \tag{8.1.74}$$

and the $G(R)$ notation is discussed below. The same integral relation holds for surface currents, in A m^{-1}, with a surface integral; and a line current, (in A), and line integral. These are convenient relations because they give the fields directly in terms of the currents in a volume in a relatively simple integral. However, they are for the far-field only.

8.1.9.2 Complete fields in terms of current distribution

The integral relations for the complete fields solutions are derived from Maxwell's equations with the potential functions (discussed below) and this is given in several texts, for example, Collin and Zucker (1969). Using $\mu_0\omega = k^2/(\omega\epsilon_0)$ with $\boldsymbol{E}(\boldsymbol{r}) = (j\omega\epsilon_0)^{-1}\nabla \times H$ (see also a summary of field quantities and relations, Section 8.1.13), the complete fields can be couched in terms of the current distribution:

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{r}) &= -\frac{1}{\mathrm{j}\omega\epsilon_0}\int_{V'} \nabla \times (\boldsymbol{J}(\boldsymbol{r}') \times \nabla G(R))\,\mathrm{d}V' \quad \left(\boldsymbol{r} \text{ not in } V'\right) \\ &= -\frac{1}{\mathrm{j}\omega\epsilon_0}\int_{V'} (k^2\boldsymbol{J}(\boldsymbol{r}')G(R) + \boldsymbol{J}(\boldsymbol{r}') \cdot \nabla^2 G(R))\,\mathrm{d}V',\end{aligned} \tag{8.1.75}$$

and

$$\boldsymbol{H}(\boldsymbol{r}) = -\int_{V'} (\boldsymbol{J}(\boldsymbol{r}') \times \nabla G(R))\, \mathrm{d}V', \quad \left(\boldsymbol{r} \text{ not in } V'\right); \tag{8.1.76}$$

where $\mathrm{d}V'$ is an elemental volume containing the current sources and the integral is over a volume V' containing the current sources. The differentiating operators (using ∇) are summarised at the end of this section. These integral relations are general, but can be difficult to solve directly and it is normal to use an extra step using the potential functions, given below.

The Green's function is

$$G(R) = \frac{\mathrm{e}^{-\mathrm{j}kR}}{4\pi R}. \tag{8.1.77}$$

The vector dependency of R is commonly written as $\boldsymbol{R}(\boldsymbol{r}|\boldsymbol{r}')$. It has been traditional to denote Green's functions as vectors using the unit dyadic $\boldsymbol{I}_{\mathrm{d}} = \hat{x}\hat{x} + \hat{y}\hat{y} + \hat{z}\hat{z}$, as $\boldsymbol{G}(\boldsymbol{r}|\boldsymbol{r}') = G(R)\boldsymbol{I}_{\mathrm{d}}$, and this allows dot product expressions with the current distribution, i.e. $\boldsymbol{G} \cdot \boldsymbol{J}$, for the integrands.

8.1.9.3 Far-field approximation

The far-field approximation for the Green's function is written

$$G(R) \approx \frac{\mathrm{e}^{-\mathrm{j}k\hat{r}\cdot r'}}{4\pi r}, \quad r \gg r'; \tag{8.1.78}$$

where $\hat{r}$ is the outward pointing unit vector, $r = |\boldsymbol{r}|$, $r' = |\boldsymbol{r}'|$. The approximations are

$$R = |\boldsymbol{r} - \boldsymbol{r}'| = \sqrt{r^2 + r'^2 - 2\boldsymbol{r}\cdot\boldsymbol{r}'} \approx r - \frac{\boldsymbol{r}\cdot\boldsymbol{r}'}{r} = r - \hat{\boldsymbol{r}}\cdot\boldsymbol{r}' = r - r'\cos\psi \tag{8.1.79}$$

in the exponent (the phase approximation) with ψ the angle between $\boldsymbol{r}'$ and $\boldsymbol{r}$; and $R \approx r$ (the amplitude approximation) in the denominator. So the far-field Green's function is

$$G(R) \approx \frac{\mathrm{e}^{-\mathrm{j}kr}}{4\pi r}\, \mathrm{e}^{\mathrm{j}kr'\cos\psi}, \quad r \gg r'; \tag{8.1.80}$$

and only the second term needs to remain in the integrals for the fields. The first term is the outward propagating spherical wave.

8.1.9.4 Rectangular, spherical and cylindrical source coordinates

For transforming the expression to correspond to general source coordinates, the relation

$$\boldsymbol{r}\cdot\boldsymbol{r}' = rr'\cos\psi = xx' + yy' + zz' \tag{8.1.81}$$

is used. The coordinate transforms are tabulated further below in Table 8.4.

In terms of the far-field spherical coordinates, the far-field approximation is expressed as a series,

$$\begin{aligned} R &= \sqrt{(r\sin\theta\cos\phi - x')^2 + (r\sin\theta\sin\phi - x')^2 + (r\cos\phi - z')^2} \\ &= r - (x'\sin\theta\cos\phi + y'\sin\theta\sin\phi + z'\cos\theta) + O(r^{-1}) \end{aligned} \tag{8.1.82}$$

and when this distance is in the denominator, the convergence is similarly expressed $R^{-1} = r^{-1}O(r^{-2})$. The far-field Green's function is now written

$$G(R) \approx \frac{\mathrm{e}^{-\mathrm{j}kr}}{4\pi r}\,\mathrm{e}^{\mathrm{j}k(x'\sin\theta\cos\phi + y'\sin\theta\sin\phi + z'\cos\theta)}, \quad r \gg r'; \tag{8.1.83}$$

and, as above, the first term comes out of the integrals of the field expressions. So for rectangular source coordinates,

$$r'\cos\psi = x'\sin\theta\cos\phi + y'\sin\theta\sin\phi + z'\cos\theta. \tag{8.1.84}$$

In cylindrical source coordinates, the expression required is

$$r'\cos\psi = \rho'\sin\theta\cos(\phi - \phi') + z'\cos\theta. \tag{8.1.85}$$

In spherical source coordinates,

$$r'\cos\psi = \rho'\left(\cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi - \phi')\right). \tag{8.1.86}$$

8.1.9.5 Practical distance to far field

For practical purposes, the far field region of a directive is usually taken as a distance $R = d$ where

$$d \geq 2\frac{D^2}{\lambda} \tag{8.1.87}$$

where D is the largest dimension of the antenna. This is based on the antenna occupying an aperture and the difference in the distance to the field point from the centre of the aperture given by D and the edge of the aperture being less than $\lambda/16$. So the criterion is somewhat arbitrary, depending on the accuracy of the phase of the (almost) plane wave.

For low directivity antennas, the criterion may take on a different basis. For an electrically small dipole, described in Section 8.2, for example, the far-field distance is often taken as the distance at which the radiating fields become larger than the reactive fields; see the *IEEE Standard Test Procedures* (1979). In the broadside direction, the reactive fields dominate within a distance $\lambda/(2\pi)$, but the radiative fields dominate outside this range. The far-field distance follows as $d \geq \lambda/(2\pi)$. Figure 8.1.24 indicates these far field relations.

Even in the far field, the integrals for the fields are still complicated and are not easily managed except for simple current distributions. For example, for current sources which do not have a changing charge distribution in the sense that the divergence of the current density is zero, $\nabla \cdot \boldsymbol{J} = 0$, called a solenoidal field, then the radiated far field reduces to the more manageable form (Collin and Zucker, 1969)

$$\boldsymbol{E}(\boldsymbol{r}) \approx -\mathrm{j}\omega\mu_0\frac{\mathrm{e}^{-\mathrm{j}kr}}{4\pi r}\int_V \mathrm{e}^{+\mathrm{j}k\hat{\boldsymbol{r}}\cdot\boldsymbol{r}'}\boldsymbol{J}\left(\boldsymbol{r}'\right)\mathrm{d}V' \quad \text{(far field)}. \tag{8.1.88}$$

Often, the propagation constant is stated as a vector, $\boldsymbol{k} = k\hat{\boldsymbol{r}}$, and so the above sign convention in the exponents corresponds to radiating antennas (propagation in the outward direction) for an outward-directed propagation vector. The above equation shows a Fourier relationship between the source distribution and the far-field pattern.

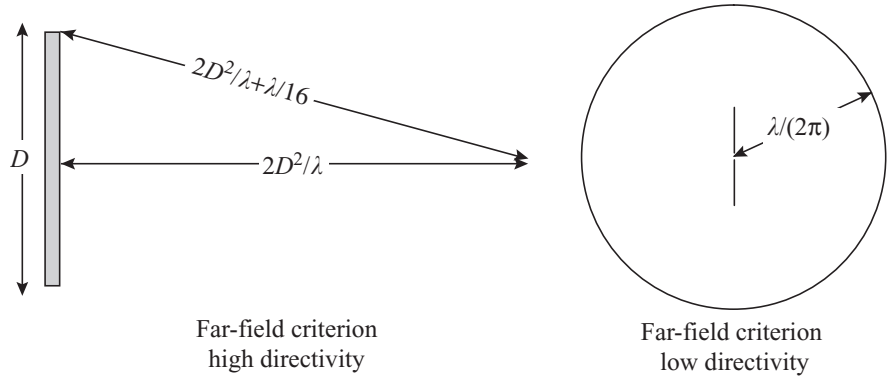


Figure 8.1.24 The practical far-field regions used for large and directive antennas and for electrically small, low-directivity antennas

8.1.9.6 Use of auxiliary potential functions

Because of the difficulties in solving the integrals above, an intermediate step is normally used for finding the radiated fields. This step is to calculate the *vector magnetic potential*, $\boldsymbol{A}$ (Wb m^{-1}), which is a mathematical function, and not a readily measured quantity. $\boldsymbol{A}$ is then used in turn to find the fields. It is defined from the magnetic field, and the electric field relation follows from $\nabla \times \boldsymbol{E} = -\partial \boldsymbol{B}/\partial t$:

$$\boldsymbol{H}(\boldsymbol{r}) = \frac{1}{\mu_0} \nabla \times \boldsymbol{A}(\boldsymbol{r}) \quad \text{(definition of } \boldsymbol{A}\text{)} \tag{8.1.89}$$

$$\boldsymbol{E}(\boldsymbol{r}) = -\mathrm{j}\omega \boldsymbol{A}(\boldsymbol{r}) - \mathrm{j}\frac{1}{\omega \mu_0 \epsilon_0} \nabla \left(\nabla \cdot \boldsymbol{A}(\boldsymbol{r})\right). \tag{8.1.90}$$

Only the first term of the electric field equation is required for the far fields. An alternative definition of $\boldsymbol{A}$, found in several texts, omits the permeability to give the vector potential of $\boldsymbol{H}$ rather than of $\boldsymbol{B}$, namely $\boldsymbol{H} = \nabla \times \boldsymbol{A}$. This form is not used here.

This extra step of introducing the vector potential, although replacing an integration with an integration plus a differentiation, is often simpler because the resulting integration of the vector potential is often easier than the integration of the direct calculation. The differentiation is not normally a difficult task.

The integral relation for the vector potential is similar to the simplified example for the electric field above and is here expressed using Cartesian parameters,

$$\boldsymbol{A}(x, y, z) = \mu_0 \iiint_{V'} \boldsymbol{J}\left(x', y', z'\right) G(R)\, \mathrm{d}x'\, \mathrm{d}y'\, \mathrm{d}z' \tag{8.1.91}$$

where $\mathrm{d}V' = \mathrm{d}x'\, \mathrm{d}y'\, \mathrm{d}z'$.

Simple current distributions give rise to simpler forms of the integral. For a thin wire, the azimuthal component of current is neglected, and the current distribution can be taken as purely in the axial direction on the wire, $\boldsymbol{I} = I\hat{L}$ (in A). If the wire

follows a contour L, the vector potential is

$$\boldsymbol{A}(x, y, z) = \mu_0 \int_L \boldsymbol{I}\left(x', y', z'\right) G(R)\mathrm{d}L, \tag{8.1.92}$$

where $\mathrm{d}L$ is the elemental length along the wire.

8.1.9.7 Fields for electric dipole

The uniform current on a wire can be approximated using top-hat loaded wire ends and short wire lengths.

8.1.9.8 'Infinitesimal', uniform dipole

For a uniform, straight current, here in the z-direction, i.e. $\boldsymbol{I}(x', y', z') = \hat{z}I_0$, and sufficiently short wire length L defined by $z' = \pm L/2$,

$$\boldsymbol{A} = \hat{z}A_z = \hat{z}\mu_0 I_0 \frac{\mathrm{e}^{-\mathrm{j}kr}}{4\pi r} L \tag{8.1.93}$$

in which the length is so small that $R \approx r$. In terms of spherical components, this is

$$\begin{aligned} A_r &= A_z \cos\theta, \\ A_\theta &= -A_z \sin\theta, \\ A_\phi &= 0. \end{aligned} \tag{8.1.94}$$

There is no azimuthal variation since the currents are purely axial on a straight wire.

The magnetic field reduces to

$$\boldsymbol{H} = \hat{\phi}\frac{1}{\mu_0 R}\left(\frac{\partial}{\partial r} r A_\theta - \frac{\partial}{\partial \theta} A_r\right), \tag{8.1.95}$$

and the electric field can be found from $\boldsymbol{A}$ as given above, or from Maxwell's equation

$$\boldsymbol{E} = \frac{1}{\mathrm{j}\omega\epsilon_0}(\nabla \times \boldsymbol{H} - \boldsymbol{J}). \tag{8.1.96}$$

The dipole fields for the uniform current distribution, here in spherical coordinates, follow as

$$\begin{aligned} E_r^{(\mathrm{u})} &= Z_0 \frac{I_0 L \cos\theta}{2\pi r^2}\left(1 + \frac{1}{\mathrm{j}kr}\right) \mathrm{e}^{-\mathrm{j}kr} \\ E_\theta^{(\mathrm{u})} &= \mathrm{j}Z_0 \frac{kI_0 L \sin\theta}{4\pi r}\left(1 + \frac{1}{\mathrm{j}kr} - \frac{1}{(kr)^2}\right) \mathrm{e}^{-\mathrm{j}kr} \\ E_\phi^{(\mathrm{u})} &= 0, \end{aligned} \tag{8.1.97}$$

and Z_0 is the free-space impedance given by the ratio of the propagating free-space electric field to the magnetic field, $Z_0 = E_\theta / H_\phi = \sqrt{\mu_0/\epsilon_0} = 120\pi\ \Omega$ (per square), as noted above.

In the far field, the terms that decrease with distance drop out, and the radiated fields reduce to

$$\begin{aligned} E_\theta^{(\mathrm{u})} &= Z_0 K_\mathrm{u} \sin\theta \\ H_\phi^{(\mathrm{u})} &= \frac{1}{Z_0} E_\theta^{(\mathrm{u})} \end{aligned} \tag{8.1.98}$$

where

$$K_\mathrm{u} = \mathrm{j}kI_0 L \frac{\mathrm{e}^{-\mathrm{j}kr}}{4\pi r}. \tag{8.1.99}$$

The radiation pattern is simply $|\sin\theta|$, and this is the effective length of the antenna, $\boldsymbol{h} = \sin\theta\hat{\theta}$, i.e. the effective length relation $V_0 = \boldsymbol{h} \cdot \boldsymbol{E}_\mathrm{inc}$ holds and equation (8.1.98) holds with $\boldsymbol{g} = \boldsymbol{h}$. This gives the directivity, defined along the horizontal plane, as 3/2, or almost 1.8 dB. If the dipole is considered as a monopole on a groundplane, the fields, considered as confined to the upper hemisphere, are otherwise the same, and so the directivity becomes almost 4.8 dB. The directivity is calculated using the formulas from the start of the section.

8.1.9.9 Sinusoidal dipole current

For the longer dipole, the more general sinusoidal current distribution is required, and this is very accurate for most purposes. The radiated fields are found from

$$\boldsymbol{A} = \hat{z} A_z = \hat{z} \mu_0 I_\mathrm{max} \frac{\mathrm{e}^{-\mathrm{j}kr}}{4\pi r} \int_{-L/2}^{L/2} \sin\left(k(L - |z|)\right) \mathrm{e}^{\mathrm{j}kr' \cos\theta} \mathrm{d}z \tag{8.1.100}$$

to be

$$\begin{aligned} E_\theta^{(\mathrm{s})} &= Z_0 K_\mathrm{s} \left(\frac{\cos\left(kL/2(\cos\theta)\right) - \cos kL/2}{\sin\theta} \right) \\ H_\phi^{(\mathrm{s})} &= \frac{1}{Z_0} E_\theta^{(\mathrm{s})} \end{aligned} \tag{8.1.101}$$

where

$$K_\mathrm{s} = \frac{\mathrm{j}I_\mathrm{max}}{2\pi} \frac{\mathrm{e}^{-\mathrm{j}kr}}{r}. \tag{8.1.102}$$

The special case of interest is for the half-wave dipole, $L/2 = \lambda/4$, which has a radiation power pattern of $|g(\theta)|^2 = |\cos\pi/2(\cos\theta)/\sin\theta|^2 \approx |\sin\theta|^3$, giving directivity of about 2.1 dB, or for the monopole, about 5.1 dB.

For dipoles longer than a wavelength, multi-lobed patterns arise as the structure of the current distribution takes on the repetitive characteristic of an array. Examples of the power patterns from sinusoidal current distributions are shown in Figure 8.1.25.

For the small dipole wire, a linearly tapered current distribution is a simple approximation to the sinusoidal distribution (small angle formula). The vector potential and the field quantities are exactly half of those from the infinitesimal uniform current distribution. So the pattern, and therefore the directivity, are the same as for the uniform current case.

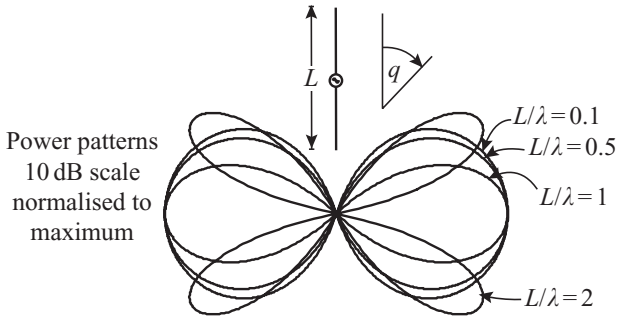


Figure 8.1.25 Power patterns of the dipole antenna for different lengths. The power patterns are normalised to have the same maximum and the full scale is 10 dB. For lengths shorter than a half wavelength, the pattern changes very little, and for lengths greater than a wavelength, multiple lobes appear

8.1.9.10 Uniform distribution dipole source

A long, uniform current distribution is difficult to realise on a wire in free space. However, the result is useful for applying to the electromagnetic duals, i.e. the slots of patch antennas, discussed below and in Section 8.2. The vector potential for a uniform current source is just

$$\boldsymbol{A} = \hat{z}A_z = \hat{z}\mu_0 I_0 \frac{\mathrm{e}^{-\mathrm{j}kr}}{4\pi r} \int_{-L/2}^{L/2} \mathrm{e}^{\mathrm{j}kr' \cos\theta} \mathrm{d}z. \tag{8.1.103}$$

This produces the far-field pattern $g(\theta) = (\sin \nu)/\nu$ with $\nu = kL/2(\cos\theta)$, a result also readily obtained from the Fourier transform relationship between the current distribution and the far-field pattern. The directivity for this pattern is plotted in Figure 8.1.26. The result is also applied in Section 8.2 for patch antennas. As the length of the uniform source increases to over a couple of wavelengths, the directivity becomes proportional to twice the line length in wavelengths.

This is fundamentally different from the sinusoidal current distribution, whose pattern breaks up with increasing length, as shown above, and the maximum directivity and its direction changes with length (for lengths over a wavelength) in a complicated way. Such long dipoles or monopoles are not used often.

8.1.9.11 Magnetic sources

If the sources are magnetic, then the process for finding the fields is followed using almost identical relationships through the principle of duality. The magnetic sources are denoted $\boldsymbol{M}$ and these are used to calculate the electric vector potential $\boldsymbol{F}$.

Magnetic sources are typically used to model the electric fields across slots in conductors, and in particular to model patch antennas where the spacings between the edges of the patch and the ground plane are taken as slots. The slots are modelled as a magnetic current, as shown in Figure 8.1.27. The electric fields across the slot result from a voltage excitation across the gap of the conductors and these fields

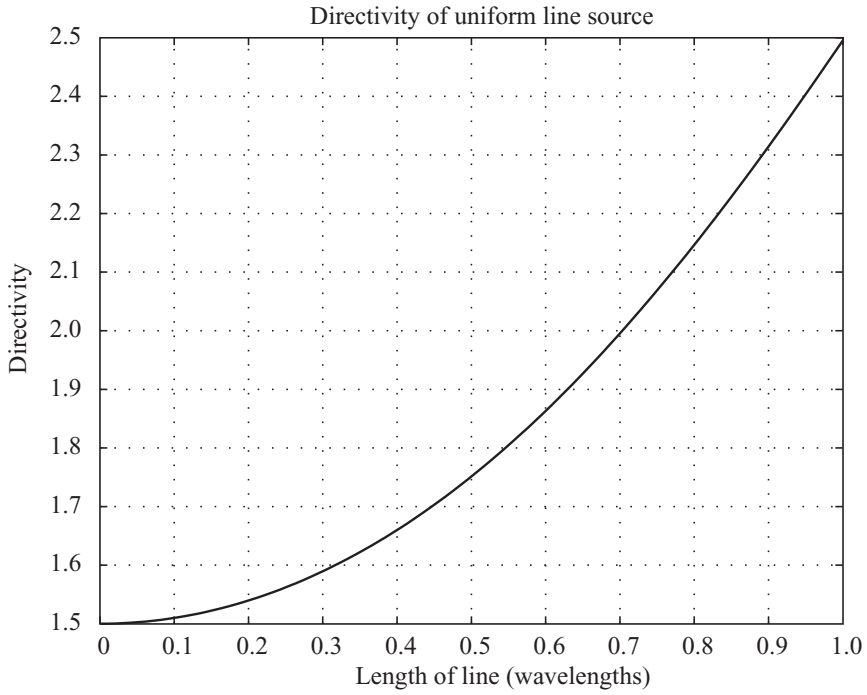


Figure 8.1.26 The directivity of a uniform line source increases with the length of the source

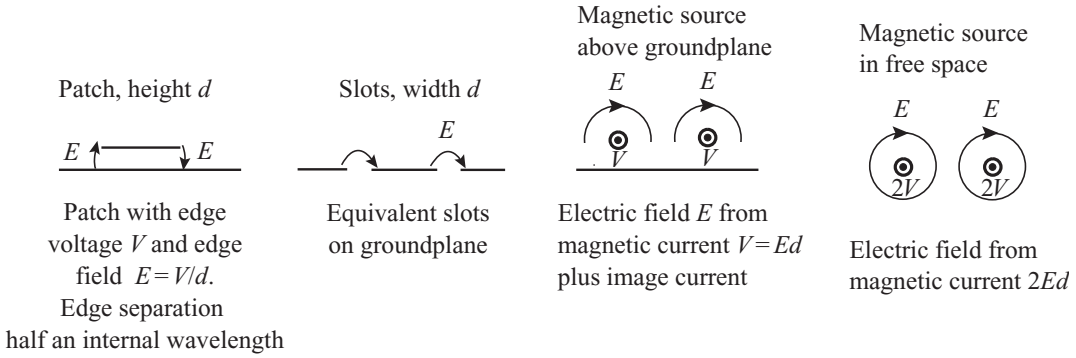


Figure 8.1.27 The modelling of a patch antenna as slots and with a magnetic current

can be considered as resulting from a filament of magnetic current. The magnetic source, i.e. the magnetic current, can then be used to calculate the fields. Often, established electric wire fields can be used when the slot structure can correspond to wire structures. A basic thin slot can correspond to a thin dipole, in a hemispheric space, for example. In particular, circular or square patch antennas in a hemisphere correspond to wire loops.

In such Maxwell duals, the field quantities are interchanged with a sign change: for magnetic sources, fields from electric wires (traditionally known results) can be used with the substitutions: $\boldsymbol{H}_F$ becomes $\boldsymbol{E}_A$ and $\boldsymbol{E}_F$ becomes $-\boldsymbol{H}_A$ (see Tables 8.2

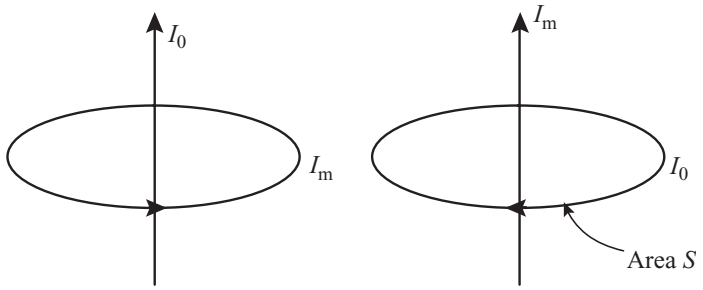


Figure 8.1.28 The small magnetic dipole with current I_m (in V) and a small loop of electric current I_0 (in A) are equivalent sources

and 8.3). For example, an electrically small loop antenna on the x-y plane, has fields given by the electric current dipole, but with the substitutions made. These fields can also be calculated from the electric currents on the conducting surfaces of the patch and the groundplane, but this process would be arduous.

The magnetic dipole moment $I_m L$ (L is the length of the 'infinitesimal' magnetic dipole) is related to a small electric loop with constant current I_0 and cross section S, as

$$I_m L = \mathrm{j} S \omega \mu_0 I_0. \tag{8.1.104}$$

The shape of the cross section is not important, but for the current distribution to be uniform, the wire diameter should be constant (see Figure 8.1.28).

8.1.9.12 Slot feeding

Slot antennas can be placed in waveguide walls and the waveguide fed to produce an array antenna. Single slots in a conductor sheet can also be fed as discrete elements. A narrow slot with a length of a half wavelength is the dual of a half-wavelength dipole, for example. The slot can be fed directly by coax at a position along the slot for the best match. The impedance seen by the coaxial feed is zero at the ends of the slot, where the electric field across the slot is short-circuited. For a half-wavelength long slot (or smaller), the impedance is a maximum at the centre of the slot where the electric field, and therefore the voltage, across the slot is a maximum. The distribution of the voltage across the slot is approximately sinusoidal along the length of the slot (cf. the current on a wire dipole) and so the coax can be placed appropriately for a match on a resonant slot: see Figure 8.1.29.

8.1.9.13 Booker's relation

The field equivalence described above holds everywhere, including the near fields. This means that there is a special relationship between the impedances of the dual antennas, called Booker's relation (Booker, 1946), which is an extension of Babinet's principle. This is usually written in terms of the admittance of the slot, Y_{slot}, and the impedance of the corresponding complementary conducting structure, Z_C, at the

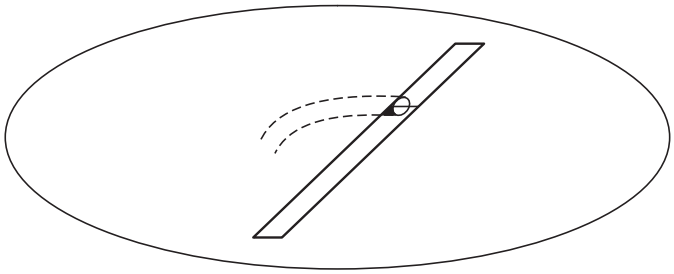


Figure 8.1.29 *Feeding a slot-on-groundplane antenna with a coaxial feed. For a narrow slot, the electric field across the slot can be replaced by a magnetic current. The electric field at the ends of the slot are zero, short-circuited by the conductor. A cavity is normally placed under the slot to confine the radiation to the upper hemisphere*

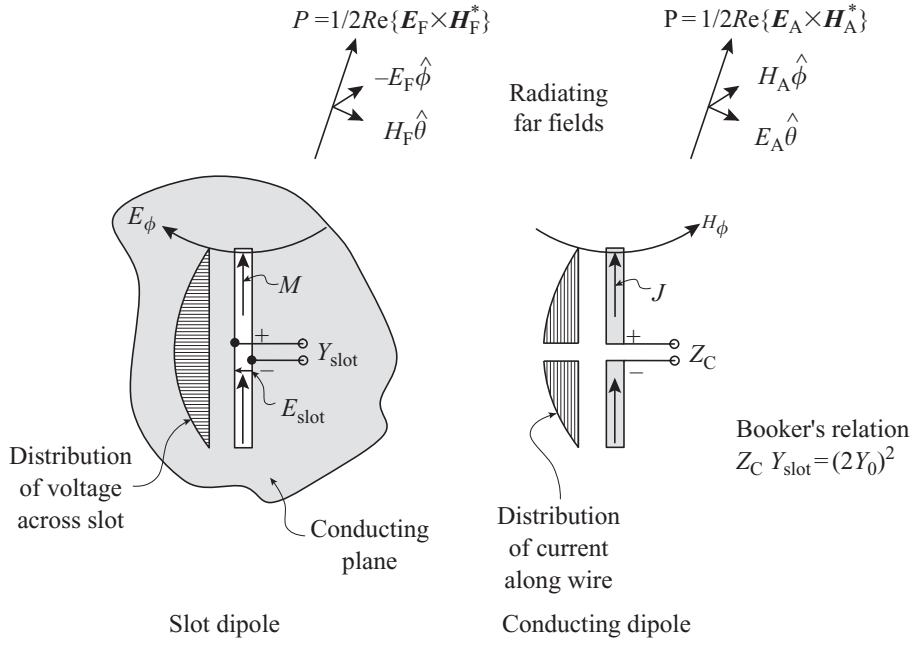


Figure 8.1.30 *The complementary pair of a slot dipole and conductor dipole. If the slot radiates both sides of the conductor, then Booker's relation holds*

corresponding feedpoint:

$$\frac{Y_{\text{slot}}}{Z_C} = \left(\frac{2}{Z_0}\right)^2. \tag{8.1.105}$$

For the thin, half-wavelength wire dipole, for example, the resonant input resistance at the centre of the feed is 73 Ω, so the resistance of the equivalent slot, fed at the centre point and allowed to radiate both sides of the slot, is about 487 Ω. The equivalent electric and magnetic dipoles are shown in Figure 8.1.30.

Booker's relation also holds for self- and mutual-impedance of complementary antenna elements in an array, and this can be expressed as $[\boldsymbol{Y}_{\text{slot}}] = 4Y_0^2[\boldsymbol{Z}_{\text{C}}]$, where [.] is the square matrix of mutual impedances or admittances. Booker's relation represents an important principle for antenna design, allowing both element and array design of slot antennas on a groundplane, to be guided by the impedances of the complementary conducting (for example, wire) structures. It can be applied to an antenna radiating into one side of the conducting plane as follows.

8.1.9.14 Cavity-backed slot

A slot antenna does not necessarily radiate to both sides of its conducting sheet. Often, the slot is closed off behind the conductor with a cavity. Impedance calculations for the slot need to account for the cavity effect (Andersen: Chapter 7 in Rudge *et al*., 1982).

The radiation is for a half-space (hemisphere) and so the radiation conductance can be determined from the external admittance, $Y_{\text{e}} = Y_{\text{slot}}/2$. This is in parallel with the internal admittance, i.e. that looking into the cavity, Y_{i}, so the overall slot admittance is $Y = Y_{\text{e}} + Y_{\text{i}}$. If the cavity is lossless then the internal impedance will be purely imaginary, $Y_{\text{i}} = \text{j}(B_{\text{slot}} + B_{\text{cavity}})$. Here, for a large enough cavity relative to the slot size, $B_{\text{slot}} \approx B_{\text{e}}$ (since this value is defined from the fringing, near fields of the slot), and B_{cavity} depends on the cavity dimensions. A simple model for the cavity is a waveguide with a short circuit at the end opposite the slot. Some experimental results for the admittances are given in Long (1975) and Cockrell (1976). The equivalent circuit is given in Figure 8.1.31, where the external admittance is in parallel with the internal admittance with the slot voltage across the slot separating them.

8.1.9.15 Self-complementary antennas

A self-complementary antenna (Rumsey, 1966) is one which, under rotation or translation, forms its own Maxwell dual; see Figure 8.1.32. Alternatively stated, if the metallic parts and dieletric parts are interchanged, then the antenna looks the same. The usual example is a spiral-type antenna, such as a planar spiral or a helix. Booker's

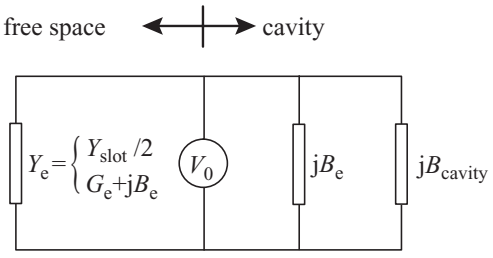


Figure 8.1.31 Equivalent circuit of admittances for a cavity-backed slot antenna. The slot voltage separates an external admittance representing the free space side of the conductor, and an internal susceptance represents the lossless cavity. The internal susceptance comprises a fringing term which is taken as the same as the external fringing value, and a term B_{cavity} *governed by the cavity details*

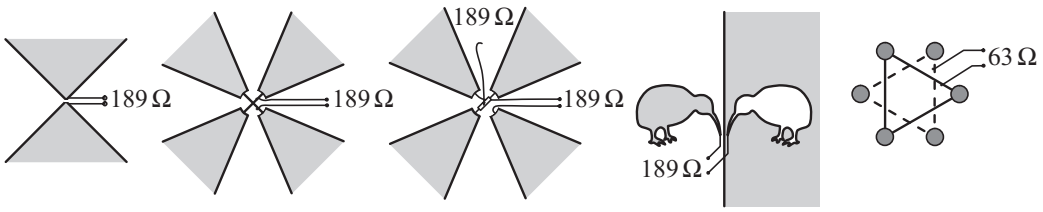


Figure 8.1.32 *The self-complementary antenna can take an infinite variety of forms. The single port has an input impedance of* 189 Ω *for all frequencies. The two-port version has* 189 Ω *at one port if the other port is terminated in* 189 Ω. *The diagram on the right depicts the six terminals of a self-complementary structure with alternate terminals interconnected to give an input impedance of* $120\pi/6 = 63\,\Omega$

relation in this case, after Y. Mushiake, and G. Deschamps, provides an important property of the single port (two-terminal) self-complementary antenna: the input impedance is $Z_0/2 = 189\,\Omega$. This holds for all frequencies if the structures are infinite.

For the N-terminal case, with alternate terminals connected, the input impedance is $120\pi/N$. Such an eight-terminal structure has an input impedance of 47 Ω, for example, an almost perfect match to 50 Ω. For a two-port antenna, with one port terminated in $Z_0/2$, the impedance at the other port is also $Z_0/2$. These results are also very helpful for checking calculation techniques and computer code, such as for the moment method.

8.1.9.16 Combination of electric and magnetic sources for field calculations

For a combination of magnetic and electric sources, the relations for finding the fields are summarised as follows:

$$\boldsymbol{A} = \mu_0 \int_{V'} \boldsymbol{J}\, G(R)\, \mathrm{d}V' \tag{8.1.106}$$

$$\boldsymbol{F} = \epsilon_0 \int_{V'} \boldsymbol{M}\, G(R)\, \mathrm{d}V' \tag{8.1.107}$$

$$\begin{aligned} \boldsymbol{E} = \boldsymbol{E}_A + \boldsymbol{E}_F &= -\mathrm{j}\left(\omega\, \boldsymbol{A} + \frac{1}{\omega\, \mu_0 \epsilon_0} \nabla\, (\nabla \cdot \boldsymbol{A})\right) - \frac{1}{\epsilon_0} \nabla \times \boldsymbol{F} \\ &= \frac{1}{\mathrm{j}\omega\, \epsilon_0} \nabla \times \boldsymbol{H}_A - \frac{1}{\epsilon_0} \nabla \times \boldsymbol{F} \end{aligned} \tag{8.1.108}$$

$$\begin{aligned} \boldsymbol{H} = \boldsymbol{H}_A + \boldsymbol{H}_F &= \frac{1}{\mu_0} \nabla \times \boldsymbol{A} - \mathrm{j}\left(\omega\, \boldsymbol{F} + \frac{1}{\omega\, \mu_0 \epsilon_0} \nabla\, (\nabla \cdot \boldsymbol{F})\right) \\ &= \frac{1}{\mu_0} \nabla \times \boldsymbol{A} - \frac{1}{\mathrm{j}\omega\, \mu_0} \nabla \times \boldsymbol{E}_F. \end{aligned} \tag{8.1.109}$$

Table 8.2 Summary of dual quantities for electric and magnetic sources

Quantities for electric source $\boldsymbol{J}$	Quantities for magnetic source $\boldsymbol{M}$
$\boldsymbol{E}_A$	$\boldsymbol{H}_F$
$\boldsymbol{H}_A$	$-\boldsymbol{E}_F$
$\boldsymbol{J}$	$\boldsymbol{M}$
$\boldsymbol{A}$	$\boldsymbol{F}$
ϵ	μ
μ_0	ϵ_0
Z_0	$1/Z_0$
$1/Z_0$	Z_0
k	k

Table 8.3 Summary of dual quantities for electric and magnetic sources. R is given in equation (8.1.79) et al.

Fields from electric sources	Fields from magnetic sources
$\nabla \times \boldsymbol{E}_A = -\mathrm{j}\omega\,\mu_0 \boldsymbol{H}_A$	$\nabla \times \boldsymbol{H}_F = \mathrm{j}\omega\,\mu_0 \boldsymbol{E}_F$
$\nabla \times \boldsymbol{H}_A = \mathrm{j}\omega\,\epsilon_0 \boldsymbol{E}_A + \boldsymbol{J}$	$-\nabla \times \boldsymbol{E}_F = \mathrm{j}\omega\,\mu_0 \boldsymbol{H}_F + \boldsymbol{M}$
$\nabla^2 \boldsymbol{A} + k^2 \boldsymbol{A} = -\mu_0 \boldsymbol{J}$	$\nabla^2 \boldsymbol{F} + k^2 \boldsymbol{F} = -\epsilon_0 \boldsymbol{M}$
$\boldsymbol{A} = \dfrac{\mu_0}{4\pi} \int\int_{V'} \int \boldsymbol{J} \dfrac{\mathrm{e}^{-\mathrm{j}kR}}{\boldsymbol{R}}\, \mathrm{d}V'$	$\boldsymbol{F} = \dfrac{\epsilon_0}{4\pi} \int\int_{V'} \int \boldsymbol{M}\, \dfrac{\mathrm{e}^{-\mathrm{j}kR}}{R}\, \mathrm{d}V'$
$\boldsymbol{H}_A = \dfrac{1}{\mu_0} \nabla \times \boldsymbol{A}$ (definition of $\boldsymbol{A}$)	$\boldsymbol{E}_F = -\dfrac{1}{\epsilon_0} \nabla \times \boldsymbol{F}$ (definition of $\boldsymbol{F}$)
$\boldsymbol{E}_A = -\mathrm{j}\omega\,\boldsymbol{A} - \mathrm{j}\dfrac{1}{\omega\,\mu_0 \epsilon_0} \nabla\,(\nabla \cdot \boldsymbol{A})$	$\boldsymbol{H}_F = -\mathrm{j}\omega\,\boldsymbol{F} - \mathrm{j}\dfrac{1}{\omega\,\mu_0 \epsilon_0} \nabla\,(\nabla \cdot \boldsymbol{F})$

The dual quantities for electric and magnetic sources are well known and are given in most antenna texts and in Table 8.2, after Harrington (1961), and the dual equations are given in Table 8.3, after Balanis (1982).

Finally, the transformations between coordinate systems are required for most antenna types. These are readily available, and compactly stated here in Table 8.4.

These can be used to transform the unit vectors either way; for example, the Cartesian–spherical are the most used, and are:

$$\begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\phi & \cos\theta\cos\phi & -\sin\phi \\ \sin\theta\sin\phi & \cos\theta\sin\phi & \cos\phi \\ \cos\theta & -\sin\theta & 0 \end{bmatrix} \begin{bmatrix} \hat{r} \\ \hat{\theta} \\ \hat{\phi} \end{bmatrix}, \tag{8.1.110}$$

$$\begin{bmatrix} \hat{r} \\ \hat{\theta} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ -\sin\phi & \cos\phi & 0 \end{bmatrix} \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{z} \end{bmatrix}. \tag{8.1.111}$$

Table 8.4 Coordinate transformations

	rectangular–spherical			rectangular–cylindrical			cylindrical–spherical			
	$\hat{r}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\rho}$	$\hat{\phi}$	$\hat{z}$		$\hat{\rho}$	$\hat{\phi}$	$\hat{z}$
$\hat{x}$	$\sin\theta\cos\phi$	$\cos\theta\cos\phi$	$-\sin\phi$	$\cos\phi$	$-\sin\phi$	0	$\hat{r}$	$\sin\theta$	0	$\cos\theta$
$\hat{y}$	$\sin\theta\sin\phi$	$\cos\theta\sin\phi$	$\cos\phi$	$\sin\phi$	$\cos\phi$	0	$\hat{\theta}$	$\cos\theta$	0	$-\sin\theta$
$\hat{z}$	$\cos\theta$	$-\sin\theta$	0	0	0	1	$\hat{\phi}$	0	1	0

8.1.10 Radiation and ohmic resistance for dipoles

In this section, some of the theories for calculating the impedance of a straight wire antenna are reviewed. In practice these theories seem to work for short (up to a quarter-wavelength long) sections, but can be unreliable for calculating the feed-point impedance of longer wire structures. Nevertheless the theories are important because they deal with the physical principles of currents and fields.

8.1.10.1 Method of moments

The method of moments is available as popular commercial software in several versions, and can offer insightful visualisation, but the user is seldom exposed to the basic principles at work. An example of moment method results is given in Section 8.4 for a monopole on a groundplane. For very thin antennas, the moment method can give reliable answers for the impedance. However, the wires often need to be impracticably thin in order to get such reliable answers.

For current excitation, the radiation resistance should be accurate, and for voltage excitation, the conductance should be accurate. (Note that one cannot be calculated from the other, unless the imaginary parts are zero.) Calculated impedances can be very dependent on the number of segments used, and the user should experiment with this, as well as with the wire thickness. Again, for short wires, the answers seem to hold well, but for longer wires which are thick enough to be mechanically robust, the calculations should be experimented with via the number of segments and wire thickness. The problem stems from the approximations in the excitation, and the fact that the impedance calculation relies on the current in the segment at which the impedance is being assessed. An inaccuracy in the current segment, particularly for points away from the current node, means an inaccurate impedance calculation. However, the inaccuracies tend to become smoothed out when integrating the calculated currents to get the far-field patterns. This means that the patterns are normally accurate even for coarse segmentation of the wire structure.

8.1.10.2 Integration of far fields

As described above, the radiated power is given by integrating the radiating far fields, and this power is equated to the power supplied to the antenna, $P_{\text{rad}} = 1/2|I_0|^2 R_{\text{rad}}$; see Figure 8.1.8. For the uniform current distribution on a dipole, this gives a radiation

resistance of

$$R_{\text{rad}}^{(\text{uniform})} = 80\pi^2 \left(\frac{L}{\lambda}\right)^2. \tag{8.1.112}$$

This is useful for very short top-hat loaded dipoles.

For the linearly tapered current distribution, the calculation yields

$$R_{\text{rad}}^{(\text{taper})} = 20\pi^2 \left(\frac{L}{\lambda}\right)^2 = \frac{R_{\text{rad}}^{(\text{uniform})}}{4}. \tag{8.1.113}$$

In both cases, the radiation resistance is proportional to the dipole's electrical length squared, so for compact antennas, the losses may turn out to be significant relative to this, depending on the frequency and the surface resistance of the conductor (see below). As an example, for a dipole with $L/\lambda = 1/50$, the radiation resistance is only $R_{\text{rad}} \approx 0.08\,\Omega$ by using the tapered current formula, and about $0.32\,\Omega$ by using the uniform current formula.

For the sinusoidal current distribution, the equation becomes complicated, and is given here related to the current maximum (cf. Collin and Zucker, 1969; Balanis, 1982):

$$R_{\text{rad}}^{(\text{sin})} = \frac{\eta_0}{2\pi}\left(\gamma + \ln(kL) - \text{Ci}(kL) + \frac{1}{2}\sin(kL)S_1 + \frac{1}{2}\cos(kL)C_1\right) \tag{8.1.114}$$

where

$$S_1 = \text{Si}(2kL) - 2\text{Si}(kL), \quad C_1 = \gamma + \ln(kL/2) + \text{Ci}(2kL) - 2\text{Ci}(kL); \tag{8.1.115}$$

$\gamma = 0.5772$ is Euler's constant, and

$$\text{Ci}(x) = -\int_x^\infty \frac{\cos y}{y}\,\text{d}y = -\frac{1}{2}(E_1(\text{j}x) + E_1(-\text{j}x)), \tag{8.1.116}$$

$$\text{Si}(x) = \int_0^x \frac{\sin y}{y}\,\text{d}y = \frac{1}{2\text{j}}(E_1(\text{j}x) - E_1(-\text{j}x)) + \frac{\pi}{2}, \tag{8.1.117}$$

in which the exponential integral, a readily available function in software such as Matlab™, is

$$E_1(x) = \int_x^\infty \frac{\text{e}^{-t}}{t}\,\text{d}t. \tag{8.1.118}$$

This form of the radiation resistance, which is referred to a current maximum of the assumed sinusoidal distribution, rather than the feed point current, is well behaved. Over the lengths of most interest (usually this is less than a half-wavelength), this radiation resistance increases as the dipole length increases. It is shown in Figure 8.1.33 referred to the current maximum (from the equation above) and also

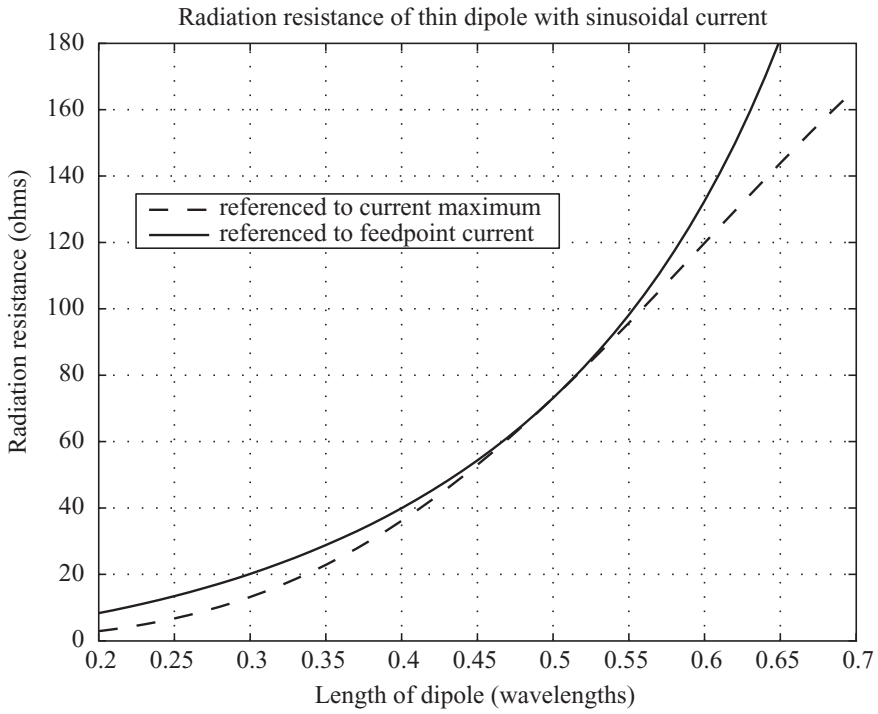


Figure 8.1.33 *The radiation resistance of a thin wire dipole in free space, which has a sinusoidal current distribution. The cases are: referred to the current maximum; and referred to the feedpoint current. For a lossless dipole, the latter case corresponds to the real part of the input impedance. The resistance is independent of the wire thickness*

referred to the feedpoint current, found by scaling by the current for fixed power, $I_{\rm IN}^2 R_{\rm IN} = I_{\rm m}^2 R_{\rm rad}$, i.e.

$$R_{\rm IN} = \left(\frac{I_{\rm m}}{I_{\rm IN}}\right)^2 R_{\rm rad}^{(\sin)} = \frac{R_{\rm rad}^{(\sin)}}{\sin^2 kL}. \tag{8.1.119}$$

Using this current approximation, the resulting expression for the radiation resistance, referred to the feedpoint, blows up at lengths which are multiples of λ. For these lengths, a more accurate expression for the current, which does not go to zero, is required (see below). In Figure 8.1.33, the two curves coincide for a half-wavelength dipole since feed current is the same as the current maximum, $\sin^2(k\lambda/4) = 1$.

For the special case of $L/\lambda = 0.5$, $R_{\rm rad} = \eta_0/4\pi\ (\ln 2\pi\gamma - Ci(2\pi)) = 73\,\Omega$. Referred to the current maximum, it reaches a local minimum of about $210\,\Omega$ at a length of about 0.8λ and then decreases to about $90\,\Omega$ at about 1.7λ, and continues to oscillate. Referred to the feedpoint, the local minima are at $(1.42\lambda, 92\,\Omega)$ and $(2.44\lambda, 111\,\Omega)$. The initial increase with length occurs as the doughnut shaped pattern sharpens. The directivity reaches a maximum of just over 5 dB at $L/\lambda = 1.25$. For longer lengths, the directivity has a complicated dependence on length as multiple lobes appear in the pattern.

8.1.10.3 Induced emf method for dipole impedance

The induced emf method (see, for example, Elliot, 1981), attributed originally to L. Brillouin in 1922, gives both real and imaginary parts of the input impedance for the wire dipole. For a wire radius of a, the tangential electric field on the vertical wire, $E_z(a, z')$, is calculated from the current distribution, which is assumed to be sinusoidal. The current itself is confined to the skin depth, which is taken here as infinitesimally small for the current density, so that $I_z = 2\pi a J_z = I_{\max} \sin\left(k(L/2 - |z'|)\right)$.

The formulation is an integral of the voltage contributions along the wire divided by the input current, to give the impedance, referred to the current maximum $I_{\max}$, as

$$\begin{aligned} Z_{\mathrm{m}} &= \frac{V}{I_{\max}} = \frac{1}{I_{\max}} \left(-\frac{1}{I_{\max}} \int_{-L/2}^{L/2} E_z(a, z')\ I_z(a, z')\ \mathrm{d}z' \right) \\ &= -\frac{1}{I_{\max}} \int_{-L/2}^{L/2} E_z(a, z')\ \sin(k(L/2 - |z'|))\ \mathrm{d}z' \\ &= R_{\mathrm{m}} + \mathrm{j}X_{\mathrm{m}}, \end{aligned} \tag{8.1.120}$$

where the input resistance is the same as that above, $R_{\mathrm{m}} = R_{\mathrm{rad}}^{(\sin)}$. The reactance, referred to the feedpoint, is

$$X_{\mathrm{IN}} = \frac{1}{\sin^2 kL} \frac{\eta_0}{4\pi} \left(2\mathrm{Si}(kL) - \cos(kL) S_1 - \sin(kL) C_2\right) \tag{8.1.121}$$

where

$$C_2 = 2\mathrm{Ci}(kL) - \mathrm{Ci}(2kL) - \mathrm{Ci}\left(\frac{2ka^2}{L}\right). \tag{8.1.122}$$

Only the last term in X_{IN} contains the dependence on the wire thickness, and this contribution drops out for the length being a multiple of a half-wavelength. So the emf formula gives the same reactance for half-wavelength antennas independent of the wire radius. The reactance is periodic with the length of the antenna. It is plotted in Figure 8.1.34 and the curves indicate the bandwidth increase (because of the decreased reactance change with frequency) and a resonant frequency decrease, with increasing wire thickness.

8.1.10.4 Wave structure approach

A general treatment for the impedance of monopoles and dipoles is given by Andersen (1968, 1971), based on a space-surface wave structure approach, and this approach also is in Jones (1979). This includes the case of lengths being a multiple of a wavelength and both the resistance and the reactance depend on the thickness. The approach gives both the radiation resistance (input resistance for a lossless structure) and the input reactance as being dependent on the wire thickness. The treatment is not reproduced here, but the basic result for the wire antenna is presented. The antenna is depicted in Figure 8.1.35.

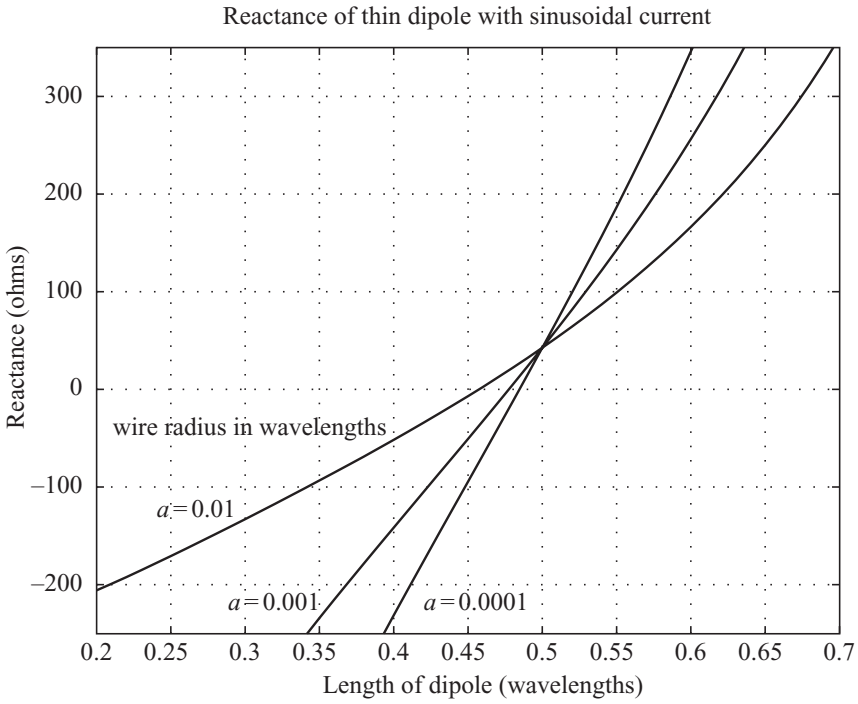


Figure 8.1.34 *The input reactance of a thin wire dipole in free space, which has a sinusoidal current distribution, calculated from the induced emf method. The increasing wire thicknesses give decreasing resonant frequencies. From this method, a half-wavelength dipole has reactance 42.5 Ω, independent of the wire thickness*

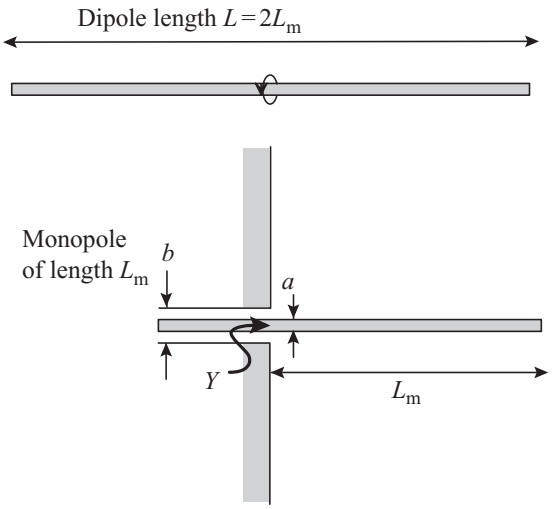


Figure 8.1.35 *Parameters for the monopole on an infinite groundplane and its dipole equivalent with a magnetic current frill as a feed model. The input admittance of the monopole is defined at the groundplane mouth of the coaxial feed*

For lossless, thin monopoles of length L_m (the equivalent dipoles have total length $L = 2L_\mathrm{m}$) the input admittance is

$$Y = \frac{-2\pi}{Z_0 \log(k_0 a \Gamma/2)} + \frac{4\pi}{Z_0 \log(a/(2L_\mathrm{m}))\,(\exp(\mathrm{j}2k_0 L_\mathrm{m}) + 1 + \log(\mathrm{j}4k_0 L_\mathrm{m}\Gamma)/(2\log(k_0 a)))}, \tag{8.1.123}$$

where here $\Gamma = \mathrm{e}^\gamma = 1.781$. The formula above is, in principle, valid for lengths greater than about 0.2 wavelengths and for wire thicknesses of $a \leq 0.01\lambda$. The feed structure is that of a coax feed and the ratio of the inner coax radius, b, and the wire radius, is $b/a \leq 2$. For infinitesimally thin antennas which are n multiples of a half-wavelength, the impedance reduces to

$$Z = 15\log(2\pi\Gamma(1+2n)) + \mathrm{j}15\frac{\pi}{2}, \quad kL_\mathrm{m} = \frac{\pi}{2}(2n+1) \quad \frac{a}{\lambda} \ll 1, \tag{8.1.124}$$

and for the quarter wavelength monopole, $(n = 0)$, $Z = 36.3 + \mathrm{j}23.6$, as above.

For lengths which are a multiple of a wavelength, the admittance reduces to the real value

$$Y = \frac{\log(k_0 L_\mathrm{m}\Gamma)}{60\ \log(k_0 a\Gamma/2)\log(a/(2L_\mathrm{m}))} \tag{8.1.125}$$

showing that the resistance for wavelength-long dipoles is bound for this formulation. The input impedance of a monopole is taken from these and plotted in Figure 8.1.36, showing the finite values at the first antiresonant point.

The dipole reactance is given in Figure 8.1.37, to allow comparison with the values from the induced emf method. The comparison is very good, particularly for the longer lengths on the graph. So either technique can be used, with the wave structure having the advantage of simple (in the computable sense), closed form expression. The differences between the results are much less than those expected as a result of using the monopole version on a relatively small groundplane, less than a wavelength or so in diameter (see below). For calculating mutual impedances, however, integral techniques like the emf method, or the method of moments, are normally used. In Section 8.4, the mutual impedance is given in terms of an integral of the far field patterns,which is a valid formulation for small antennas.

For short monopole lengths such as a quarter wavelength, the wave structure formulation is useful for calculating detailed effects of the finite thickness of the wire, and of course the effect of the length. Figure 8.1.38 shows the impedance against the wire thickness for lengths around a quarter wavelength. Moment methods are hard pressed to calculate these effects accurately because the monopole needs

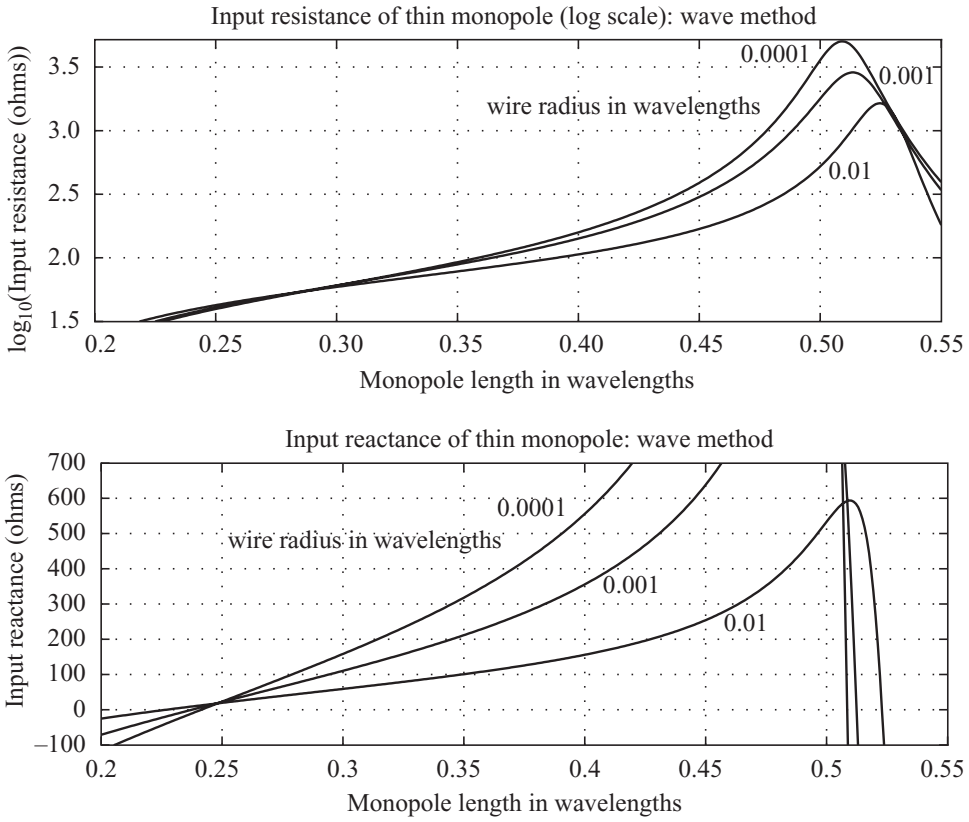


Figure 8.1.36 *The input impedance of a thin wire monopole on a groundplane in free space as a function of the length of the antenna, as calculated from the wave structure method. The first resonance is at about a quarter-wavelength, and the first antiresonance is at about a half-wavelength. Note that the monopole length on the abscissa is half of the corresponding dipole length* $L_{\mathrm{m}} = L/2$

to have many segments (several tens), and for monopoles of practical thickness the length-to-diameter ratio of each wire segment becomes small, violating a typical moment method assumption for wires.

8.1.10.5 Effect on monopole impedance of finite circular groundplane

The effects of finite groundplane are normally calculated from diffraction formulas evaluated at the groundplane edges. Weiner *et al*. (1987) give a thorough treatment of the effect of the quarter wavelength monopole on a finite circular groundplane. They apply several calculation techniques, depending on the groundplane size.

The radiation resistance and the input reactance are shown in Figure 8.1.39 (Weiner *et al*., 1987). The value for the infinite groundplane case ($\sim$36 Ω) is shown as a dotted line. Note that in this reproduced figure, the radii a and b are swapped according to the convention of this section.

The actual values of the radiation resistance oscillate about this as the ground plane size increases, with a period that approaches a wavelength along the groundplane

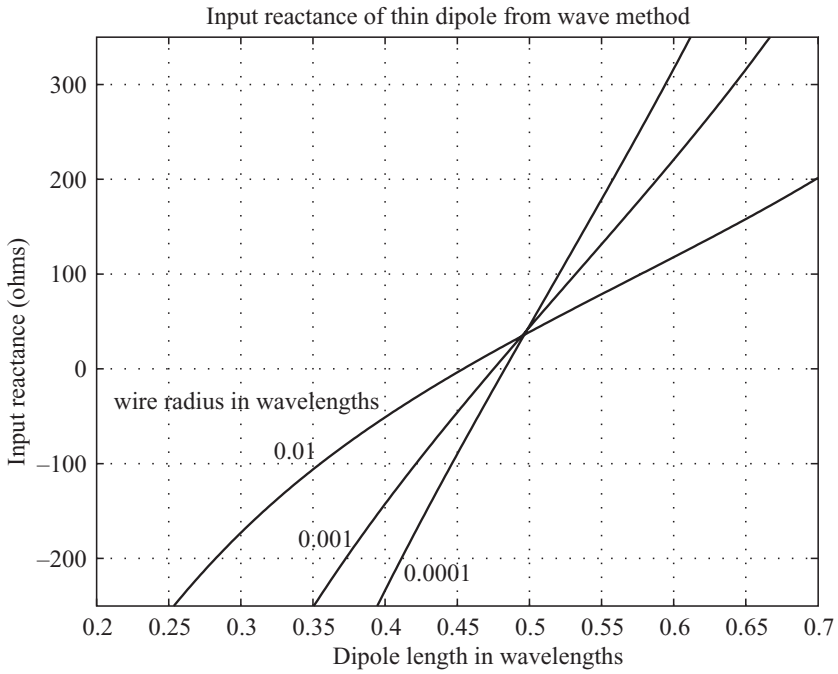


Figure 8.1.37 A closer view of the input reactance of a thin wire dipole around resonance, calculated from the wave structure method, allowing a comparison with the values from the emf method of Figure 8.1.34

radius. The plot is accurate even for very small ground planes of radius approaching 0.1λ. At a groundplane size of 0.5λ, the resistance peaks at about $45\,\Omega$, some 25 per cent above the asymptotic value. Similarly, with a groundplane of about 0.9λ, the resistance is some 14 per cent low, and with a groundplane of about 1.1λ, the resistance is some 18 per cent high, and so on. The reactance follows the same type of behaviour.

The plot is helpful since it offers the designer some feel for the finite groundplane effect. Even for a relatively large groundplane of a couple of wavelengths across, the impedance may be over 10 per cent from its value with an infinite groundplane. These variations do not reduce to less than a few per cent until the groundplane is about 6λ in diameter. So the best operating frequency (for impedance match) of a monopole on a small groundplane depends on the groundplane size. Square and rectangular groundplane shapes tend to cause smaller oscillations because the edge locus is not all at the same distance from the element. This example is a worst case in the sense that the radiation pattern from the element is strongly directed at the edge of the groundplane. For an element that radiates at broadside to the groundplane, such as a fundamental mode patch, the effect of the finite groundplane size is much less.

The patterns tend to be raised by the finite groundplane. The effect is shown in Figure 8.1.40 (Weiner *et al*., 1987), where the elevation patterns are on a linear scale, and these give a feel for the influence of a small groundplane. The plots range from the sleeve monopole (no groundplane) to the infinite groundplane pattern. The groundplane with a radius of one wavelength (bottom left in the figure) shows that

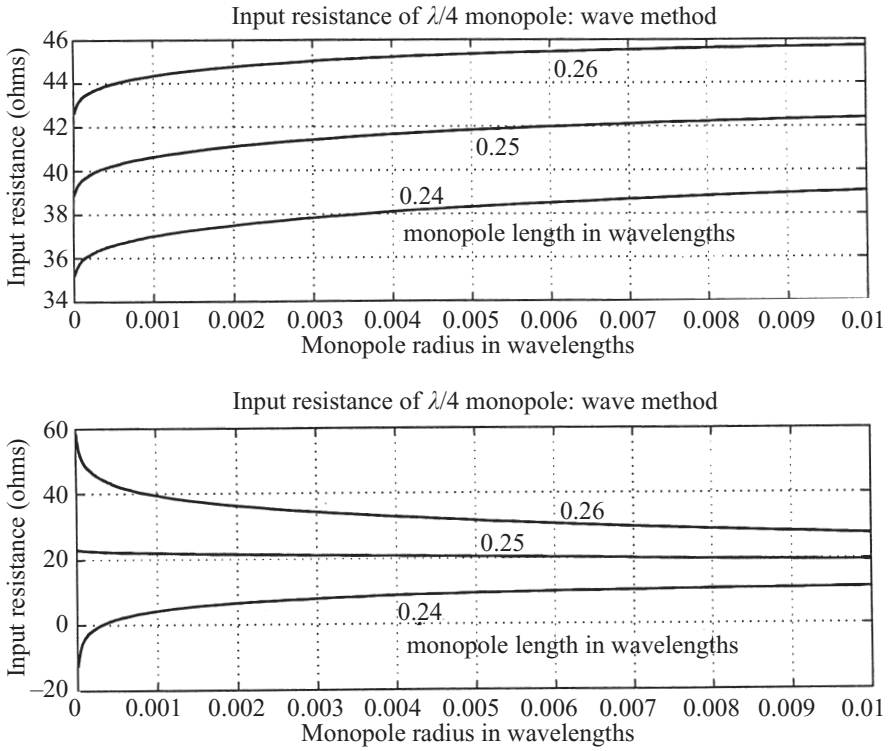


Figure 8.1.38 The impedance of a lossless, quarter wavelength monopole showing the effect of the wire thickness and the length. This impedance is hard to calculate accurately from moment methods for practical monopole thicknesses

even relatively large groundplanes (in the context of compact antennas, for example for handheld terminals, even a wavelength in diameter is large) still change the pattern a great deal.

8.1.11 Ohmic loss for wires

The ohmic loss on a wire antenna is usually very small. Here the losses are estimated for a monopole (cf. Collin, 1985) to demonstrate the calculation.

The input impedance of an antenna is defined from the total dissipated power, and here for a current excitation with peak current I_0, the impedance is

$$Z_a = \frac{P_{rad} + P_{ohmic} + j2\omega\left(W_m - W_e\right)}{(1/2)|I_0|^2}, \tag{8.1.126}$$

where P_{rad} is the radiated power, P_{ohmic} is the power dissipated in the ohmic losses, and $W_m - W_e$ is the difference between the stored energy in the magnetic and electric fields (see Section 8.2).

The ohmic losses are calculated as follows. The magnitude of the current flow in a conductor is a maximum at the surface and decays exponentially with conductor

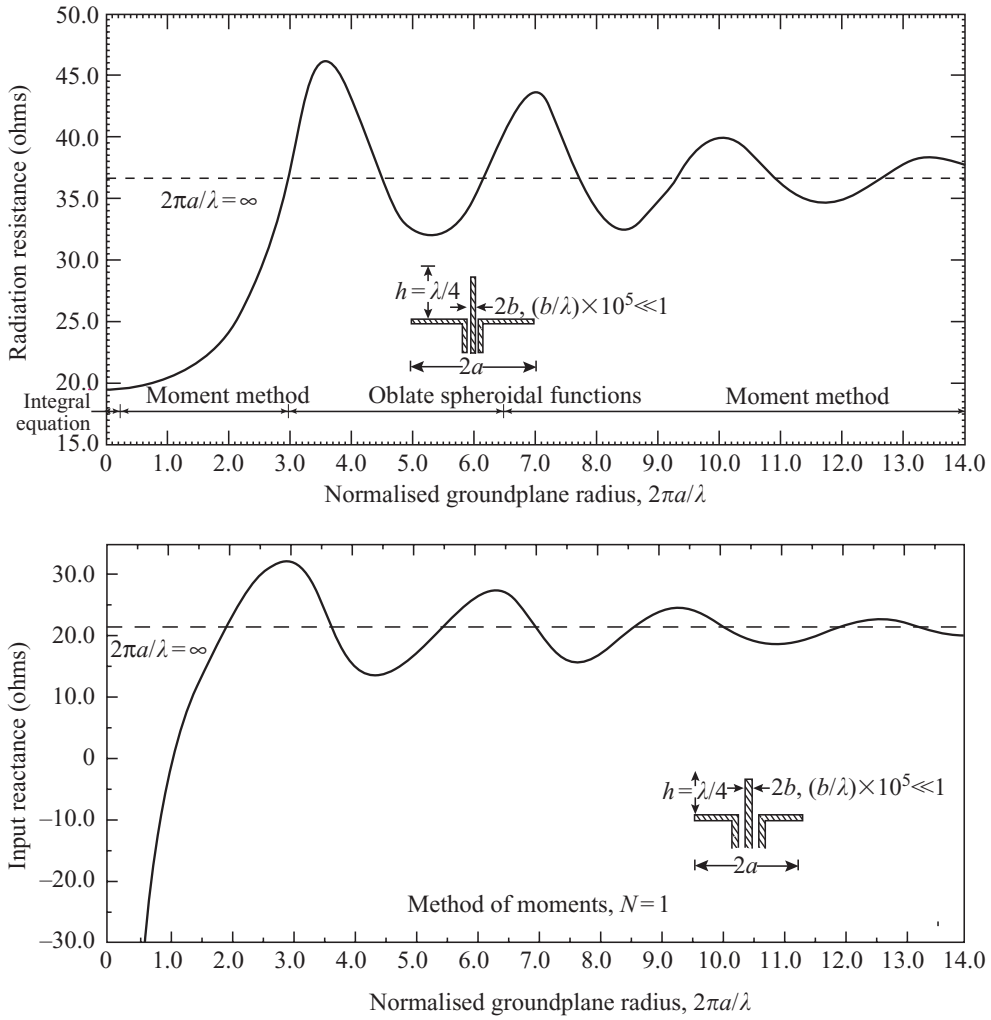


Figure 8.1.39 The radiation resistance and input reactance of a quarterwave monopole on a finite, circular groundplane. In this figure from Weiner et al. (1987), 'a' is the groundplane radius and 'b' is the monopole radius, a swap of terminology from the rest of the section

depth, i.e. as e^{-x/δ_S} (x is the depth). The skin depth, δ_S, is when the current has reduced to 1/e of its surface value, and

$$\delta_S = \sqrt{\frac{2}{\omega\mu\sigma}} \tag{8.1.127}$$

where σ is the conductivity and the surface resistance is $R_S = 1/(\sigma\delta_S)$. Note that the skin depth is inversely proportional to the square root of frequency and the surface resistance is directly proportional to the square root of the frequency. The skin depth is depicted in Figure 8.1.41.

The current distribution is now approximated as uniform within the skin depth. The power loss follows as $P_{\text{ohmic}} = 1/2|I_0|^2 R_S$. For a round wire of radius a, the

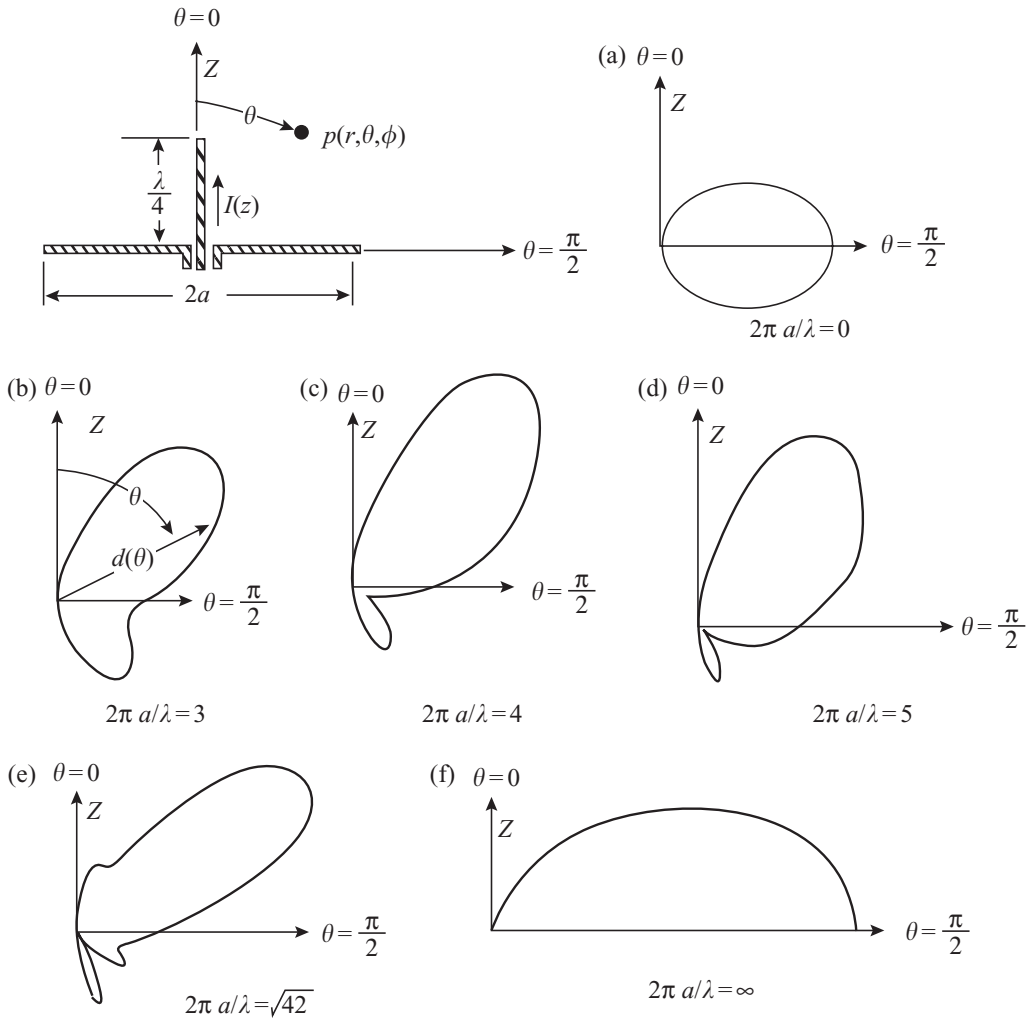


Figure 8.1.40 Radiation patterns on a common linear scale for a quarterwave monopole on a finite, circular groundplane (Weiner et al., 1987). Here, 'a' is the groundplane radius and 'b' is the monopole radius, a swap of terminology from the rest of the section

cross section of the current is about $2\pi a\delta_S$ and the ohmic power loss for a wire of length L with a sinusoidal current density along the length is

$$P_{\text{ohmic}} = \frac{1}{2}\int_0^L \left(\frac{I_0}{2\pi a}\right)^2 \cos^2 kz \frac{1}{\sigma\delta_S}\,\mathrm{d}z \tag{8.1.128}$$

and equating this to a network model with losses $1/2I_0^2 R_{\text{ohmic}}$ gives the resistance contribution to the input impedance for a quarter wavelength monopole as

$$R_{\text{ohmic}}^{(\lambda/4)} = \frac{\lambda_0}{16\pi a\sigma\delta_S}. \tag{8.1.129}$$

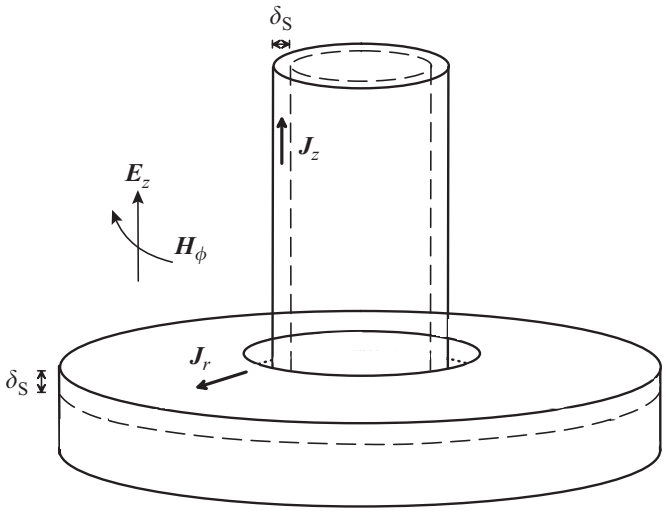


Figure 8.1.41 The skin depth approximates the depth of a sheet of constant current density, here indicated for a monopole on a circular groundplane

Table 8.5 The loss resistance of a quarter wavelength monopole element with wire diameter 1 mm and frequency 1.9 GHz, for various materials. The value is doubled for the dipole equivalent

Wire material	Conductivity (S m^{-1} or mho m^{-1})	Resistance (Ω)
Silver	$6.17(10)^7$	0.069
Copper	$5.80(10)^7$	0.071
Aluminium	$3.72(10)^7$	0.089
Brass	$1.57(10)^7$	0.137
Stainless steel	$1.35(10)^7$	0.148
Solder	$0.71(10)^7$	0.204

For the $\lambda/2$ dipole, this value is doubled. Table 8.5 gives some resistance values for different wire materials. Note that the monopole losses do not include losses from some groundplane or other feeding structure.

The groundplane for the monopole element can contribute more losses than in the element itself. The surface current is found from the tangential magnetic field, $\boldsymbol{J}_S = \hat{n} \times \boldsymbol{H}_S$. For a circular groundplane having a radial extent of r_b to r_u (r_b is the outer diameter of the coax feed, say) the dissipated power in the groundplane is found from

$$P_{\text{ohmic}} = \int_0^{2\pi} \int_{r_b}^{r_u} \frac{1}{2} \left| \boldsymbol{J}(r) \right|^2 \frac{1}{R_S} \, \mathrm{d}r. \tag{8.1.130}$$

8.1.12 Radiation and ohmic resistance for small loops

For electrically small loop antennas, the radiation resistance in ohms is found as

$$R_{\text{rad}}^{(\text{single loop})} = 20\pi^2 \left(\frac{C}{\lambda}\right)^4, \tag{8.1.131}$$

where C is the circumference of the circular loop. For example, a circular loop of diameter $\lambda/10$ has a radiation resistance of just $1.9\,\Omega$. Clearly the resistance is critically dependent on the loop size, the dependence being a quartic, as opposed to the resistance of the short dipole, which has a quadratic dependence on its length. However, if N identical loops are wound so that the same magnetic flux passes through each one, then the radiation resistance can be increased by a factor of N^2, i.e.

$$R_{\text{rad}}^{(N\text{ loops})} = 20\pi^2 N^2 \left(\frac{C}{\lambda}\right)^4. \tag{8.1.132}$$

The total electrical length of the wire should be kept small for this equation to hold, i.e. $NC/\lambda \ll 1$.

When the loops are not in close proximity (spacing between conductors greater than five times the wire radius), the ohmic losses of the loop are calculated directly from the losses on a wire, using the skin depth ($\delta = \sqrt{2/(\omega\,\mu_0\sigma)}$ m) times the perimeter for the current-carrying cross section. This gives the ohmic resistance (in Ω), as

$$R_0 = \frac{NC}{P}\sqrt{\frac{\omega\mu_0}{2\sigma}}, \tag{8.1.133}$$

where NC is the total length of the loop (all turns), P is the perimeter of the wire cross section, and the square root term is the surface resistance $R_{\text{S}} = 1/(\sigma\delta)$, which includes the conductivity, σ mhos. For N circular loops spaced well apart, having conductor radius a, the resistance per length is

$$R_0^{(\text{l})} = \frac{N}{2\pi a} R_{\text{S}}\ \Omega\ \text{m}^{-1}. \tag{8.1.134}$$

For the above example, with a $\lambda/10$ diameter loop, here with a thin copper ($\sigma = 5.8(10)^7$ mho m^{-1}) circular wire of diameter $\lambda/1000$, the ohmic resistance is $0.83\,\Omega$. The radiation efficiency is then $1.9/(1.9+0.83) = 70$ per cent, and this excludes any matching circuit losses. The wire diameter should be made as large as practicable and the number of turns, N, is kept as low as possible, in order to minimise the ohmic losses. In practice, however, many turns are used because of the overriding effect of the strongly increased (proportional to N^2) radiation resistance. If the turns are closely packed, then the proximity effect needs to be considered, as follows.

8.1.12.1 Proximity effect on ohmic resistance and radiation efficiency

If the multiple loops are spaced by less than four to five times the wire radius, then the mutual coupling between the conductors creates uneven surface currents. This causes the ohmic losses to be greater than those expected from the skin depth formulas above, and the radiation efficiency will decrease.

For closely spaced loops, the effect of the close proximity of the wires, and the radiation resistance, is given in Smith (1972). Figures 8.1.42 and 8.1.43 show these effects. The additional radiation resistance per length is denoted $R_\mathrm{p}\ \Omega\ \mathrm{m}^{-1}$, and the radiation efficiency, neglecting any matching losses, can be expressed as

$$\eta_\mathrm{rad} = \frac{20\pi^2 N^2 k_0^2 b^4}{20\pi^2 N^2 k_0^2 b^4 + (Nb/a)(R_\mathrm{p}/R_0 + 1)} \quad \text{(zero matching loss)}. \qquad (8.1.135)$$

The resistance increases very quickly as the wires approach touching, which corresponds to the left axis position in Figure 8.1.42. As an example from the curve, for a 1 mm diameter wire, coated with lossless insulator of thickness $c - a = 0.05$ mm so that the wire centres are spaced by $c/a = 1.1$, then the additional resistance is almost 2.5 times the resistance calculated from the above formula for well-spaced wires.

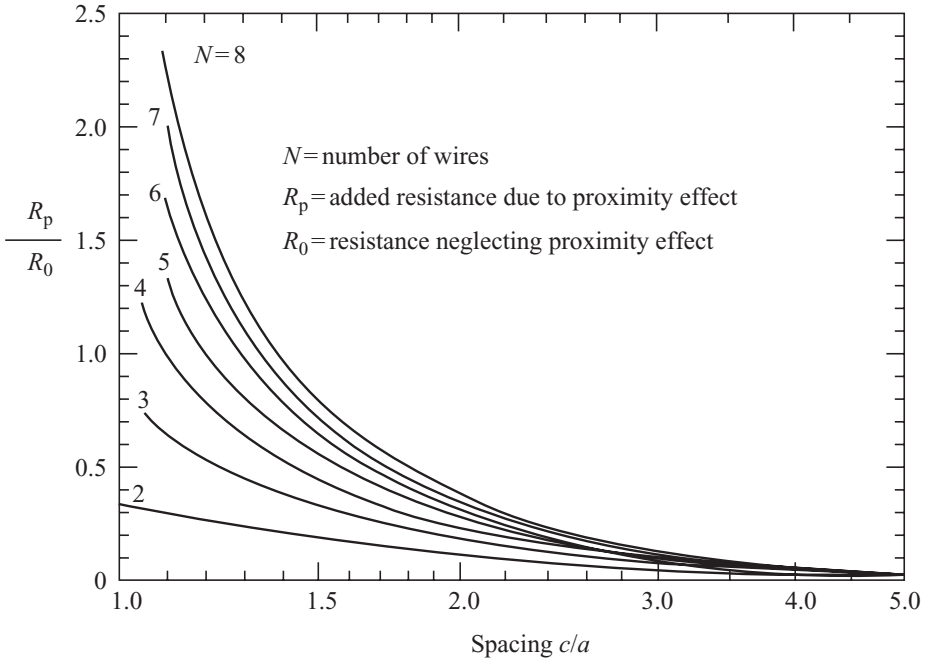


Figure 8.1.42 The proximity effect of closely spaced parallel wires in a multi-turn loop antenna which causes an additional resistance per length. Here the wire diameter is 2a and the spacing between conductor centres is 2c, and the separation of the wire surfaces is 2(c − a) (Smith, 1972)

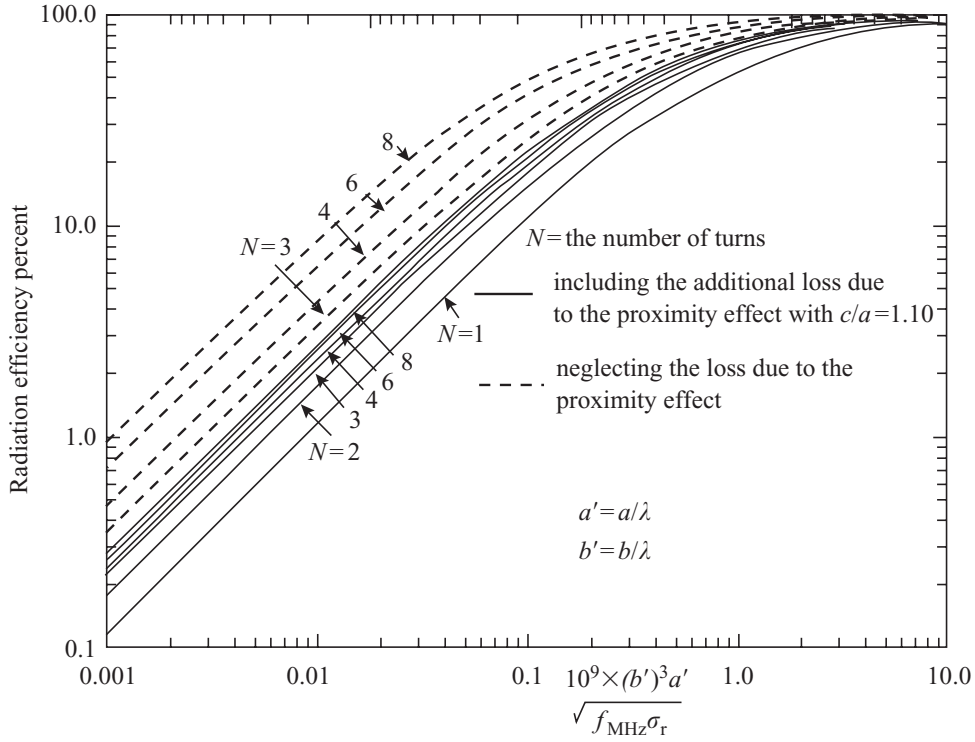


Figure 8.1.43 The additional resistance from the proximity effect of closely spaced wires decreases the radiation efficiency of the small loop. Here b is the radius of the loop and the abscissa is dimensionless (Smith, 1972)

The radiation efficiency, with zero matching loss, is plotted in Figure 8.1.43. Here the above equation is rearranged to be

$$\eta_{\mathrm{rad}} = \frac{1}{1 + (8.48(10)^{-10}\sqrt{f_{\mathrm{MHz}}\sigma_r}/(Na'b'^3))(R_p/R_0 + 1)} \quad \text{(zero matching loss)}, \tag{8.1.136}$$

where σ_r is the conductivity of the wire relative to the conductivity of copper, and $a' = a/\lambda$ and similarly for b'. The dashed lines are for no proximity effect, and the solid lines include the additional losses from a close conductor spacing, $c/a = 1.1$. The plot shows that a three-turn loop which is well spaced has a better efficiency than the closely spaced eight-turn loop.

8.1.12.2 Ferrite loaded loops

The radiation resistance of the small loop is very small, and this combined with the ohmic losses in the loop plus matching circuit, makes the efficiency of this antenna low. Making the loop larger is one solution to increasing the radiation resistance, and this can be effectively accomplished by placing a high permeability core, made of ferrite material, in the wire loop. The magnetic flux passing through the loop is increased without necessarily making the physical dimensions larger.

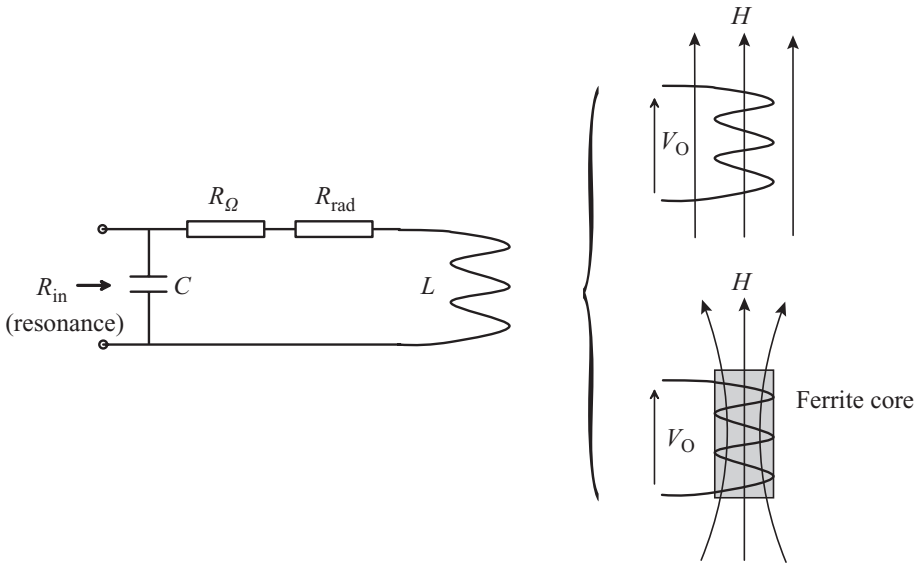


Figure 8.1.44 *The small loop antenna is usually resonated with a parallel capacitor to assist with matching. The inclusion of a ferrite core allows the antenna to be more compact, through a reduction in the number of turns or the physical cross section area*

The ferrite core loop typically requires a matching circuit which includes a parallel capacitor for attaining resonance. It has a narrow band (for a fixed capacitor) with low efficiency. Figure 8.1.44 shows the equivalent circuit with matching capacitor. The practical advantage is that it can be extremely compact, and for this reason, it is widely used for broadcast reception for portable radios. Some basic design criteria are given below. The inefficiency of the broadcast receive antennas is compensated for by the large powers used in broadcasting. In fact, for the low frequencies used in broadcasting (AM stations as low as 500 kHz, for example), the antenna noise temperature is much higher than that at microwave frequencies. This means, from a reception viewpoint, that the antenna can be allowed to be inefficient because the dominant noise source is not from the thermal sources at the receiver, but stems from the received signal. The antenna can be made increasingly compact by allowing decreasing efficiency, until the antenna noise and the thermal noise are at comparable levels.

With ferrite with relative permeability $\mu_{\rm fr} = \mu_{\rm f}/\mu_0$ occupying the core of the loop, the antenna is analysed in terms of a relative effective permeability $\mu_{\rm er} = \mu_{\rm e}/\mu_0$. The radiation resistance of the N-turn circular ferrite core loop modifies to

$$R_{\rm f} = 20\pi^2 \left(\frac{C}{\lambda}\right)^2 N^2 \mu_{\rm er}^2. \tag{8.1.137}$$

So for a given excitation current, the radiated power of a ferrite loop increases relative to an air-cored loop by $\mu_{\rm e}^2$.

In the receiving case, with the loops in the $\theta = \pi/2$ plane, the effective height (Section 2.1), is $\boldsymbol{h}_{\rm L} = -{\rm j}\omega(\mu_{\rm er}\mu_0/Z_0)NS\sin\theta\ \hat{\phi}$. With a magnetic field passing

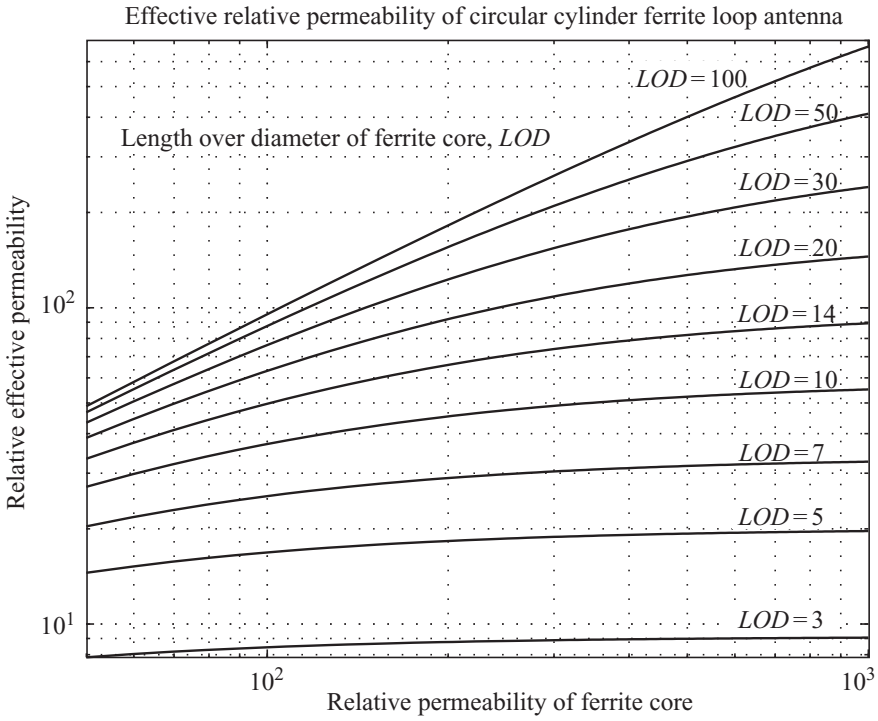


Figure 8.1.45 Design curves for the effective relative permeability for a cylindrical ferrite core

directly through the loops, the open circuit voltage from a free space field, say $\boldsymbol{H} = H\hat{\theta}$, $\boldsymbol{E} = Z_0 H\hat{\phi}$, is from $V_\mathrm{O} = \boldsymbol{h}_\mathrm{L} \cdot \boldsymbol{E}$,

$$V_\mathrm{O} = -\mathrm{j}\omega\mu_\mathrm{e}\mu_0 NSH \tag{8.1.138}$$

in which S is the projected cross-section area of one loop.

The effective permeability has been measured by Wolff (1966) for a cylindrical ferrite core, and is given by

$$\begin{aligned}\mu_\mathrm{e} &= \frac{\mu_\mathrm{f}}{1 + D_\mu(\mu_\mathrm{f} - 1)}, \\ &\approx \frac{1}{D_\mu} \quad (\mu_\mathrm{f} \text{ large});\end{aligned} \tag{8.1.139}$$

where the empirical demagnetisation factor D_μ is modelled by

$$D_\mu \approx \mathrm{e}^{-(1.54\ \log(LOD)+0.52)} \tag{8.1.140}$$

and LOD is the length over diameter of the ferrite core. The formula is claimed to hold for LODs between about 3 and 100. Some design curves for the effective relative permeability are given in Figure 8.1.45. The demagnetisation factor is a function of the geometry of the core and so the formula is not accurate for other than circular

Table 8.6 The usual symbols for the field quantities

Physical quantity	Usual symbol	Dimensions	Units
force	f	MLT^{-2}	N
electric charge	q	Q	C
surface charge density	ρ or ρ_S	QL^{-2}	C m^{-2}
volume charge density	ρ or ρ_V	QL^{-3}	C m^{-3}
voltage scalar potential electro-motive force	V	$ML^2T^{-2}Q^{-1}$	V
electric current	I	QT^{-1}	A
surface current density	$\boldsymbol{K}$ or $\boldsymbol{J}$ or $\boldsymbol{J}_S$	$QT^{-1}L^{-1}$	A m^{-1}
volume current density	$\boldsymbol{J}$	$QT^{-1}L^{-2}$	A m^{-2}
electric field	$\boldsymbol{E}$	$MLT^{-2}Q^{-1}$	V m^{-1}
electric flux	Ψ	Q	C
electric flux density	$\boldsymbol{D}$	QL^{-2}	C m^{-2}
magnetic field	$\boldsymbol{H}$	$QL^{-1}T^{-1}$	A m^{-1}
magnetic flux	Φ	$ML^2T^{-1}Q^{-1}$	Wb
magnetic flux density	$\boldsymbol{B}$	$MT^{-1}Q^{-1}$	Wb m^{-2}
vector magnetic potential	$\boldsymbol{A}$	$MLT^{-1}Q^{-1}$	Wb m^{-1}
vector electric potential	$\boldsymbol{F}$	QL^{-1}	C m^{-1}
conductivity	σ	$Q^2TM^{-1}L^{-3}$	mho m^{-1}
permittivity	$\epsilon = \epsilon_0\epsilon_r$	$Q^2T^2M^{-1}L^{-3}$	F m^{-1}
permeability	$\mu = \mu_0\mu_r$	MLQ^{-2}	H m^{-1}
permittivity of free space	$\epsilon_0 \approx 1/(36\pi)\ (10)^{-9}$ F m^{-1} $(= CV^{-1}m^{-1} = AsV^{-1}m^{-1})$		
permeability of free space	$\mu_0 = 4\pi(10)^{-7}$ H m^{-1} $(= WbA^{-1}m^{-1} = VsA^{-1}m^{-1})$		
speed of light in free space	$c = 1/\sqrt{\mu_0\epsilon_0} \approx 3(10)^8$ m s^{-1}		
intrinsic impedance of free space	Z_0 or $\zeta_0 = \sqrt{\dfrac{\mu_0}{\epsilon_0}} \approx 120\pi\ \Omega$ per square		

cylindrical shapes. Other cylindrical shapes (rectangular, etc.) are not as effective as circular cross-sections, and there seems little advantage in using ferrite cores which are shaped along their length.

8.1.13 Summary of field quantities and relations

8.1.13.1 Table of field quantities and units

Table 8.6 gives the usual symbol and the units for some of the electromagnetic field quantities. The units are in length in metres (L), mass in kilograms (M) and time in seconds (T). Other quantities can be deduced from these. The table is useful for checking results (or rather the units of results) from integral and differential equations for field calculations.

Table 8.7 The del operator for the differentiation of scalars and vectors

Differential operation	Result	Rectangular coordinate equation for scalar A and vector $\boldsymbol{A}$
gradient of a scalar (grad)	vector	$\text{grad } A = \nabla A = \hat{\boldsymbol{x}}\frac{\partial A}{\partial x} + \hat{\boldsymbol{y}}\frac{\partial A}{\partial y} + \hat{\boldsymbol{z}}\frac{\partial A}{\partial z}$
divergence of a vector (div)	scalar	$\text{div } \boldsymbol{A} = \nabla \cdot \boldsymbol{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$
curl of a vector (curl)	vector	$\text{curl } \boldsymbol{A} = \nabla \times \boldsymbol{A} = \hat{\boldsymbol{x}}\left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right) + \hat{\boldsymbol{y}}\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right) + \hat{\boldsymbol{z}}\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)$
Laplacian of a scalar (div grad)	scalar	$\text{div grad } A = \nabla \cdot \nabla A = \nabla^2 A = \frac{\partial^2 A_x}{\partial^2 x} + \frac{\partial^2 A_y}{\partial^2 y} + \frac{\partial^2 A_z}{\partial^2 z}$
Laplacian of a vector (div grad)	vector	$\text{div grad } \boldsymbol{A} = \nabla \cdot \nabla \boldsymbol{A} = \nabla^2 \boldsymbol{A} = \hat{\boldsymbol{x}}\nabla^2 A_x + \hat{\boldsymbol{y}}\nabla^2 A_y + \hat{\boldsymbol{z}}\nabla^2 A_z$
curl of a curl of a vector (curl curl)	vector	$\nabla \times \nabla \times \boldsymbol{A} = \hat{\boldsymbol{x}}\left\{\frac{\partial}{\partial y}\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) - \frac{\partial}{\partial z}\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial y}\right)\right\} + \hat{\boldsymbol{y}}\ldots$ $= \nabla(\nabla \cdot \boldsymbol{A}) - \nabla^2 \boldsymbol{A}$ $= -\nabla^2 \boldsymbol{A}, \quad \text{if } \nabla \cdot \boldsymbol{A} = 0.$

8.1.13.2 Vector differential operators

The differential operators are used extensively in antenna design and analysis. Only the rectangular coordinates versions are given here (Table 8.7). The grad, div and curl operators all act to differentiate in space, so the units after the operation are modified by L^{-1}.

8.1.13.3 Gradient (Grad)

The gradient is the familiar differentiation, but in all spatial coordinates. ∇A has a direction orthogonal to any isosurface of A (i.e. surfaces of constant $A = A_0$), and so the magnitude, $|\nabla A|$, gives the maximum rate of change of A at that point in space.

8.1.13.4 Divergence (Div)

The divergence represents the difference of flux lines entering and exiting a vanishing small volume. So for source-free regions, the flux lines may be curving but no new lines are added or subtracted, and the divergence is zero. Such source-free fields are called solenoidal fields, the usual example being the divergence of the magnetic field always being zero, $\nabla \cdot \boldsymbol{B} = 0$, since it is assumed that there are no magnetic monopoles.

Table 8.8 Maxwell's equations are postulates for linking the fields and sources of charge distribution and currents, here for isotropic, free space. For harmonic time dependence, the time derivative is replaced with $j\omega$. All the equations must be satisfied simultaneously for any closed contours L and surfaces S

	Integral form of Maxwell's equations		Differential form of Maxwell's equations
(MI1)	$\int_S \nabla \times \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S} = \int_V \rho_v \mathrm{d}V$ (C) (Gauss's law)	(MD1)	$\nabla \cdot \boldsymbol{D} = \rho_v$ (C m^{-3})
(MI2)	$\int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S} = 0$ (Wb) (Gauss's law for magnetic fields)	(MD2)	$\nabla \cdot \boldsymbol{B} = 0$ (Wb m^{-3})
(MI3)	$\int_L \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{L} = -\frac{\mathrm{d}}{\mathrm{d}t} \int_S \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{S}$ (V) (Faraday's law)	(MD3)	$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}$ (V m^{-2})
(MI4)	$\int_L \boldsymbol{H} \cdot \mathrm{d}\boldsymbol{L} = \int_S \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S} + \frac{\mathrm{d}}{\mathrm{d}t} \int_S \boldsymbol{D} \cdot \mathrm{d}\boldsymbol{S}$ (A) (Ampere's law)	(MD4)	$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}$ (A m^{-2})

If there are sources (charges) present within the volume, then the divergence operation may yield a non-zero result, e.g. for the electric field, giving the corresponding Maxwell equation, $\nabla \cdot \boldsymbol{D} = \rho_v$ C m^{-3}.

8.1.13.5 Divergence theorem

The divergence theorem states that the sum of the divergences calculated within a volume (V) is the same as the sum of the flux lines exiting the surface (S) of the volume, $\int_V \nabla \boldsymbol{A}\, \mathrm{d}V = \int_S \boldsymbol{A} \cdot \mathrm{d}S$. Applying this to the electric and magnetic fields in two of Maxwell's equations in their integral form, and dividing through by a vanishing small volume element, ΔV, allows conversion of the equations to their differential form. These are denoted MI1 and MI2; and MD1 and MD2 respectively, in Table 8.8.

8.1.13.6 Curl

The curl is the measure of how 'curled up' a field is. It originated from fluid motion studies and is also known as 'rot', short for rotation. A way to envisage the operator is to compare it to the rotating action on a paddle wheel in the fluid (flux lines of field). If the fluid flow can act to turn the wheel then the curl is non-zero and the component of the curl of the field is in the direction of the wheel axis, given by the right hand screw rule. The curl operation gives the vector sum of the contributions from three orthogonal paddle wheels (see Figure 8.1.46).

In a propagating wave, the electric and magnetic fields are bound together and have a unique velocity, c. The curl operator relates the fields, and this is depicted with harmonic relations in Figure 8.1.47.

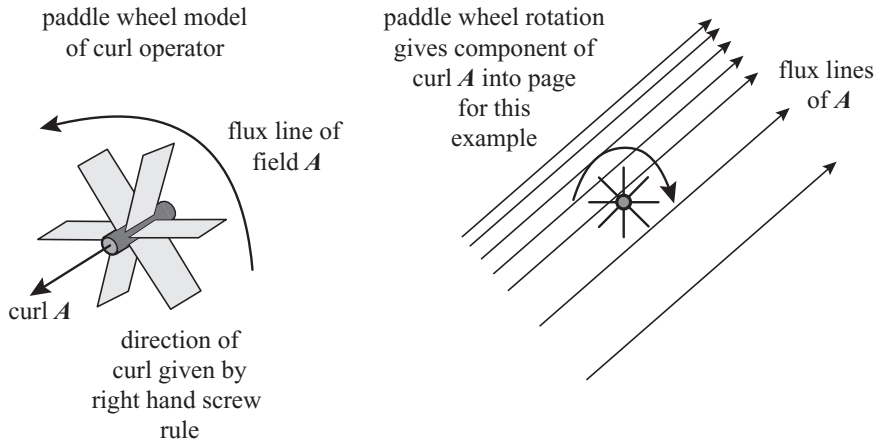


Figure 8.1.46 *The paddle wheel model to envisage the curl operator. The flux lines of a field are like a fluid flow whose net effect on the rotation of the submerged paddle wheel gives the curl. The direction of each component of the curl is from the right-hand screw rule, and the vector sum of these gives the direction of* $\nabla \times \boldsymbol{A}$

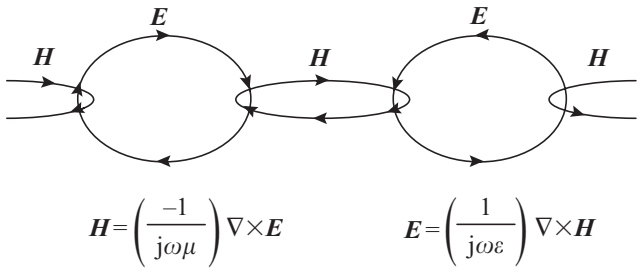


Figure 8.1.47 *The curl operation links the electric and magnetic fields, here depicted using single flux lines*

8.1.13.7 Stoke's theorem

Stoke's theorem states that the sum of the curls, or circulations, calculated in an area (S, with elemental direction $\mathrm{d}\boldsymbol{S}$ perpendicular to S) is the same as the circulation around the area, $\int_S \nabla \times \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{S} = \int_L \boldsymbol{A} \cdot \mathrm{d}\boldsymbol{L}$. Here $\mathrm{d}\boldsymbol{L}$ is the elemental length along the perimeter L in direction and magnitude. The theorem can be applied to the remaining two Maxwell equations (MI3 and MI4) and dividing through by a small surface element, ΔS, allows a conversion of them to their differential form (MD3, MD4).

8.1.13.8 Laplacian

The Laplacian operator is a double differentiation in space. The wave equation is compactly expressed using the Laplacian. In free space, with no sources (no charges or currents), then both the electric and magnetic fields follow the homogeneous wave equation, here for the electric field, $\nabla^2 \boldsymbol{E} - k^2 \boldsymbol{E} = 0$, and this holds for each scalar component as well.

8.1.13.9 Maxwell's equations

Maxwell's equations are given in Table 8.8. Here, the electric and magnetic fields are space- and time-dependent: $\boldsymbol{E}(\boldsymbol{r},t) = \epsilon_0 \boldsymbol{D}(\boldsymbol{r},t)$, $\boldsymbol{H}(\boldsymbol{r},t) = \mu_0 \boldsymbol{B}(\boldsymbol{r},t)$; and the net charge inside any closed surface is $\int_V \rho_v \mathrm{d}V = q(t)$; and the net current flowing through any open surface bounded by the closed line L is $\int_S \boldsymbol{J}\mathrm{d}S = I(t)$. The latter two quantities are taken as time dependent only, i.e. not space-dependent. This difference is reflected in the use of full derivatives and partial derivatives in the appropriate Maxwell's equations.

8.2 Compact elements

In mobile and personal communications, it is normally important to have very compact antennas. Physical compactness means a less intrusive presence at both handsets and base stations. However, there is a tradeoff between antenna performance, in the sense of impedance bandwidth, and its electrical size. For high directivity, an antenna must become electrically large, occupying apertures of several wavelengths. Similarly, both the bandwidth and the directivity of an antenna cannot be large if the antenna is electrically too small.

For personal terminals, it is necessary to have physically small antennas, and with current carrier frequencies of about 1 GHz for cellular systems, for example, this means the handsets must bear electrically small antennas as well. The performance of compact handset antennas is influenced by their immediate surroundings, in particular, the operator's hand and head. In practice, for currently used handset antennas, these effects are almost impossible to isolate from the handset performance. As carrier frequencies increase in future systems, it becomes more feasible to mount large numbers of uncorrelated antennas on a handset because the structure becomes electrically large. However, the performance of these antennas will still be influenced by the presence of the immediate surroundings.

In this section, fundamental limits of the compactness of element antennas are discussed, and some basic compact element designs are compared.

8.2.1 Introduction to antenna Q from network considerations

The Q, or *quality factor*, is a fundamental parameter for the impedance bandwidth. In order to compare various antennas it is important to understand the basic definitions of Q.

8.2.1.1 Basic definitions

In the classic text by Montgomery *et al*. (1948) the Q of a circuit is defined by

$$Q = 2\pi \frac{\text{energy stored}}{\text{energy dissipated per cycle}} \tag{8.2.1}$$

or alternatively

$$Q = \frac{\omega\ (\text{time averaged energy stored})}{\text{energy dissipated per second}}. \tag{8.2.2}$$

If we assume that the losses are purely from radiation (i.e. a lossless antenna), the Q thus defined is the unloaded Q, sometimes denoted Q_u where a lack of subscript may be ambiguous.

The averaged energy stored by the antenna may be both electric energy, W_e, and magnetic energy W_m. These are expressed as integrals over the antenna volume,

$$W_e = \frac{1}{4}\int \epsilon\, |E|^2 \,dv, \quad W_m = \frac{1}{4}\int \mu\, |H|^2 \,dv \tag{8.2.3}$$

where E and H are the electric and magnetic field peak values respectively. The time-dependent energy in a capacitor C for example, is $W_e(t) = 1/2C|V_{pk}(t)|^2 \cos^2(\omega t)$, and so the time-average is $W_e = 1/4C|V_{pk}|^2$, where V_{pk} is the peak voltage.

Denoting the ohmic power loss and the radiated power with P_{ohmic} and P_{rad}, respectively, the Q can be written

$$\begin{aligned} Q &= \frac{\omega\,(W_m + W_e)}{P_{ohmic} + P_{rad}} \\ &= \frac{\omega\,(W_m + W_e)}{P_{rad}} \quad \text{(lossless antenna)}. \end{aligned} \tag{8.2.4}$$

8.2.1.2 Relation to reactance

The stored energies are also related to the reactance or susceptance, so for an input current-defined system the reactance is (Montgomery *et al*., 1948)

$$X = \frac{2(W_m - W_e)}{|I_{eff}|^2} \tag{8.2.5}$$

where I_{eff} is the effective, i.e. the rms current value, and the derivative is

$$\frac{\partial X}{\partial \omega} = \frac{2(W_m + W_e)}{|I_{eff}|^2}. \tag{8.2.6}$$

This is the basis of Foster's reactance theorem which states that the slope of the reactance of a lossless network as a function of frequency is always positive. The radiation losses (or any other losses) allows the slope to be negative.

Denoting the radiation resistance with R, the power is $P = R|I_{eff}|^2$, and the Q can be expressed in terms of the slope of the reactance, using equations (8.2.4) and (8.2.6), as

$$Q = \frac{\omega}{2R}\,\frac{\partial X}{\partial \omega}. \tag{8.2.7}$$

The dual expression for a voltage defined terminal is $Q = \omega/2G\ \partial B/\partial\omega$ where the slope of the susceptance B is $\partial B/\partial\omega = 2(W_m + W_e)/|V_{eff}|^2$ and V_{eff} is here the rms voltage.

These expressions are general and exact. They emphasise that the Q is a narrow-band parameter defined at one frequency. In a measurement situation, the absolute value may need to be taken of the slope of the reactance in some situations. This is because the slope may not always be positive in a real-world, lossy circuit.

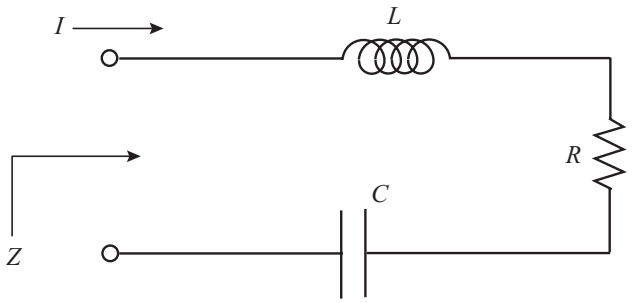


Figure 8.2.1 A series RLC circuit is the simplest resonant circuit and models the impedance of an antenna (or antenna plus resonating network) near resonance

8.2.1.3 Q at resonance

Usually, the Q is only of interest when there is resonance, i.e. when $X = 0$ and the stored energies are the same, namely $W_e = W_m = W$, so

$$Q = \frac{2\omega\, W}{P} \quad \text{(resonance)}. \tag{8.2.8}$$

Often a small antenna is not in itself resonant, e.g. $W_m > W_e$, in which case a network element with electrical stored energy is added, until $W_m = W_e$, and resonance is achieved. Equation (8.2.7) may then be used at resonance to find the Q, where the frequency dependence of the total reactance (the resonating network plus the antenna) is included. This is the method used in the following for numerically defined impedances (or admittances).

8.2.1.4 Circuit model near single resonance

Near resonance the impedance equation takes on the characteristics of a series *RLC* network, see Figure 8.2.1. The resonant frequency is $\omega_o = 1/\sqrt{LC}$. The input impedance can be written

$$\begin{aligned} Z &= R + \mathrm{j}\left(\omega\, L - \frac{1}{\omega\, C}\right) \\ &= R\left(1 + \mathrm{j}Q\left(\frac{\omega}{\omega_o} - \frac{\omega_o}{\omega}\right)\right). \end{aligned} \tag{8.2.9}$$

This agrees with the expression using the slope of the reactance above (equation (8.2.7)).

8.2.1.5 Bandwidth near single resonance

The 3 dB double-sided relative frequency bandwidth, $\Delta\omega_{3\,\mathrm{dB}}$, (or $\Delta f_{3\,\mathrm{dB}}$ in Hz) about resonance (i.e. where $Z = R + \mathrm{j}R$) follows as

$$\frac{\Delta\omega_{3\,\mathrm{dB}}}{\omega_0} = \frac{\Delta f_{3\,\mathrm{dB}}}{f_0} = \frac{1}{Q}. \tag{8.2.10}$$

This relation assumes that the real part of the impedance is fixed while the reactance changes, as frequency is changed. However, this is seldom the case except perhaps over very narrow bandwidths.

For a matched antenna, the loaded bandwidth is twice this value, or the loaded Q is half the unloaded Q,

$$Q_{\mathrm{L}} = \tfrac{1}{2} Q \quad \text{(matched)}. \tag{8.2.11}$$

Often it is more convenient from measurements to use the *VSWR* or the voltage reflection coefficient, Γ, to estimate the Q. The relation is $VSWR = (1 + |\Gamma|)/(1 - |\Gamma|)$. The antenna is loaded during the measurement, typically by 50 Ω. (A match is not strictly necessary for estimating the unloaded Q.) The cases for $\Gamma = -10\,\mathrm{dB}$ ($VSWR = 1.925$) and $VSWR = 2$ ($\Gamma = -9.54\,\mathrm{dB}$) are almost the same; the expressions for the (unloaded) Q are:

$$Q = \frac{1}{\sqrt{2}} \left(\frac{\Delta\omega_{VSWR=2}}{\omega} \right)^{-1} \quad (VSWR = 2) \tag{8.2.12}$$

$$Q = \frac{2}{3} \left(\frac{\Delta\omega_{\Gamma=-10}}{\omega} \right)^{-1} \quad (\Gamma = -10\,\mathrm{dB}) \tag{8.2.13}$$

where $\Delta\omega_{\{\cdot\}}$ is the appropriately defined double-sided observed bandwidth.

In terms of the relative bandwidths, the above equations state that the loaded (matched) bandwidth is twice the unloaded bandwidth; the $VSWR = 2$ bandwidth is $1/\sqrt{2}$ of the unloaded bandwidth; and the $\Gamma = -10\,\mathrm{dB}$ bandwidth is 2/3 of the unloaded bandwidth.

For finding the relative bandwidth for other values of the *VSWR* for a single resonance, the following formula is used:

$$\left(\frac{\Delta\omega_{VSWR}}{\omega} \right) = \left(\frac{\Delta\omega_{VSWR=2}}{\omega} \right) \frac{\sqrt{2}(VSWR - 1)}{\sqrt{VSWR}}. \tag{8.2.14}$$

8.2.1.6 Measured Q

The measured Q of an antenna can be deceptive because the measured Q contains both radiation losses and ohmic losses, $Q_{\mathrm{meas}} = 2\omega W/(P_{\mathrm{rad}} + P_{\mathrm{ohmic}})$, or

$$\frac{1}{Q_{\mathrm{meas}}} = \frac{1}{Q_{\mathrm{rad}}} + \frac{1}{Q_{\mathrm{ohmic}}}. \tag{8.2.15}$$

For a particularly lossy antenna, the measured Q may well indicate a large impedance bandwidth. By adding losses, the impedance bandwidth increases, but of course the Q, or the impedance bandwidth, is not well related to the power being transferred to radiation. This is particularly relevant for antennas on handheld terminals, where the lossy human, in close proximity to the antenna, can enhance the measured impedance bandwidth, and decrease the radiated power. For purely antenna losses, the power dissipated in the metallic parts should be calculated from the surface current estimate and the surface resistivity. For dielectric parts, such as in patch antennas, the dielectric ohmic losses are calculated from the electric field estimates and the imaginary part of the dielectric constant.

Table 8.9 Comparison of the Q of some basic small antennas, where a is the radius of the inscribing sphere. The patch and Goubau antennas are above the groundplane, and the size of the groundplane is not included in the ka dimension. See Figure 8.2.2 for a plot of the values

	Antenna	ka	Q	Q_{theory}	Q/Q_{theory}
a	dipole, $L/\lambda = 0.2$	0.63	51.2	5.58	9.17
b	spherical helix, $D_s/\lambda = 0.38$	0.44	17.3	13.9	1.25
c	patch, $D_p/\lambda = 0.19$, $h/\lambda = 0.63$	0.58	10.6	6.84	1.55
d	low profile dual layer patch $D_p/\lambda = 0.26$, $h/\lambda = 0.043$	0.83	8.43	2.95	2.85
e	Goubau, $D_G/\lambda = 0.28$, $h/\lambda = 0.05$	0.87	5.18	2.64	1.96
f	circular polarised patch, $\epsilon = 10$	0.58	29	4.28	6.76

(a) Dipole length/diameter $L/D_w = 50$, length/wavelength $L/\lambda = 0.2$; calculated using method of moments.
(b) Spherical helix with four wires, sphere radius a, wire diameter $D_w = a/12$, pitch angle 0.3 (16.7°); calculated using method of moments.
(c) Patch (circular, radius $R = 10.65$ mm, $\epsilon_r = 1.07 + j0.16$), 10 dB bandwidth 6.3 per cent, $f = 1.9$ GHz, radius $a = 14.61$ mm ($a/\lambda = 0.093$), height $h = 10$ mm ($h/\lambda = 0.063$); measured (Waterhouse, 1998).
(d) Low profile dual layer patch (circular, single resonance), $VSWR = 2$ bandwidth = 8.4 per cent, $f = 882$ MHz, radius $a = 43$ mm ($a/\lambda = 0.13$), height $h = 14.5$ mm ($h/\lambda = 0.043$), measured.
(e) Goubau antenna, $f = 550$ MHz, $VSWR = 2$ bandwidth = 75 MHz, radius $a = 7.58$ cm, ($a/\lambda = 0.14$).
(f) Circular polarised (degenerate mode pair), narrowband patch, $f = 1575$ MHz, $VSWR = 2$ bandwidth = 38 MHz, radius $a = 1.7$ cm, ($a/\lambda = 0.092$); theoretical Q calculated from dual-mode equation; see Section 8.3.

8.2.1.7 Comparisons of small antennas

Some basic small antennas, which are discussed further below, are compared in Table 8.9. These are the dipole, helix, a wideband patch, and the Goubau antenna. In Table 8.9 the Qs are compared with their theoretical single-mode value (see below), although these antennas, except the dipole, probably have a significant amount of both modes. A feature of the table is that the smallest antenna in the context of an enclosing sphere is the spherical helix. However, other antennas are more practicable for typical situations. If a low profile on a groundplane is required for wall- or roof-mounting, for example, then a patch antenna is much more suitable, despite its spherical enclosure being relatively large.

8.2.1.8 Groundplane effects

The patch antennas and the Goubau antenna are antennas above a groundplane, so it is difficult to know exactly what radius of a sphere is required to contain the near fields. As a rule of thumb, the groundplane needs to be about a half-wavelength square

($a_g/\lambda \sim 1/4$ or $ka_g \approx \pi/2$) for the impedance to be essentially independent of the groundplane size. For an antenna that radiates principally along the groundplane direction, such as a quarter-wavelength monopole over a groundplane, the groundplane size needs to be larger than a half-wavelength square for the impedance to be essentially independent of the groundplane size. The circular groundplane with a centrally located monopole is a worst case, and the groundplane needs to be wavelengths in diameter; see Section 8.1.

In principle, the required groundplane should be included in the spherical volume; however, this makes the groundplane antennas electrically large, namely $ka \approx \pi$. However, the patch antenna can nevertheless be built as a back-to-back image pair (shown below), so the groundplane is removed and the radiation space is both hemispheres, and then the spherical extent of the antenna in the table holds.

8.2.2 Basic limitations of small antennas in isolation

The fundamental aspects of electrically small antennas have been investigated in detail since the pioneering work of Wheeler (1947), Chu (1948), Collin and Rothschild (1964), and a re-examination of the topic by McLean (1996). Electrically small is taken to mean $ka \lesssim 1$ or $a/\lambda \lesssim 1/6$, where a is the radius of the spherical volume containing the antenna. Wheeler demonstrated that as the size of an electrically small antenna decreases, the radiation resistance decreases relative to the ohmic losses in the antenna plus matching network, which means a decreasing antenna efficiency. Wheeler also showed that a small antenna could be efficiently matched over a narrow bandwidth, but when attempting to increase the bandwidth, the antenna (plus matching network) efficiency again decreased owing to the losses. Finally, Wheeler also found that ferrite or dielectric loading of an antenna allowed a decrease in size. However, this did not apply in general, and it was Chu who addressed the problem by analysing the fields about a spherical volume containing an arbitrary radiating antenna. Chu derived theoretical limits for the Q and the size of the sphere. Consequently, the size of a small antenna is taken as that of the enclosing sphere. Chu's work showed that a dipole-type radiated field implied the smallest possible Q.

The basic volume of an electrically small antenna is taken as an enclosing sphere about the antenna. Although a sphere is seldom the most practical shape for an antenna, a spherical antenna has the lowest Q for a fixed antenna volume. For outward travelling spherical waves, the radial behaviour is given by the spherical Hankel function of the first kind, denoted

$$h_n^{(1)}(kr) = \mathrm{j}_n(kr) + \mathrm{j}n_n(kr). \tag{8.2.16}$$

When $n > kr$, the nth mode is cut-off, which means heavily attenuated in the far field. The lowest order ($n = 1$) spherical modes are the TM_{01} and TE_{01}, which correspond exactly to an electrical dipole and a magnetic dipole, respectively. (The zero subscript indicates zero azimuthal dependence.) Higher-order spherical modes will have inherently higher Q values and should thus be avoided, if possible, if the aim is to have the largest bandwidth within a certain volume. The cut-off for the higher modes occurs when $kr < 1$, which is the basis for the 'electrically small' criterion, $ka \lesssim 1$ above.

For an infinitesimal electric dipole in the z-direction, the non-zero fields are

$$\begin{aligned} H_{\phi}^{(TM)} &= \sin\theta \left(1 - \frac{\mathrm{j}}{kr}\right) \frac{\mathrm{e}^{-\mathrm{j}kr}}{r} \\ E_{\theta}^{(TM)} &= Z_0 \sin\theta \left(1 - \frac{\mathrm{j}}{kr} - \frac{1}{(kr)^2}\right) \frac{\mathrm{e}^{-\mathrm{j}kr}}{r} \\ E_{r}^{(TM)} &= Z_0\, 2\cos\theta \left(-\frac{\mathrm{j}}{kr} - \frac{1}{(kr)^2}\right) \frac{\mathrm{e}^{-\mathrm{j}kr}}{r} \end{aligned} \tag{8.2.17}$$

and for a magnetic dipole, the electric and magnetic fields are interchanged;

$$\begin{aligned} E_{\phi}^{(TE)} &= \sin\theta \left(1 - \frac{\mathrm{j}}{kr}\right) \frac{\mathrm{e}^{-\mathrm{j}kr}}{r} \\ H_{\theta}^{(TE)} &= Z_0\ \sin\theta \left(1 - \frac{\mathrm{j}}{kr} - \frac{1}{(kr)^2}\right) \frac{\mathrm{e}^{-\mathrm{j}kr}}{r} \\ H_{r}^{(TE)} &= Z_0\, 2\cos\theta \left(-\frac{\mathrm{j}}{kr} - \frac{1}{(kr)^2}\right) \frac{\mathrm{e}^{-\mathrm{j}kr}}{r}. \end{aligned} \tag{8.2.18}$$

Here, Z_0 is the free-space impedance and the expressions have a normalisation which gives the simple far-field expressions, here for just the electric dipole case,

$$\begin{aligned} E_{\theta}^{(TE)} &= \eta\ \sin\theta \\ H_{\phi}^{(TE)} &= \sin\theta. \end{aligned} \tag{8.2.19}$$

Note that close to the antenna (small kr), the electric dipole has mostly electric field energy, much of which is stored, and for the magnetic dipole, the stored energy is mainly in the magnetic fields.

The theoretical Q value for an antenna (or circuit) is defined as (cf. equation (8.2.4))

$$Q = Q_{\mathrm{th}} = \begin{cases} 2\omega\, W_{\mathrm{e}}/P & W_{\mathrm{e}} > W_{\mathrm{m}} \\ 2\omega\, W_{\mathrm{m}}/P & W_{\mathrm{m}} > W_{\mathrm{e}}. \end{cases} \tag{8.2.20}$$

As noted above, when one of the stored energies is larger than the other, some of the other is added, using a matching arrangement, to make the antenna resonant. The total stored energy is not normally of interest. It is customary to define the minimum Q of an antenna as one where all the stored energy is outside a given radius a. It seems impossible to design an antenna with no internal stored energy, so the above Q value should be considered a theoretical lower bound.

For the dipolar fields, the stored energy, or non-propagating energy, is found by subtracting the radiated energy density from the total stored energy density and integrating the result over all space outside a sphere of radius a. The result is (McLean, 1996)

$$Q = Q_{TM} = \frac{1}{(ka)^3} + \frac{1}{ka}. \tag{8.2.21}$$

The result for a magnetic dipole is exactly the same since the fields are dual.

8.2.2.1 Combined TE and TM modes

It is of interest to combine the *TM* and the *TE* mode into one system and evaluate the Q. If it is assumed that the modes are excited so that the far field magnitudes are the same, then the radiated power is doubled since the modes are independent. However, the stored energy is only increased slightly since the electric energy of the *TE* mode is smaller than that of the *TM* mode, and the result is (McLean, 1996)

$$Q = Q_{TM+TE} = \frac{1}{2(ka)^3} + \frac{1}{ka}. \tag{8.2.22}$$

For ka small, the Q for the double mode system is roughly halved relative to a single mode antenna. This is because of the nature of the definition of Q (equation (8.2.4)) and that the TE_{01} mode stores mainly magnetic energy in the non-radiating fields, but stores some electric energy as well, and vice versa for the TM_{01} mode. So adding the second mode adds only a small amount of the larger of the stored electric energy or the stored magnetic energy.

Note that only powers of the modes have been added, so the relative phase between the two modes is unimportant. This means that the polarisation is arbitrary. It has been customary to think of the *TM* (or *TE*) case as linearly polarised, and the joint $TM+TE$ as circularly polarised, and although this could well be the case, it is not necessary. It can be shown that two orthogonal dipoles, both electric or both magnetic, may be combined to give circular polarisation, at least in some directions. However, such an antenna would still have a Q given by equation (8.2.21).

8.2.2.2 Combined mode antenna synthesis

The synthesis problem of how to design an antenna so it comes as close as possible to the minimum value of Q, given by equation (8.2.20), is an interesting exercise in itself. One solution would be to use metallic wires (no magnetic sources) with current running at, or close to, the spherical surface. This excites both the *TE* mode and the *TM* mode, and also occupies the spherical volume, or rather the surface, efficiently. Using the electric dipole fields, the ratio of the magnetic far fields is proportional to the surface current densities, here expressed as ratios,

$$\begin{aligned} \frac{J_\phi}{J_\theta} = \frac{H_\theta^{(TE)}}{H_\phi^{(TM)}} &= \alpha \, \frac{1 - \mathrm{j}/ka - 1/(ka)^2}{1 - \mathrm{j}/ka} \\ &\approx -\alpha \frac{\mathrm{j}}{ka} \quad (\text{small } ka). \end{aligned} \tag{8.2.23}$$

In order to have a real ratio between the two magnetic fields at the surface of the sphere, the proportion of the far field components (and the current density components), α, must have unit magnitude. The wires on the sphere will have a pitch angle ν given by $\tan \nu = |J_\theta / J_\phi|$, and this describes a spherical helix. For a small sphere (ka small), there would be many turns for realising dense wiring on the surface. An example is included in Table 8.9 and Figure 8.2.2.

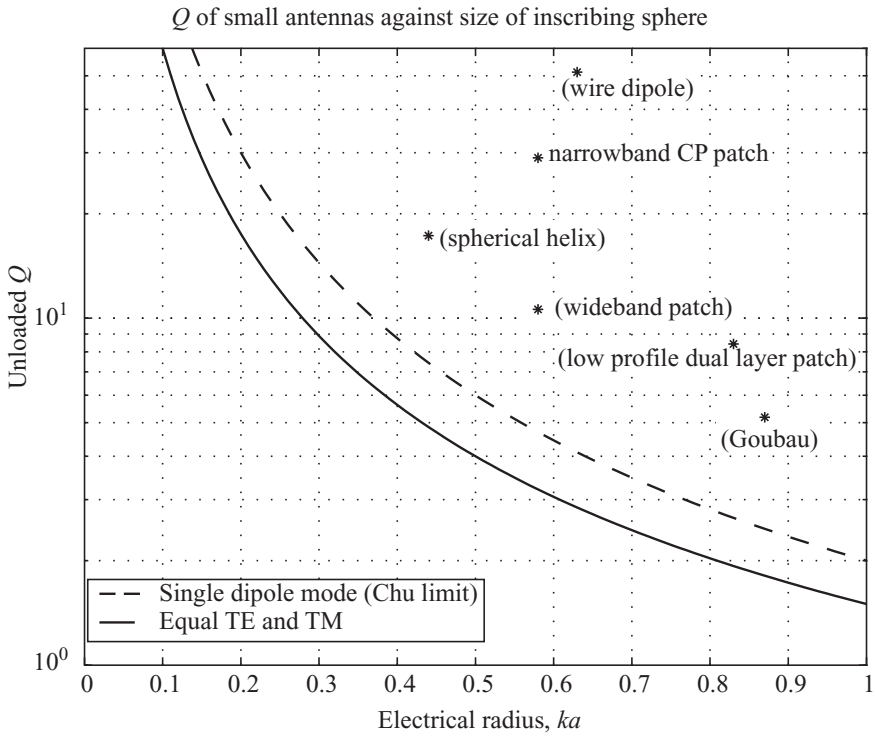


Figure 8.2.2 Minimum unloaded Q of small antenna consisting of a dipole mode (electric or magnetic) and of the two modes in equal excitation as a function of the radius of the surrounding sphere. The wideband patch is from Waterhouse (1998)

From Figure 8.2.2, it is apparently difficult to get below the one-mode curve for real antennas. The spherical helix comes within 25 per cent of the theoretical one-mode curve, although this belongs more appropriately to the two-mode family. A spherical shape is not normally practical to build, and any deviations from it will increase the Q. Also, any ohmic losses will lead to decreased Q-values.

The spherical helix antennas give a benchmark for antenna compactness, but the spherical volume metric does not provide the whole story for a particular application for compact antennas, as noted above.

8.2.3 Bandwidth enhancement of small antennas

The Q values of small antennas described in the previous section are related to the intrinsic lower bound, or Chu limit, connected with the unavoidable external, reactive fields. The bandwidths (such as for $VSWR < 2$) are the direct ones assuming a matched source or load connected to the antenna. But the bandwidth can be enhanced, to be larger than the Chu limit, by applying a more complicated matching network. The bandwidth may still be limited by the intrinsic Q, but it can be larger than that for the simple, single resonance circuit.

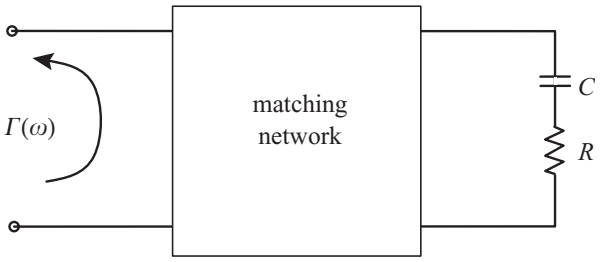


Figure 8.2.3 A lossless matching circuit for the series RC network representing a single resonance antenna. The matching circuit can be designed to introduce extra resonances which increase the impedance bandwidth, and these may be incorporated in the antenna structure

The technique may be borrowed from microwave filter theory (Matthaei *et al.*, 1964; Pozar, 1998) for designing bandpass filters with a given ripple. Figure 8.2.3 shows the arrangement.

Depending on the number of idealised filter sections applied, the bandwidth approaches a limit for an infinitely complex filter. This limit may be described by the Fano (1950) theory, which gives a theoretical limit on the minimum reflection coefficient for a given canonical load impedance. In practice, we can only approximate the performance, but measure how well relative to the theoretical limit. For a series RC circuit for the load, the Fano criterion states that the integral of the log of the reflection coefficient over the frequency range is limited, given by

$$\int_0^\infty \ln\left(\frac{1}{|\Gamma(\omega)|}\right) \mathrm{d}\omega < \pi\omega_0^2 RC, \tag{8.2.24}$$

where ω_0 is the centre frequency of the bandwidth. The right-hand side can be modified appropriately for other load circuit models, such as LR and for the parallel circuit arrangements.

Typically, a maximum value of reflection coefficient, Γ_m, is specified for a finite bandwidth $\Delta\omega$, and the Fano criterion reduces to $\ln(1/\Gamma_\mathrm{m}) \leq \Delta\omega\pi/(RC)$. This shows that the reflection coefficient can be zero only over an infinitesimally small bandwidth (cw case), although this can occur at multiple frequencies with multiple sections in the matching circuit; the bandwidth and the reflection coefficient (or rather $1/\Gamma_\mathrm{m}$) must be traded off with each other for a given RC load; and the magnitude of the load, RC, trades off with either the bandwidth or the reflection coefficient (or rather $1/\Gamma_\mathrm{m}$).

The rippling of the VSWR (see Figure 8.2.4) corresponds to the introduction of additional filter resonances, and it is clear that if these additional resonances could be integrated in the antenna, and the currents be radiating instead of just being internal currents leading to extra losses, some additional benefit has been achieved. A common application is to have multiple, resonant-type patches stacked on top of each other.

An example is given in Figure 8.2.4, where two additional resonances have been added externally to the antenna, which has a Q of 20. The $VSWR = 2$ bandwidth has been more than doubled.

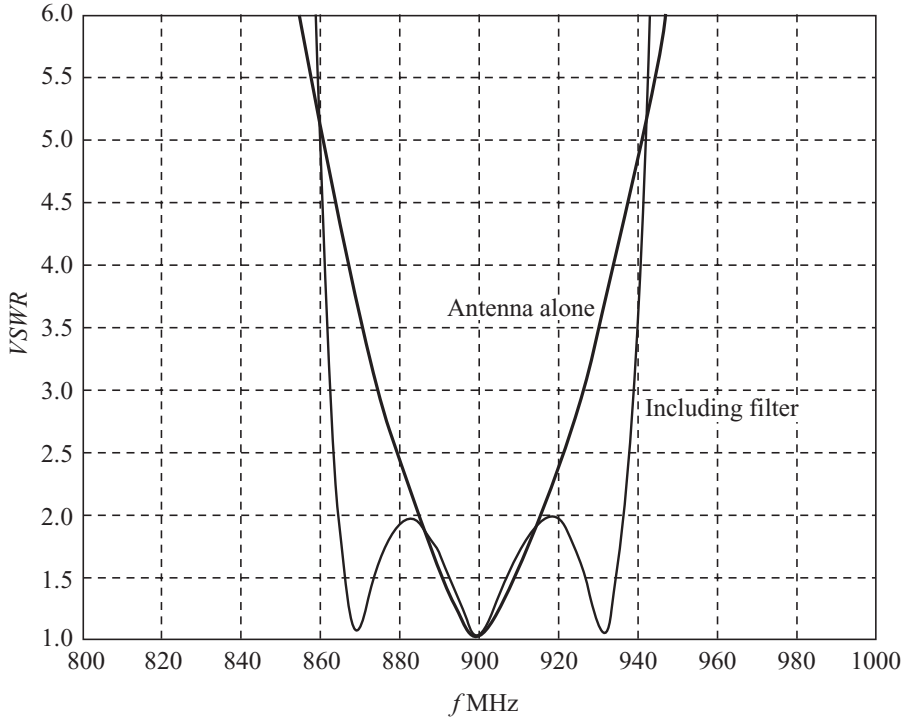


Figure 8.2.4 *A single resonance with a Q of 20 has an increased bandwidth at VSWR = 2 of a factor of 2.25 by adding additional resonances in a matching filter*

Table 8.10 *Maximum bandwidth improvement factors resulting from the addition of matching circuits, including parasitic patch elements*

Number of additional matching circuits	$VSWR = 2$ bandwidth improvement factor	Half-power bandwidth improvement factor
1	2.3	2.0
2	2.8	2.4
3	3.1	2.7
4	3.3	2.8
∞	3.8	3.2

After Hansen (1998)

Such bandwidth improvement factors are given in Table 8.10, from Hansen (1998). The above example with two additional resonances gives a $VSWR = 2$ bandwidth improvement of about 2.3, whereas the table suggests a factor of 2.8. An example further below shows a bandwidth increase factor of about 3 to 4 with a single additional antenna element, so in practice the result from the theory of resonant circuits is not the end of the story for antenna bandwidth.

A final point with the matching circuit theory is that the classical antenna efficiency factor does not include the losses introduced by a matching circuit, as discussed in

Section 8.1. However, most matching circuits will impose a loss which is significant at microwave frequencies.

8.2.3.1 Discrete component basic matching circuit

The lossless matching circuit can be arranged by stub matching or by discrete components. In general, both the real and the imaginary part need to be adjusted independently, so two independent variables are required. An 'L' section comprising an inductor and capacitor represents a single matching section. The circuits and their region of applicability are shown in Figure 8.2.5. Normally, lumped components can be used conveniently at frequencies less than 1 GHz or so.

The values for the matching circuit, with the usual notation load Z_L and required characteristic impedance Z_0 (different from the intrinsic free space impedance, and usually 50 Ω), for $R_L \neq Z_0$, are readily found as (cf. Pozar, 1998)

Case $R_L > Z_0$

$$B = \frac{X_L \pm \sqrt{R_L/Z_0}\sqrt{R_L^2 + X_L^2 - Z_0 R_L}}{R_L^2 + X_L^2} \tag{8.2.25}$$

$$X = \frac{1}{B} + \frac{X_L Z_0}{R_L} - \frac{Z_0}{B R_L}. \tag{8.2.26}$$

Case $R_L < Z_0$

$$X = \pm\sqrt{R_L(Z_0 - R_L)} - X_L \tag{8.2.27}$$

$$B = \pm\sqrt{\frac{(Z_0 - R_L)/R_L}{Z_0}}. \tag{8.2.28}$$

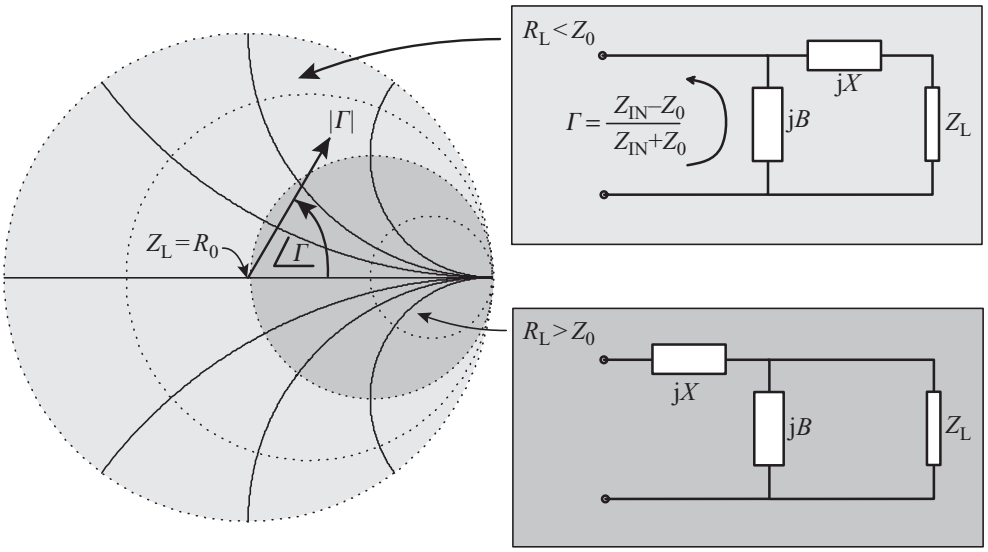


Figure 8.2.5 Basic matching circuit for $R_L \neq Z_0$*, and the regions on the Smith chart for each circuit type*

The multiple solutions give different frequency responses away from the centre frequency but results in the same bandwidth. One solution may have more practicable values for realising the circuit than the other.

For multiple matching circuits for increasing the bandwidth, Chebyshev solutions are normally used, which give the maximum bandwidth for a given maximum reflection coefficient. If required, a maximally flat solution can be obtained by using binomial solutions.

8.2.3.2 Transmission line transformers

If the load is purely real, a transmission line transformer can be used. These are compact physically at microwave frequencies, but may be bulky in the sense of compact antennas, in particular for multiple section realisations. The simplest line transformer is the quarterwave section of constant impedance line. The characteristic impedance of the line section, Z_l, is the geometric mean of the load and the desired value, i.e. $Z_l = \sqrt{Z_\mathrm{L} Z_0}$ and the length is a quarter wavelength at the centre frequency. The frequency response near the frequency given by the line length being a quarterwavelength, $l = \lambda_\mathrm{g}/4$, is

$$\Gamma_\mathrm{IN} = \Gamma(k_\mathrm{g}) \approx \frac{|Z_\mathrm{L} - Z_0|}{2\sqrt{Z_0 Z_\mathrm{L}}} \, |\cot k_k l|. \tag{8.2.29}$$

8.2.3.3 Multiple quarterwave sections

Multiple quarterwave transformers can be used to increase the bandwidth (Slater, 1942), but acts to increase the bandwidth of the match, but uses more space of course. With multiple transmission line sections, each of length l, propagation constant k_g, and impedance Z_n, then the input refection coefficient is

$$\Gamma_\mathrm{IN} = \sum_{n=0}^{N} \Gamma_n \, \mathrm{e}^{-\mathrm{j}2nk_\mathrm{g}l} \tag{8.2.30}$$

where for each section, $\Gamma_n = (Z_{n+1} - Z_n)/(Z_{n+1} + Z_n)$. Only N of the $N + 1$ impedances can be chosen independently since in the small length limit $(l = 0)$, $\Gamma_\mathrm{IN} = (Z_\mathrm{L} - Z_0)/(Z_\mathrm{L} + Z_0)$.

For a maximally flat type frequency response, the reflection coefficients of each section relative to that at the input follow a binomial pattern,

$$\frac{\Gamma_n}{\Gamma_\mathrm{IN}} = \frac{N!}{(N-n)!n!}. \tag{8.2.31}$$

The intermediate characteristic impedances follow as the binomial coefficients. For example, for three sections, transforming Z_L to Z_0, the binomial coefficients are {1,3,3,1}, and the line impedances are found from

$$\ln\left(\frac{Z_1}{Z_\mathrm{IN}}\right) = 3 \, \ln\left(\frac{Z_2}{Z_1}\right) = 3 \, \ln\left(\frac{Z_3}{Z_2}\right) = \ln\left(\frac{Z_3}{Z_L}\right). \tag{8.2.32}$$

The frequency response is given by the reflection coefficient as a function of k_g,

$$\Gamma(k_g l) = \left(\frac{Z_L - Z_0}{Z_L + Z_0}\right) |\cos k_g l|^N, \tag{8.2.33}$$

which has a maximum at the band edges ($k_g l = 0, \pi$) of $(Z_L - Z_0)/(Z_L + Z_0)$. The minimum occurs for $l = \lambda_g/4$, and the bandwidth for a given reflection coefficient can be found directly from the frequency response equation.

Many other distributions for the multiple sections can be synthesised for a given type of required frequency response; see for example, Elliott (1993).

8.2.3.4 Tapered section

Tapered transmission lines offer the continuous form of the multi-sectioned transmission lines. The total reflection, seen at the input, is the summation along the length of the line (cf. the summation of discrete reflection coefficients in the multiple sections, above), and can be approximated from small-reflection theory as a Fourier integral form (Klopfenstein, 1956). So for a given impedance change function along the tapered line, $Z(l)$, the reflection coefficient can be calculated and vice versa,

$$\Gamma_{IN} = \Gamma(2k_g) = \int_{-L/2}^{L/2} \left(\frac{1}{2}\frac{d}{dl} \ln Z(l)\right) e^{-j2k_g l}\, dl \tag{8.2.34}$$

$$\frac{1}{2}\frac{d}{dl} \ln Z(l) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Gamma(2k_g)\, e^{j2k_g l}\, d(2k_g). \tag{8.2.35}$$

Here k_g is taken as independent of the position, which holds for *TEM* modes, and here l is the length along the line from $-L/2$ to $L/2$.

An exponential taper is the simplest design, comprising $Z(l) = Z_0 e^{al}$, and this results in $\Gamma(k_g) = 1/2\, \ln(Z_L/Z_0)\, e^{-jk_g L}\, \sin(k_g L)/(k_g L)$.

The Klopfenstein taper is the continuous equivalent of an infinite number of Chebyshev sections, and gives the lowest reflection coefficient for a given bandwidth, for a fixed length of tapered line, L. The reflection coefficient looking into the matching section is defined from the transform of the Dolph–Chebyshev taper, and this is the frequency response,

$$\begin{aligned} \Gamma_{IN}(k_g) &= \Gamma_0\, e^{-jk_g L} \frac{\cos\sqrt{(k_g L)^2 - A^2}}{\cosh A}, \quad k_g L > A, \\ \Gamma_{IN}(k_g) &= \Gamma_0\, e^{-jk_g L} \frac{\cos\sqrt{A^2 - (k_g L)^2}}{\cosh A}, \quad k_g L < A, \end{aligned} \tag{8.2.36}$$

and this, like the Chebyshev sections, is constant ripple. Here, the reflection coefficient at the centre frequency is $\Gamma_0 = (Z_L - Z_0)/(Z_L + Z_0) \approx 0.5 \ln(Z_L/Z_0)$. The parameter A controls the maximum value of the reflection coefficient, Γ_0, through $\Gamma_0 = \Gamma_m \cosh A$ and the passband is given by $\beta L \geq A$, which can be considered as defining the required length.

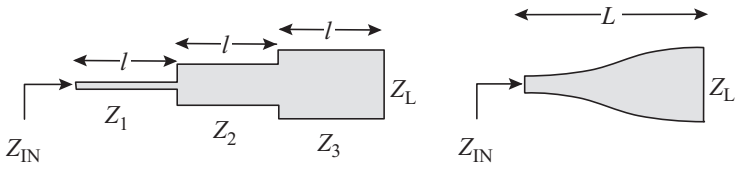


Figure 8.2.6 Stripline profiles for a multi-section line and a tapered line

The transform for transmission line, length $-L/2$ to $L/2$, has an impedance taper

$$\begin{aligned}\ln Z(l) &= \frac{1}{2}\ln(Z_0 Z_L) + \frac{\Gamma_0}{\cosh A}\, A^2 \int_0^{2l/L} \frac{I_1\left(A\sqrt{1-x^2}\right)}{A\sqrt{1-x^2}}\,\mathrm{d}x \\ &= \ln Z_L \quad \text{for } l > L/2; \\ &= \ln Z_0 \quad \text{for } l < L/2. \end{aligned} \tag{8.2.37}$$

The integral requires straightforward numerical solution. The impedance taper function is not truly a continuous taper, with discontinuities (relative to the load and source impedances) at each end. Using the approximation $\Gamma_0 \approx 0.5\ln(Z_L/Z_0)$, as seen in the above equation, improves this situation. A normal procedure is: given Z_0, Z_L, L, Γ_L/Γ_0, check the reflection coefficient; then calculate Γ_0 and then the impedance taper to be synthesised using stripline, for example. Figure 8.2.6 shows stripline representation of a multi-section line and a tapered line.

8.2.4 Goubau antenna

The Goubau antenna (Goubau, 1976) is basically a top-loaded monopole on a groundplane with extra feed complexity which allows a decrease in its height while maintaining a match to 50 Ω. Although its construction details make it complicated to build, its development from a simple top-loaded monopole introduces illuminating concepts and paves the way for the feed design of more practicable patch antennas.

The starting point is a capacitive disc on a monopole mounted on a groundplane, shown in Figure 8.2.7. The idea is to reduce further the length of the monopole, making the antenna profile lower and in terms of the inscribing sphere, more compact than a quarter-wave monopole.

The monopole element and its disc are now split symmetrically into N common sub-monopoles with disc segments as shown in Figure 8.2.8 for $N = 4$. The symmetry maintains purely vertical polarisation which stems from the total current in the monopole and symmetric currents (whose far field contributions cancel) on the disc segments. Strictly speaking, this cancellation occurs only if the antenna is infinitesimally small, and the Goubau is not particularly small electrically (see Figure 8.2.2). The far fields cancel only in the $\theta = 0$ direction. Figure 8.2.9 below shows the inclusion of wire connections between the sub-monopoles at the disc. Goubau notes

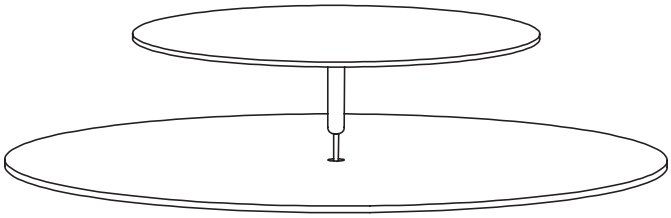


Figure 8.2.7 The top-loaded monopole on a groundplane with coaxial feed

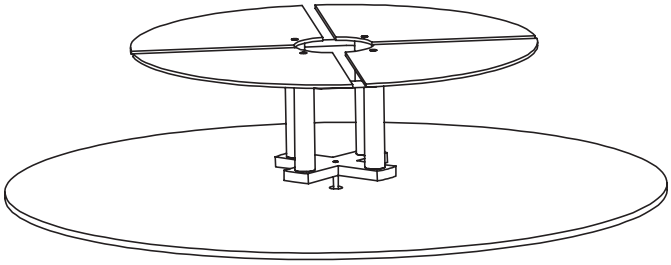


Figure 8.2.8 The segmented monopole and disc on a groundplane with coaxial feed common to the sub-monopoles

that these are for inductive coupling between the sub-monopoles at the disc and do not affect the electrical (radiation) characteristics. Again, this is not strictly true in practice.

The sub-monopoles are now considered to have their own excitation at the groundplane. The currents $\boldsymbol{I}_{\mathbf{S}} = [I_1,\ I_2,\ \ldots,\ I_N]^T$ and voltages, $\boldsymbol{V}_{\mathbf{S}} = [V_1,\ V_2,\ \ldots,\ V_N]^T$, of the N sub-monopoles are related by the admittance matrix, $\boldsymbol{Y}$, as seen from the feeds, i.e.

$$\boldsymbol{I}_{\mathbf{S}} = \boldsymbol{Y}\boldsymbol{V}_{\mathbf{S}}. \tag{8.2.38}$$

The total current on the monopole (cf. Figure 8.2.7) is the sum of the sub-monopole currents,

$$I_{\mathrm{T}} = \sum I_n = \sum_{n=1}^{N}\sum_{m=1}^{N} Y_{nm} V_m. \tag{8.2.39}$$

From the symmetry, the sum of admittances, Y_N, is common for each sub-monopole, namely

$$\sum_{n=1}^{N} Y_{nm} = \sum_{m=1}^{N} Y_{mn} = Y_N \tag{8.2.40}$$

and so the total current is

$$I_{\mathrm{T}} = Y_N \sum_{m=1}^{N} V_m. \tag{8.2.41}$$

For the case shown in Figure 8.2.8, where all the sub-monopoles are connected at the feed end and so have the same voltage V_{mono}, we have

$$I_{\text{T}} = NY_N V = Y_{\text{mono}} V_{\text{mono}}, \tag{8.2.42}$$

where Y_{mono} is the input admittance of the assembly of Figure 8.2.8.

The case where all the sub-monopoles are grounded except for one, which has excitation $V_1 = V_{\text{mono}}$, is now considered.

The coupling inductance, realised using wire connections between the tops of the sub-monopoles (see Figure 8.2.9), is now used to arrange the currents. The details are as follows: each voltage excitation is considered as a combination of common and differential components, here for the voltage,

$$V_m = V'_m + V''_m. \tag{8.2.43}$$

The common voltage is

$$V'_1 = V'_2 = \cdots = V'_N = V' \tag{8.2.44}$$

and so the differential components are

$$\begin{aligned} V''_m &= -V', \quad m \neq 1; \\ V''_1 &= -\sum_{m=2}^{N} V''_m \\ &= (N-1)V'; \end{aligned} \tag{8.2.45}$$

and

$$V_1 = V'_1 + V''_1 = NV' = V. \tag{8.2.46}$$

The currents are denoted the same way with $I_m = I'_m + I''_m$ and $I'_m = I/N$ independent of the coupling inductance. The I'_m differ (between the m different sub-monopoles) and are dependent on the coupling inductance. The coupling inductance is now chosen so that $I''_1 = 0$ at some operating frequency. Using the configuration for the single monopole (equation (8.2.42)), results in $I'_m = Y_{\text{mono}} V'/N = Y_{\text{mono}} V/N^2$.

So using the coupling inductances to arrange the currents in this way, the input current can be shown to be

$$I_1 = \frac{I_{\text{T}}}{N} = \frac{V_{\text{mono}} Y_{\text{mono}}}{N^2}. \tag{8.2.47}$$

The input impedance at the excited sub-monopole is

$$Z_1 = \frac{V_1}{I_1} = \frac{V_1 N^2}{Y_{\text{mono}} V_{\text{mono}}} = \frac{N^2}{Y_{\text{mono}}}, \tag{8.2.48}$$

i.e. this is seen to be N^2 times the input impedance of the common monopole structure, $1/Y_{\text{mono}}$. The effect is discussed in Section 8.1, and can be seen, for example, in

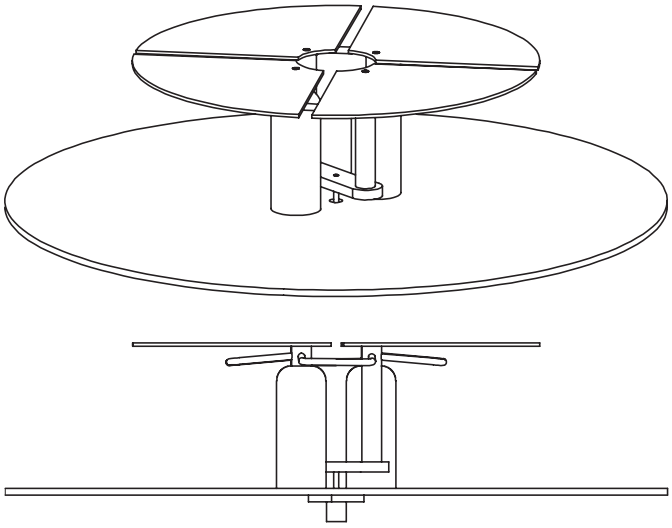


Figure 8.2.9 The Goubau antenna with non-uniform sub-monopoles for very wide bandwidth. The upper diagram shows the inductive coupling wires

a folded monopole on a groundplane whose impedance (about 150 Ω) is some four times that of a quarter-wave monopole on a groundplane (about 36 Ω). Alternatively stated, the effective height of the antenna (proportional to the voltage, see Section 8.1 and Chapter 2) is reduced by a factor of N. So the height of the antenna can be reduced for the resistance to suit a 50 Ω match, but the reactance needs to be tuned out, and this is achieved using inductive coupling wires, shown in Figure 8.2.9.

Goubau's (1976) example has a height of $h/\lambda = 0.05$ and the disk radius is $r/\lambda = 0.115$, so the radius of the inscribing sphere is $a = \sqrt{r^2 + h^2}$, which gives $k_C a = 0.87$.

Goubau also gives a more complex design using sub-monopoles of differing thickness, shown in Figure 8.2.9. This has very wide bandwidth, but does not qualify as electrically compact ($ka \lesssim 1$) over the upper frequencies of its bandwidth. The arrangement maintains purely vertical polarisation by suppressing dipolar currents in the disc segments and this is done by having two planes of symmetry. The structure is effectively double resonant, having two intertwined symmetric structures: one with the thinner sub-monopoles, and one with the thicker sub-monopoles. This antenna has $k_C a = 0.866$, $h/\lambda = 0.1$, and a $VSWR = 2$ bandwidth of 1 : 1.9. If Q is expressed in terms of the centre frequency (recall that Q is strictly a narrowband parameter) the result is $Q = 1/\sqrt{2}\,(f_c/\Delta f_{VSWR=2}) \approx 1.13$. At the lowest frequency in the band, $k_l a = 0.73$ and at the uppermost frequency, $k_u a = 1.4$. The extraordinary bandwidth of this antenna is shown as a measurement in Figure 8.2.10, quoted from Goubau (1976).

In summary, the Goubau antenna demonstrates how the height of a basic monopole can be reduced. A constraint for the antenna is that it maintains vertical polarisation, but for compact antennas for mobile communications, this polarisation purity is seldom an issue and so the symmetry of the structure can be relaxed. The feeding arrangement also offers some hints of how to feed an asymmetric antenna, and this is seen below with the low-profile patch antenna example.

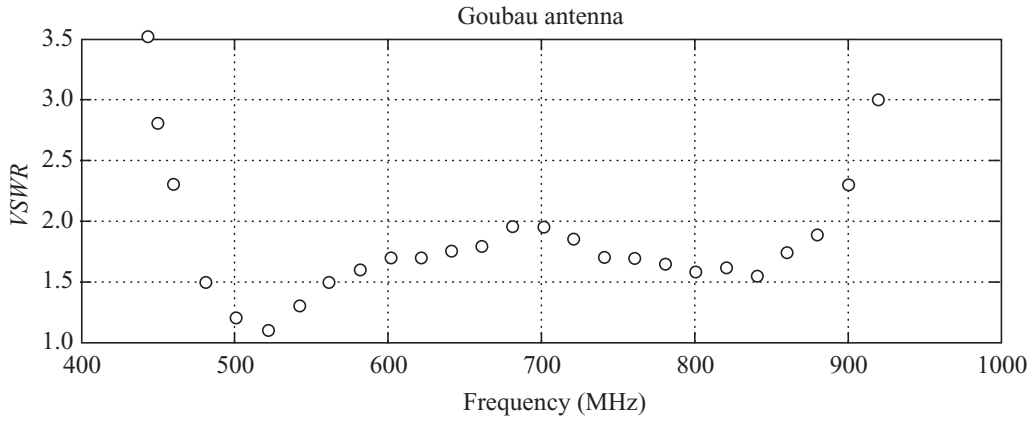


Figure 8.2.10 The measured, wideband impedance of the Goubau antenna (Goubau, 1976)

8.2.5 Patch antennas

Patch antennas are practical and popular owing to their ease of manufacturing together with other circuit elements. The PIFA (Planar Inverted-F Antenna) and the square patch are reviewed here. The circular patch is discussed in Section 8.3. There seems little advantage in using shapes other than these, although for circular polarisation, perturbations from these basic shapes are used. The PIFA offers a very compact, in the low profile sense, radiator which is capable of bandwidths covering typical cellular system bands. The square and circular patches are easier to build, not normally requiring a short circuit along a wall, as in the PIFA.

In developing a patch antenna (or most antennas for that matter) to given frequency, bandwidth or pattern requirement, it is normally necessary to iterate the design experimentally. For patch antennas, a first cut design is based on the basic cavity principles. The basic formulas do not account for the fringing fields of the slots at the radiating edges of the patch. The fringing fields act to make the patch effectively larger, or alternatively stated, lower the frequency of operation relative to that expected from the basic physical dimensions of the patch and its basic cavity model. Fringing formulas are available which provide a correction, typically in the order of a few percent of the operating frequency. However, the practical constraints found with most antenna situations, such as finite groundplane, substrates with a variable permittivity, metals and plastics of varying types, and manufacturing inaccuracies, combined with the many assumptions in the electromagnetic analysis, mean that the first cut design seldom works exactly as expected, even with the fringing formulas. When the antenna specification is pushing the limits of what is possible, which is a typical trait of trading off parameters such as antenna compactness with acceptable bandwidths, experimental iteration is necessary. This means building and measuring a design, measuring and scaling it to its required operating frequency, and trying again. So basic principles only are presented here, although some designs are detailed.

8.2.5.1 PIFA

The PIFA is one of the most compact patch-type antennas. One element often used on handsets is the so-called PIFA (Planar Inverted-F Antenna), originally called an open microstrip antenna (James *et al*., 1989), shown in Figure 8.2.11. This is basically a half of the usual square patch antenna. It has the advantage of being more compact than the square patch and has very good bandwidth capability. It is not compact in the spherical volume sense, but it occupies a profile.

The PIFA has height h and width W, and for a full short circuit across the width, the length is approximately a quarter of a wavelength in the dielectric material between the patch and the groundplane. The short circuit across one of the slots of the patch is a manufacturing limitation since the design is no longer a purely planar patch. The PIFA is normally coaxially fed from the groundplane.

The radiation is mainly in the broadside direction away from the patch, dominated by the dipole-like fields of the end slot. The size of the groundplane is therefore much less critical that for the monopole antenna example quoted in Section 8.1, where the radiation is mainly along the groundplane direction.

8.2.5.2 Cavity model for radiation

The fields under the patch have a field distribution which is modelled by the parallel plate waveguide of the groundplane and patch. Using the coordinate system shown in

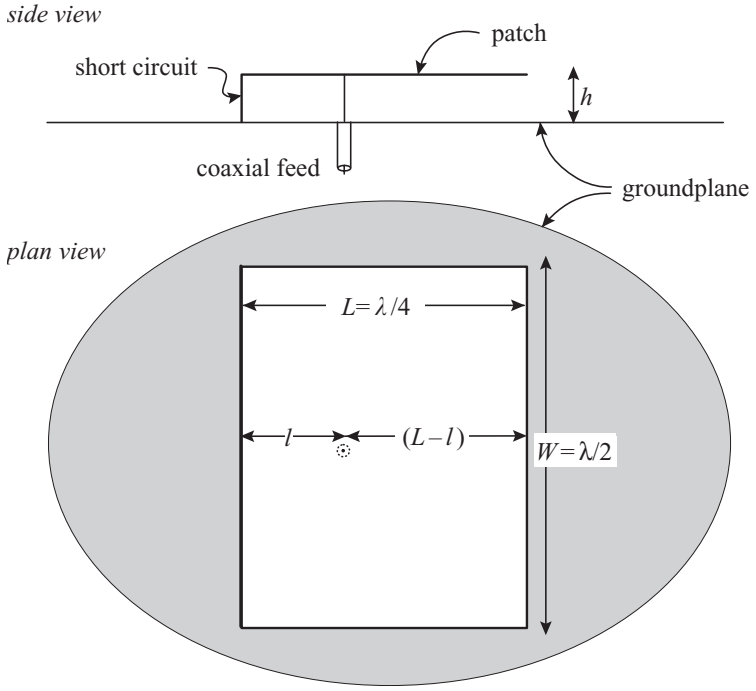


Figure 8.2.11 Side and plan view of a PIFA of height h. The substrate is air in this figure but the patch can be built with a dielectric of permittivity ϵ. Typically, the size is a quarter by a half of the internal wavelength

Figure 8.2.12, the electric fields under the patch are assumed to follow the low-height cavity distribution

$$E_y = \hat{y}\, V \cos\left(k\frac{z}{L}\right), \tag{8.2.49}$$

and so the fields are zero along the short circuit, uniform along the end slot, and sinusoidal along the edge slots. V is the voltage at the edge of the end slot and $k = k_0\sqrt{\epsilon}$ is the wave number in the patch substrate which has relative permittivity ϵ. The equivalent linear magnetic currents follow as:

$$\begin{aligned}
&\{1\} \quad I_x^{(\mathrm{m})} = 2V\hat{\boldsymbol{x}}, && x = -W/2 \text{ to } W/2,\ z = 0,\ y = 0;\\
&\{2\} \quad I_z^{(\mathrm{m})} = 2V\cos\left(2\pi\frac{z}{L}\right)\hat{\boldsymbol{z}}, && z = 0 \text{ to } L,\ x = W/2,\ y = 0;\\
&\{3\} \quad I_z^{(\mathrm{m})} = -2V\cos\left(2\pi\frac{z}{L}\right)\hat{\boldsymbol{z}}, && z = 0 \text{ to } L,\ x = -W/2,\ y = 0;
\end{aligned} \tag{8.2.50}$$

where the current direction is defined as positive along the positive axis direction. The end slot {1} is the main radiator, but the side slots {2},{3} also contribute. The equivalent magnetic current wire circuit is depicted in Figure 8.2.12.

From the magnetic currents, the far fields can be found from the contributing electric vector potentials. The total electric potential (in $\mathrm{C\,m^{-1}}$) is, using Table 8.3 from Section 8.1,

$$\boldsymbol{F} = \frac{\epsilon_0}{4\pi}\frac{e^{-\mathrm{j}k_0 r}}{r}\int_x\int_y\int_z \boldsymbol{M}\mathrm{e}^{\mathrm{j}k_0(x\sin\theta\cos\phi + y\sin\theta\sin\phi + z\cos\theta)}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z. \tag{8.2.51}$$

For the end slot, {1}, the volumetric magnetic current (in $\mathrm{V\,m^{-2}}$) is $\boldsymbol{M}_{\{1\}} = 2V\delta(y)\delta(z)\hat{\boldsymbol{x}}$ for $-W/2 < x < W/2$. The contribution of the end slot to the total

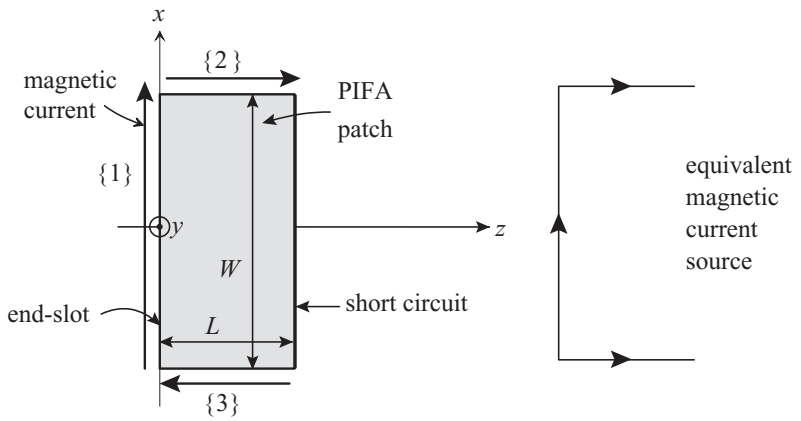


Figure 8.2.12 A plan view of a PIFA in a coordinate system and the magnetic currents of the slots. The end slot {1} is the main radiator although the side slots {2} {3} also contribute to make a radiating magnetic current source {3}{1}{2}. The magnetic current goes to zero at the ends

electric vector potential is

$$\boldsymbol{F}_{\{1\}} = \frac{\epsilon_0}{4\pi} 2V \frac{\mathrm{e}^{-\mathrm{j}k_0 r}}{r} \int_{-W/2}^{W/2} \mathrm{e}^{\mathrm{j}k_0 x \sin\theta\cos\phi}\, \mathrm{d}y\, \hat{\boldsymbol{x}}$$
$$= \frac{\epsilon_0}{2\pi} V \frac{\mathrm{e}^{-\mathrm{j}k_0 r}}{r} W \left(\frac{\sin u_W}{u_W}\right) \hat{\boldsymbol{x}}, \tag{8.2.52}$$

where

$$u_W = \frac{k_0 W}{2} \sin\theta\cos\phi. \tag{8.2.53}$$

Similarly, for the side-slots, $\boldsymbol{M}_{\{2\}} = 2V\cos(\pi z/L)\delta(x - W/2)\delta(y)\hat{\boldsymbol{z}}$ for $0 < z < L$, for example. The two side slots give

$$\boldsymbol{F}_{\{2\}} + \boldsymbol{F}_{\{3\}} = \frac{\epsilon_0}{\pi} V \frac{\mathrm{e}^{-\mathrm{j}k_0 r}}{r} L^2 \left(\frac{(-\mathrm{j}k_0\cos\theta)(1 + e^{\mathrm{j}k_0 L\cos\theta})}{(\pi)^2 - (k_0 L\cos\theta)^2}\right) \cos u_W\, \hat{\boldsymbol{z}}. \tag{8.2.54}$$

These vector potentials can be used to find the fields from the equations of Section 8.1, i.e.

$$\boldsymbol{H} = -\mathrm{j}\omega\boldsymbol{F} - \mathrm{j}\frac{1}{\omega\mu_0\epsilon_0}\nabla(\nabla\cdot\boldsymbol{F}), \quad \boldsymbol{E} = -\frac{1}{\epsilon_0}\nabla\times\boldsymbol{F}. \tag{8.2.55}$$

and the far field requires just the first term for $\boldsymbol{H}$. The far field is dominated by the end slot, which has the (uniform current source) dipole pattern in the half-plane.

8.2.5.3 Feed impedance

The feed point is chosen to give a suitable impedance. A model for finding this is to have transmission lines for the length and width directions, see Figure 8.2.13. The presence of the transverse transmission lines suggests that the parallel plate waveguide will not have a truly uniform field distribution as the fields model above suggests. However, the power flow to the side slots is accounted for using this approximation.

As noted above, the end slot is a uniform voltage source and is modelled with a uniform wire of magnetic current. Equivalence theory is now used to estimate the edge admittances.

From electric wire theory, the (anti-) resonant impedance of the uniform electric current wire a half wavelength long, is $Z_\mathrm{C} = 80\pi^2(L/\lambda)^2 = 20\pi^2$. From Booker's relation, the impedance of the equivalent slot in full spherical space is $Z_\mathrm{SF} = 1/Z_\mathrm{C}(Z_0/2)^2$, and we double this for operating in the half-space above the groundplane, so the end slot impedance is $Z_\mathrm{es} = 2Z_\mathrm{SF} = 360\,\Omega$, and we have $G_\mathrm{es} = Y_\mathrm{es} = 1/Z_\mathrm{es}$.

This value for the end slot admittance can also be found from a direct derivation of the admittance of a uniformly illuminated slot. The directivity integral is plotted in Figure 8.1.26 in Section 8.1. The technique of equating the integral of the Poynting vector giving the radiated power, with the power at the slot terminal, gives

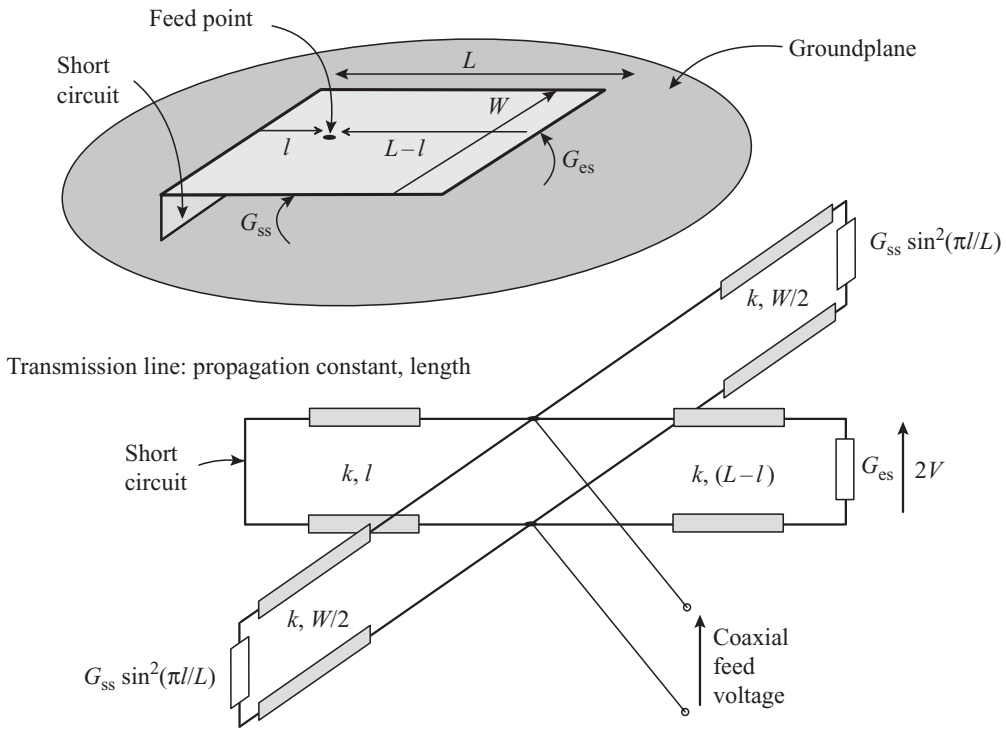


Figure 8.2.13 *The PIFA patch with a transmission line model for power flow to the end slot and side slots. The edge admittances are G_{es} for the end slot, and G_{ss} for each of the edge slots, and combining these into a single line gives an edge admittance of $G_{2ss} = 2G_{ss}$*

$P_{rad} = 1/2 G_{es} V^2$, where V is the slot voltage; see Section 8.1. This gives the general formula for the uniform slot, of

$$G_S^{(U)} = \frac{1}{30\,D}\left(\frac{W}{\lambda}\right)^2 \quad \text{(hemisphere)}, \tag{8.2.56}$$

where D is the maximum directivity. For a small electrical slot length W/λ, the directivity in a hemisphere is $D = 3$, and if $W/\lambda = 1/2$, then $G_{es} = 1/360$ mhos. Using this small-length directivity gives the small-length formula for the admittance of a uniform slot in a half space. Similarly, the asymptotic directivity of an electrically long, uniform line source in a hemisphere, is $D = 4W/\lambda$, and this gives the long-length admittance formula. These limiting admittances are:

$$\begin{aligned} G_S^{(U)} &= \frac{1}{90}\left(\frac{W}{\lambda}\right)^2 \quad \text{(hemisphere)}, \quad W \ll \lambda; \\ G_S^{(U)} &= \frac{1}{120}\left(\frac{W}{\lambda}\right) \quad \text{(hemisphere)}, \quad W \gg \lambda; \end{aligned} \tag{8.2.57}$$

and the small length form gives the correct answer for $W/\lambda = 1/2$.

The conductance of the side slots of the PIFA can be estimated in the same way as above, i.e. from the sinusoidal current distribution of a quarter-wave wire, which at resonance has $Z_C = 36\,\Omega$. Again using Booker's relation for the half-space, and combining the symmetrical transverse transmission lines into a single line, the single impedance at resonance representing both edges is $Z_{2ss} = 4Z_C = 144\,\Omega$. The corresponding conductance is related as $Y_{2ss} = G_{2ss} = 2G_{ss} = 1/Z_{2ss}$ and this value is referred to the point of maximum magnetic current, i.e. the end slot position. This corresponds to the corner of the magnetic current wire.

Mutual coupling between the edge slots has been neglected in this model. For a high-permittivity substrate, the free space spacing between the edge slots reduces significantly, and the mutual coupling may need to be included. This involves a current source in parallel with the edge admittances, with voltage proportional to the slot mutual admittance. Mutual coupling equivalent circuits are discussed in Section 8.4.

The position of the feed point does not affect the radiated power (calculated from the integral of the Poynting vector of the slot), and so the feed point can be placed for an impedance match in the usual way for patch antennas. The feed is placed distance l along the PIFA from the short circuit wall. Note that l is in the negative z direction and measured from $z = L$. The power radiated from each side slot is

$$P_{\text{rad}}^{(\text{ss})} = \tfrac{1}{2}G_{\text{ss}}|V|^2 = \tfrac{1}{2}G(l)|V(l)|^2 = \frac{1}{2}G_{\text{ss}}\,\sin^2\left(\frac{\pi l}{L}\right)|V|^2 \tag{8.2.58}$$

and so the contribution of both of the side slots to the feed admittance is G_{2ss} $\sin^2(kl/L)$, referred to the side edge of the patch, at some position l from the short circuit end of the PIFA.

The feedpoint impedance can be calculated from the transmission line model with the estimates of the edge conductances. The admittance form is simpler for the parallel combination, and is

$$Y_{\text{F}} = -\text{j}Y\cot(kl) + Y\frac{Y_{\text{es}} + \text{j}Y\tan(k(L-l))}{Y + \text{j}Y_{\text{es}}\tan(k(L-l))} + Y\frac{2Y_{\text{ss}} + \text{j}Y\tan(kW/2)}{Y + \text{j}2Y_{\text{ss}}\tan(kW/2)}, \tag{8.2.59}$$

where $Y = 1/Z = \sqrt{\epsilon}/Z_0$. This is given in Figure 8.2.14 with some measurements. Table 8.11 gives the experimental details.

For the particular height of these experimental PIFAs, the feedpoint for a best match to 50 Ω is about $l/L = 0.58$.

The antenna may be very compact with high permittivity, ϵ, and small height, but in general the basic constraints above must apply where the inscribing sphere is now a hemisphere above the groundplane. For the PIFA, however, experimental results demonstrate the bandwidth capability.

For the 160 mm wide PIFA, a centre frequency of $f_c = 854$ MHz with the feedpoint at $l = 34.8$ mm ($l/L = 0.44$, $l/\lambda_c = 0.1$) gives a $VSWR = 1.03$, and a $VSWR = 2$ bandwidth of 51 MHz, and a $VSWR = 1.5$ bandwidth of 29 MHz; corresponding to a Q of 11.8. The minimum Q found experimentally was 10.7, for the

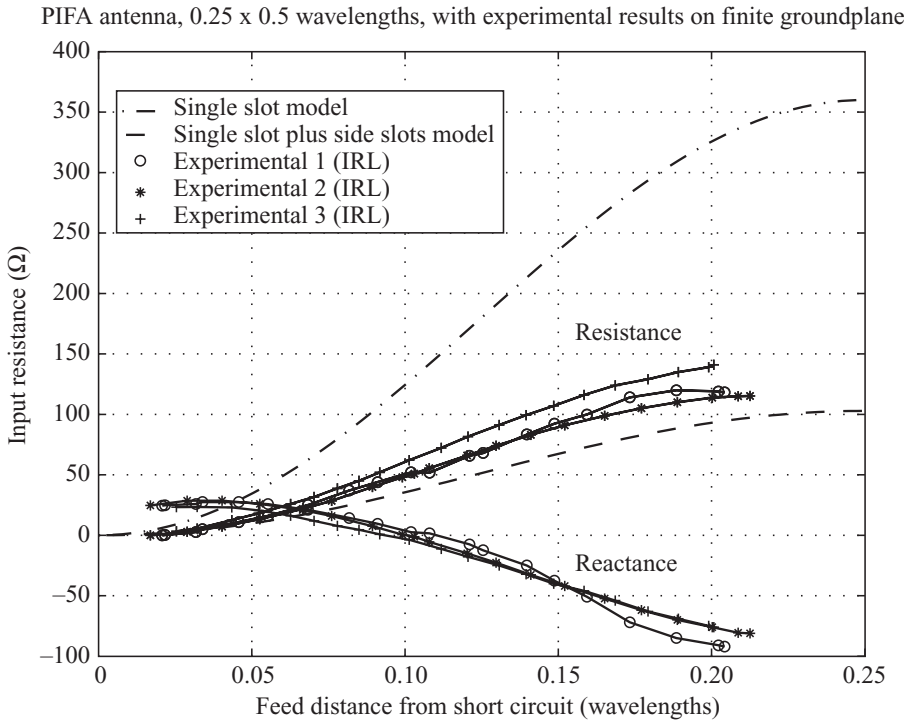


Figure 8.2.14 *The input resistance at the feedpoint as calculated from the transmission line model, as a function of feed position along the length of the PIFA. The feed is in the centre of the transverse direction (W/2). The measurements are for three different brass PIFAs operating at the frequency at which the feed impedance best matched* 50 Ω *at resonance, and the different measurement sets give a feel for typical experimental error*

Table 8.11 *Experimental parameters for the PIFA impedance measurements. The internal height in each case is 7.2 mm, and the patch thickness 0.9 mm. The width is an external measurement. The groundplane is 300 mm square, and larger sizes have no effect on the impedances*

	Width (mm, λ_c)	Length (mm, λ_c)	Height (7.2 mm)	Feed position $l/L, l/\lambda_c$	Measurement frequency, f_c
1	75 mm, $0.2230\lambda_c$	150 mm, $0.446\lambda_c$	$0.0214\lambda_c$	0.48, $0.120\lambda_c$	892.0 MHz
2	70 mm, $0.2275\lambda_c$	140 mm, $0.455\lambda_c$	$0.0234\lambda_c$	0.45, $0.113\lambda_c$	975.0 MHz
3	80 mm, $0.2332\lambda_c$	160 mm, $0.466\lambda_c$	$0.0210\lambda_c$	0.41, $0.102\lambda_c$	847.5 MHz

configuration: $l = 40$ mm, $(l/L = 0.5,\ l/\lambda_c = 0.12)$; $f_c = 862$ MHz, $VSWR_c = 1.26$; $VSWR = 2$ bandwidth of 57 MHz, corresponding to a loaded relative bandwidth of 6.6 per cent. A wider bandwidth can be obtained from increasing the height of the PIFA, and the maximum $VSWR = 2$ bandwidth is about 14 per cent, or $Q = 5$,

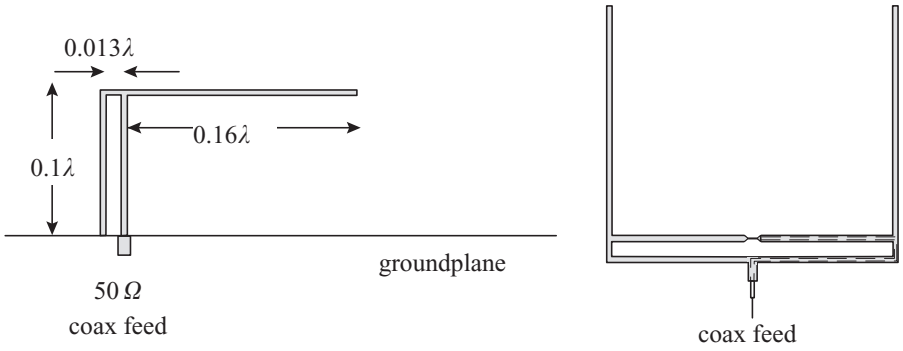


Figure 8.2.15 The wire inverted F antenna configured for a 50 Ω *input impedance, and its full space counterpart shown with a coaxial feed arrangement*

without adding extra parasitic conductor layers (see below for an example of a PIFA with parasitic element).

For the square patch which corresponds to the above PIFA design (without the parasitic element), which is twice the area but the same height, the Q formula (see Section 8.2.5.4) which assumes a small electrical height, gives $Q = 0.6\lambda/h = 29$, or a relative bandwidth of $\sqrt{2}/Q = 2.4$ per cent.

The inscribing sphere for the PIFA is $k_0 a \approx \pi/2$, so the PIFA is not compact in the spherical volume sense, although a high-permittivity substrate such as Alumina ($\epsilon = 9.9$) reduces the physical dimension to about $k_0 a = 0.5$, but the bandwidth is also reduced.

The name PIFA stems from the first forms of this antenna which were wire structures above a groundplane, forming an F-shape. These were also called inverted L antennas from early work by R.W.P. King. An example is given in Figure 8.2.15 which matches to 50 Ω, calculated from the moment method, and the full-space version with a coaxial feed through the conducting arms. The bandwidth of these has been enhanced by adding plates to the wires (Fujimoto *et al.*, 1987), which later evolved to the PIFA.

8.2.5.4 Square patch

The square patch antenna is common because of its simple construction. It is normally a half-wavelength (in the substrate) long and wide, which allows coaxial feed placement to match to a load, typically 50 Ω. The patch can be fed by stripline as shown in Figure 8.2.16. The edge impedance is typically several hundred ohms, but a stripline transformer can be included to convert to a match. The advantage of the patch over the PIFA is clear from this: there is no need to fabricate a short-circuit wall along one of the edges. With a stripline feed, the antenna and its matching section can be made on a single layer. Similarly, several antennas can be connected in an array, complete with impedance matching.

The following is from Andersen (1982). The fields inside the patch are taken to be in the vertical direction only and vary as $\boldsymbol{E} = \hat{\boldsymbol{y}} V \cos(\pi z/L)$, i.e. uniformly across the width W and sinusoidally along the length L.

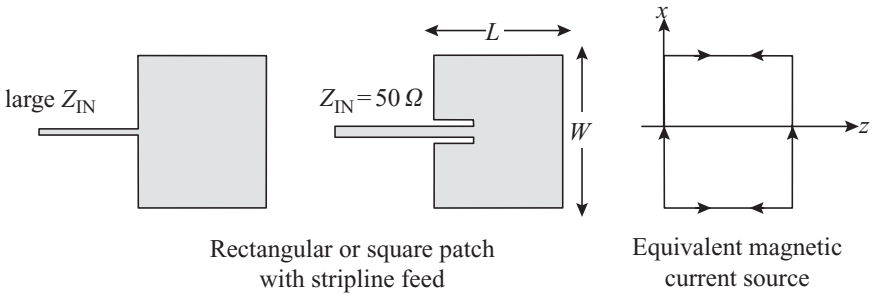


Figure 8.2.16 *The rectangular and square patch fed with a high impedance stripline matched to the edge (several hundred ohms), or to a location to match to a lower impedance, typically 50 Ω, or by coaxial feed from the groundplane (not shown). Normally, the patch dimensions are half by half wavelength in the substrate material*

The magnetic loop of current has a far-field pattern, using the coordinate system of Figure 8.2.16 and normalised in the broadside ($\hat{\boldsymbol{y}}$) direction with $E_\theta(\pi/2,\ \pi/2) = 1$, of

$$\begin{aligned} E_\theta(\theta, \phi) &= \frac{\sin u_W}{u_W} \cos u_L \sin\phi \\ E_\phi(\theta, \phi) &= \frac{\sin u_W}{u_W} \cos u_L \sin\phi \cos\theta \left(1 - \frac{(k_0 L \sin\theta)^2}{\pi^2 - (k_0 L \cos\theta)^2}\right) \end{aligned} \tag{8.2.60}$$

where $u_W = k_0 W/2\ \sin\theta \cos\phi$ and $u_L = k_0 L/2\ \cos\theta$. The first term in the brackets is due to the end slots of length W, and the second term is for the edge slots of length L. The procedure is given below for the PIFA. The edge slots contribute very little to the radiation pattern, with the uniformly illuminated end slots dominant. The maximum directivity is 10 dB for an air substrate, and reduces as the permittivity increases, approaching a limit of about 5 dB. It is about 5.5 dB for a relative permittivity of 10.

From the cavity model, the Q can be calculated from its definition, $Q = 2\omega U/P_{\text{rad}}$. U is the stored energy under the patch,

$$U = \frac{1}{4}\epsilon_r \epsilon_0 \int_V |E|^2 \mathrm{d}V, \tag{8.2.61}$$

where E is the electric field in the patch. $P_{\text{rad}} = 1/2 G_{\text{fsq}} |V|^2$ is the power at the feedpoint with the feedpoint conductance, G_{fsq}. The ohmic loss of both the patch and groundplane, and any loss in the substrate, is neglected here.

Using the (maximum) directivity formula, $G = 1/(30\,D)(W/\lambda)^2$, the Q can be written

$$Q = \frac{\sqrt{\epsilon_r}}{32} \frac{\lambda^2}{hW} D; \tag{8.2.62}$$

D needs to be calculated from an integration of the far fields for a specific patch size. However, for a small width (large permittivity) D is constant at about 5 dB, and so Q

is proportional to the cross-section area of the patch, hW. For larger patch sizes, the directivity is proportional to W and the formula reduces to

$$Q = \frac{\sqrt{\epsilon_r}}{32}\frac{\lambda^2}{hW}D$$
$$\approx 0.6\frac{\lambda}{h}, \quad \epsilon_r = 1. \tag{8.2.63}$$

Recall that this is for small heights only, and does not include ohmic losses.

8.2.5.5 Array of square patches

The advantage of the square patch is the ability to build it using a single substrate. Matching striplines can be included to connect an array of elements. An example is given in Figure 8.2.17. Although a planar array is not compact in the spherical volume sense, it is very low profile in the planar sense. The bandwidth can be made surprisingly large, considering the elements are low bandwidth patches, as seen in the following example (Teal *et al.*, 1997). The patches are edge-fed, and are strip-line joined to give a single 50 Ω feed. The substrate permittivity is urethane foam, and there is also a plastic radome above the patch. Consequently there are some losses in the antenna: in the copper, in the dielectric radome (1 mm thick) and in the foam substrate. The patches are all in phase, giving high gain pattern. A pattern example and the impedance match details are given in Figures 8.2.18 and 8.2.19 (Teal *et al.*, 1997). The impedance is arranged for a maximum impedance bandwidth, rather than a perfect match at a single frequency. Because of the parallel nature of the feed, the feedpoint impedance is quite insensitive to small impedance variations at each patch element. The measured Q is, from the $VSWR = 2$ bandwidth of 0.42 GHz and a centre frequency of 2.4 GHz, $Q = 1/\sqrt{2}(2.4/0.42) \approx 4$.

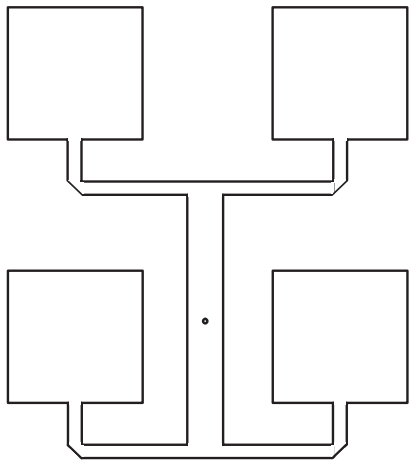


Figure 8.2.17 A four-element array of square patch antennas co-phased for broadside directivity. The dot indicates a coax feed position matched to 50 Ω, with the central stripline 100 Ω and the patch feed lines 400 Ω. The patches and feed lines are all one print layer of conductor

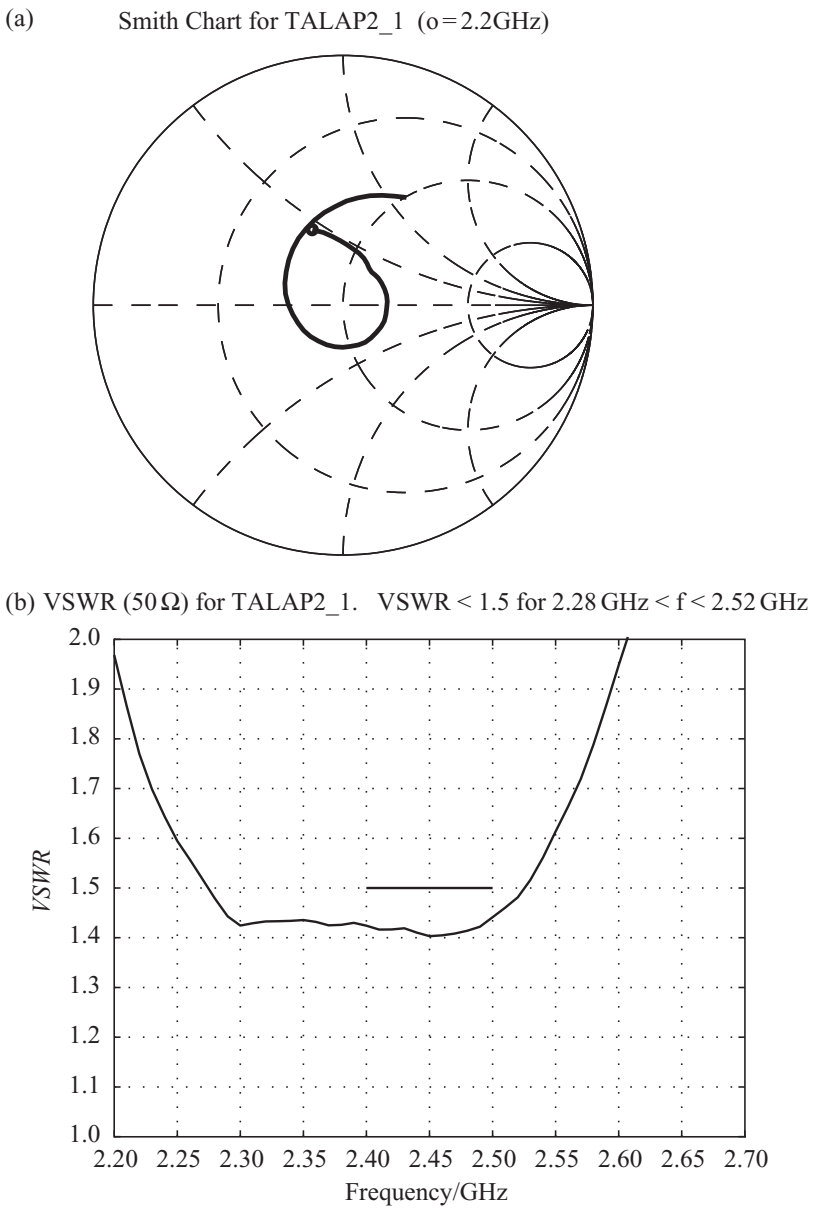


Figure 8.2.18 (a) The Smith chart impedance plot and (b) VSWR measured for a four-patch array. The VSWR = 2 bandwidth is 17.5 per cent and the VSWR = 1.5 bandwidth is 10 per cent

The good polarisation purity results from the radiation from the patch end slots dominating radiation from the patch edges and the groundplane edges.

8.2.5.6 Multiple, stacked patches

The bandwidth of a patch antenna may be increased considerably by having multiple patches closely spaced, shown for example by Waterhouse (1999), but, as noted above, the antenna is no longer electrically small. Most published wideband patch antennas have $ka > 1$, in which case it is possible to make double resonance structures, which are inherently wideband. In practice it is difficult to decide exactly how large a part

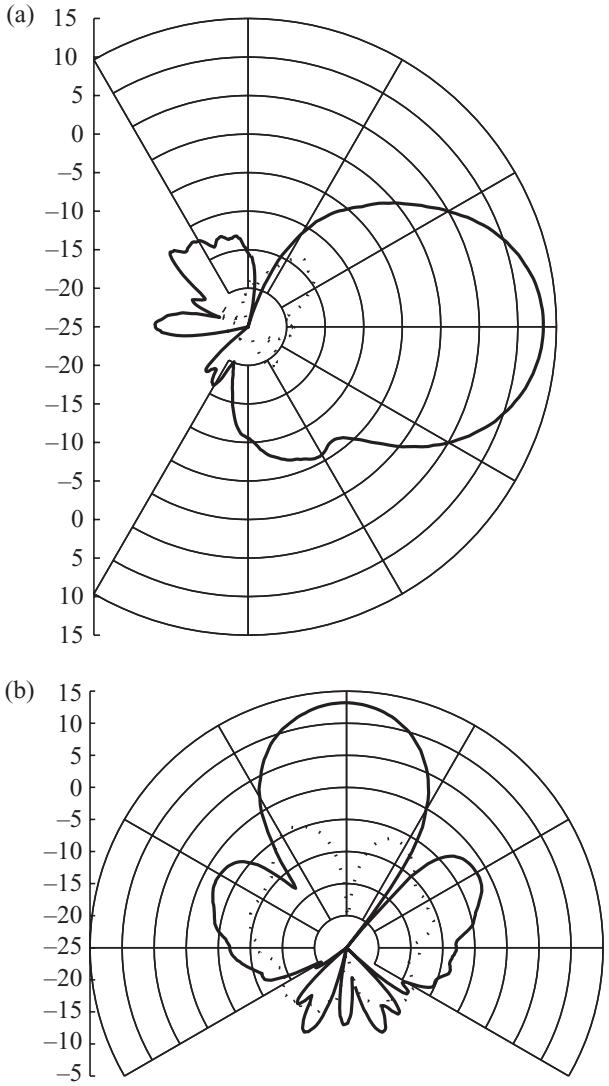


Figure 8.2.19 Measured patterns for the four-patch array. (a) Azimuthal pattern cut, (b) elevation cut. The solid line is co-polar (vertical polarisation) and the dotted line is the cross-polar (horizontal) (courtesy IRL)

of the groundplane should belong to the antenna, and this problem remains with handset-mounted antennas, where the handset case is an integral part of the antenna.

A PIFA design with single 'parasitic' patch element giving a double resonant structure (the *VSWR* is similar to Figure 8.2.4 but has only two minima near $VSWR = 1$), reported by Bonefačić *et al*. (2001), has been confirmed as having at least a 44 per cent $VSWR = 2$ bandwidth. This relative bandwidth corresponds to $Q = 1.6$ if calculated from the single resonance formula of equation (8.2.12). Since the bandwidth of this particular PIFA (without the parasitic element) is given as 10 per cent, this bandwidth increase well exceeds that expected from Table 8.10. The parasitic element is a sloping plate which is the same size as the PIFA top, and is fixed (has zero height) at the top of

the short circuit PIFA edge and slopes to an opening at the open slot edge. The dielectric is air. This design has $W = 0.444\lambda_c$, $L = 0.257\lambda_c$, $(L-l) = 0.0533\lambda_c$, height $= 0.0444\lambda_c$, and height of parasitic element at its open end $= 0.0621\lambda_c$ (total height at end slot edge $= 0.1065\lambda_c$). The inscribing sphere is not significantly larger than the basic PIFA, $ka \approx \pi/2$.

8.2.5.7 Low-profile dual layer patch

Of particular interest to vehicular-mounted and wall or ceiling mounted antennas is very low profile antennas, as given above. Here compactness in a spherical-containment sense is less important because there is plenty of 'flat area' available for mounting, but quite a 'cost' associated with any height. However, the spherical volume measure remains of interest to see how compact a simple design can be made, but still maintaining a low profile.

While the Goubau antenna is relatively low profile, it is possible to maintain the low profile but make a more easily constructed antenna by moving towards patch techniques. Such a hybrid concept is shown in Figure 8.2.20. This is basically a centre-fed patch, although it has a much smaller diameter than a resonant mode patch would have, and so it may be also viewed as a top-loaded monopole. (Even the lowest mode air-dielectric circular patch requires a diameter of about 0.6λ for resonance.) Also,

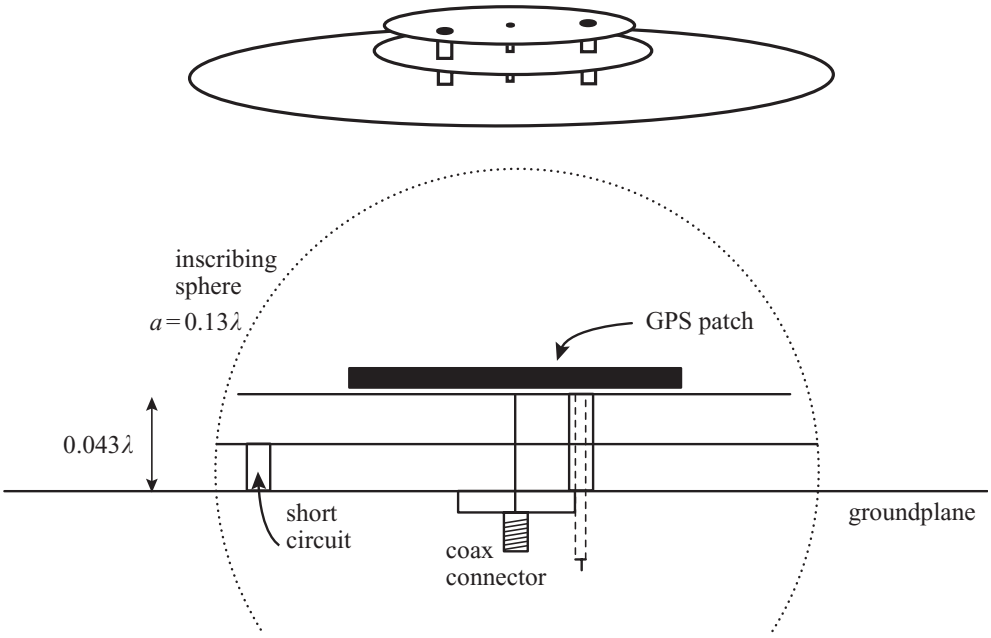


Figure 8.2.20 An example of an antenna which combines patch and Goubau concepts; a low profile patch comprising centrally fed, dual circular patches, with shorting posts. The GPS patch antenna on the top (with ceramic dielectric), does not affect the operation. The inscribing sphere is somewhat misleading in the sense that it does not cover the radiating currents on the groundplane

the antenna height required (for making the volume larger) means that a two layer patch is required to support the required bandwidth. The two layers do not act as separate resonant structures, and this can be seen from the fact that its impedance locus comprises a smooth change throughout the bandwidth. There are shorting posts in the patch, two near the edge and one close to the feed, which are reminiscent of the Goubau structure and the PIFA. The manufacture of such designs is more difficult than for single layer patches, but simpler than for Goubau type antennas.

The patch example is for cellular applications, and operates within the AMPS band. This particular design also has a GPS patch antenna (1575 MHz) mounted on top. The GPS antenna is described in Section 8.3. The two antennas operate at different frequencies and do not couple strongly, so the addition of the GPS patch does not affect the cellular patch behaviour. The design offers a compact GPS/cellular antenna for rooftop type mounting.

The details are: centre frequency is $f_c = 882$ MHz; relative bandwidth measured at $VSWR = 2$ is 74 MHz; giving a relative bandwidth of $74/882 = 0.0839$; a loaded Q factor of $Q_L = 11.9$; and an unloaded Q factor of $Q = 1/\sqrt{2}(f_c/\Delta f_{VSWR=2}) \approx 8.43$. The height is 14.5 mm; diameter, 85 mm; giving the radius of the inscribing sphere, $a = 43$ mm, and electrical radius, $k_c a = 0.83$. The theoretical Q factor is $Q_{th} = 1/(ka)^3 + 1/(ka) \approx 2.95$; and finally the ratio of Qs is $Q_u/Q_{th} \approx 2.85$.

The inscribing sphere for estimating the spherical compactness is somewhat misleading for the groundplane antenna. This is because the radiation depends on groundplane currents, and in practice the input impedance is noticeably affected by the groundplane size when it is less than a diameter of about a half-wavelength. However, a full-spherical space version of this type of antenna can be realised, and in this case, the inscribing sphere size holds (Figure 8.2.21).

8.2.6 Antennas on a handset (free space)

Modern portable mobile phones are so small (or rather large, relative to the carrier wavelength) that the total structure must be considered when designing an antenna. The standard small helix and monopole antennas serve as well as coupling elements to the chassis or phone as antennas in their own right. The patterns and impedances and bandwidths are determined by the total structure. In general this means that it is necessary to perform a numerical analysis of the antenna plus the structure, and general design principles are scarce. An important factor is the user influence which we shall treat in the following section.

An example is given in Toftgård *et al*. (1993) where typical patterns and impedances are shown (Figure 8.2.22). Figure 8.2.22(b) shows that the θ-component as a function of θ has a major lobe in the downward direction (for a vertically oriented antenna). This is indicative of a major contribution from currents running along the chassis, so the antenna plus phone may be considered as an asymmetric dipole with a fat lower part. The structure is thus also part of creating a low-Q antenna as is apparent from the almost constant real part of the impedance.

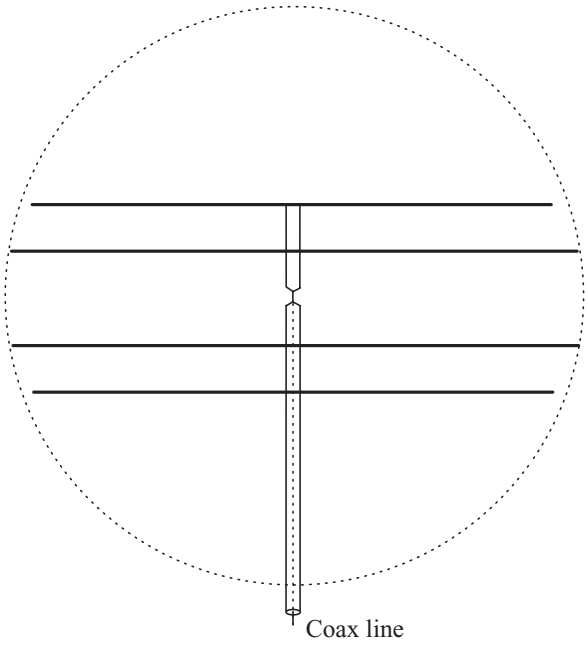


Figure 8.2.21 The dual-layer patch antenna as a full-spherical space version. The shorting post details are omitted, and the feed is now by coaxial line with the inner connected to an upper feed line of about the same width as the coaxial outer

From a user point of view an integrated antenna without any protruding parts is an advantage and was demonstrated by Pedersen and Andersen (1994), where it was shown that it was possible to match the antenna over the whole GSM band around 900 MHz. A PIFA antenna is essentially a slot antenna as described above, and the currents are largest near the end slot. A more general overview is given in Jensen and Rahmat-Samii (1995).

8.2.7 Antennas on a handset near a person

There are two important aspects of the user proximity to the radiating antenna, a communication aspect and a perceived health risk aspect. They are partly related because the main effect of human tissue close to the antenna is absorption, and absorption density, the so-called SAR (specific absorption rate) in a 1 g or 10 g tissue mass is the determining parameter in the health discussion. The total absorption is vividly demonstrated in Figure 8.2.23 (courtesy G.F. Pedersen) showing the total absorption measured for 43 different users, who were asked to hold the antenna as was found usual for them. In the figure the persons are sorted after absorption ranging for the helix antenna from 98 per cent to 82 per cent. The figures do not reveal where the absorption takes place or how high the SAR values are, but from a communication point of view, this contribution to the loss budget is serious. The reason for the high losses from the helix is apparently the current flowing down the surface of the phone inducing currents in the hand. This also explains the variability among persons, since

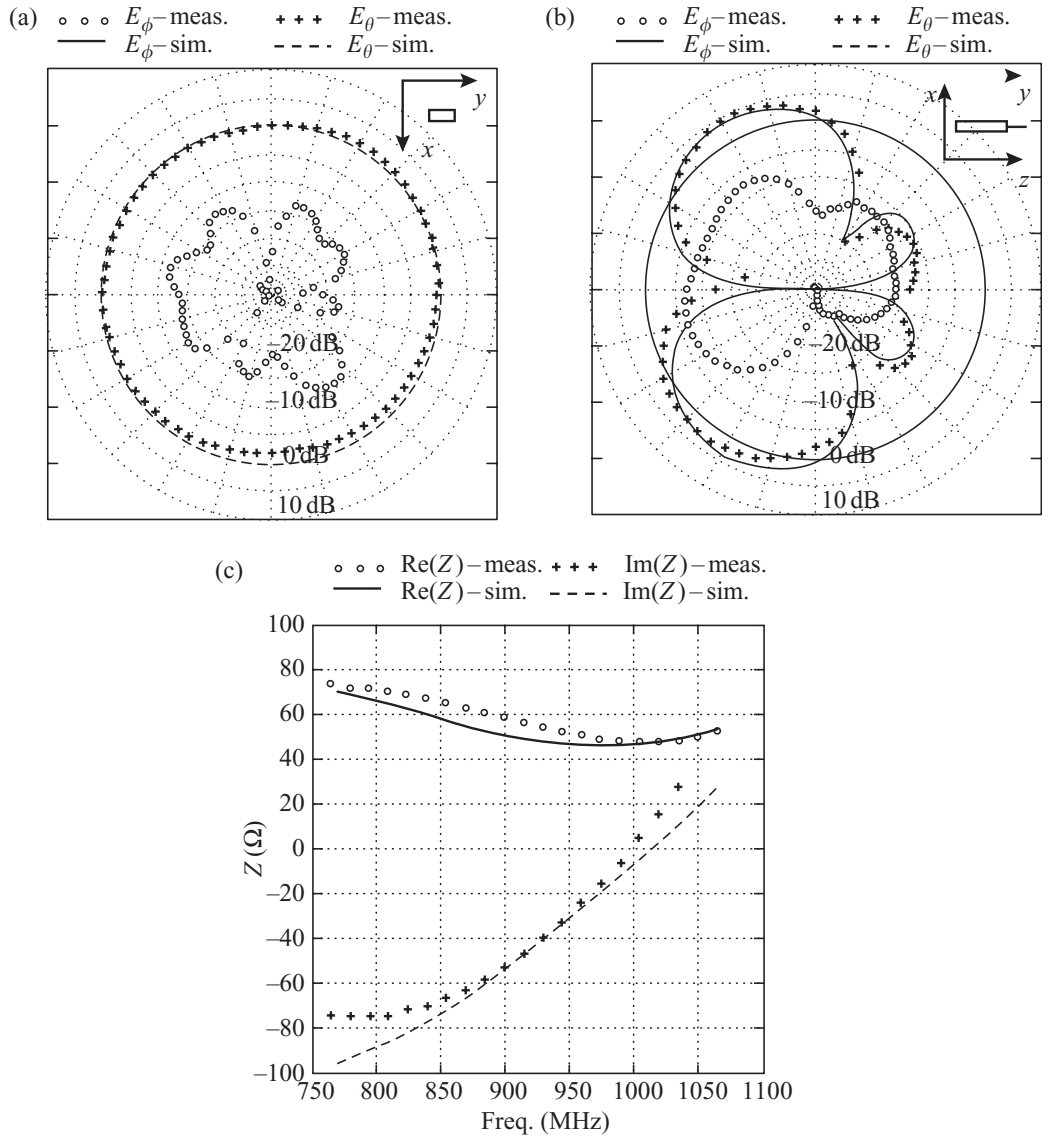


Figure 8.2.22 *Simulated and measured results for a λ/4 monopole mounted on a box (15*6*2.5 cm) at 900 MHz (Toftgård et al., 1993). (a) Azimuth and (b) elevation patterns; (c) impedance*

the losses increase the closer the hand is to the top of the phone where the antenna is positioned.

The absorption in the head is automatically followed by a shadowing effect, shown in Figure 8.2.24, similar to Figure 8.2.22 but with a person included. For E_θ there is a difference on the two sides of the head of a few dB, and a major change in polarisation. These calculations and measurements were in an anechoic chamber, a situation quite far from the random scattering environment in the real world (cf. Vaughan, 1996; Vaughan and Scott, 1999; Toftgård *et al.*, 1993).

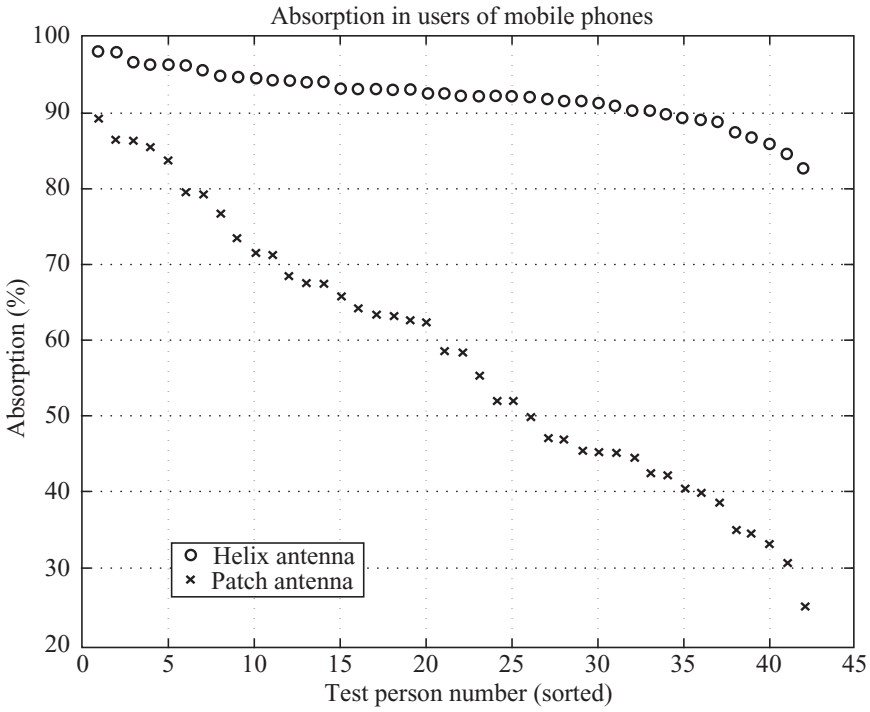


Figure 8.2.23 Absorption for handheld antennas (a helix and a patch) for a large group of persons (courtesy G.F. Pedersen)

8.2.8 Handset antennas in a random environment

As noted in Chapter 2 the time-dependent complex signal at an antenna port is given by the open circuit voltage

$$V(t) = \int \boldsymbol{E}(\Omega, t) \cdot \boldsymbol{h}(\Omega)\, \mathrm{d}\Omega \tag{8.2.64}$$

where the integral is over real space, $\boldsymbol{h}(\Omega)$ is the electric far field (complex height) of the antenna, which here includes the person, and $\boldsymbol{E}(\Omega, t)$ is proportional to the electric field of the incident waves, and t indicates that the environment is changing with time, usually by movement of the user. By assuming that the phases of the incoming fields are random and independent of polarisation and direction of arrival, the average power, $P = \left\langle V(t)V^*(t)\right\rangle_t$ (here V is the rms value) is given by

$$P = \int (P_1 P_\theta(\Omega) + P_2 P_\phi(\Omega) G(\Omega))\, \mathrm{d}\Omega \tag{8.2.65}$$

where P_1 is the total power in the θ-polarisation, and P_θ is normalised so

$$\int P_\theta(\Omega)\, \mathrm{d}\Omega = 1 \tag{8.2.66}$$

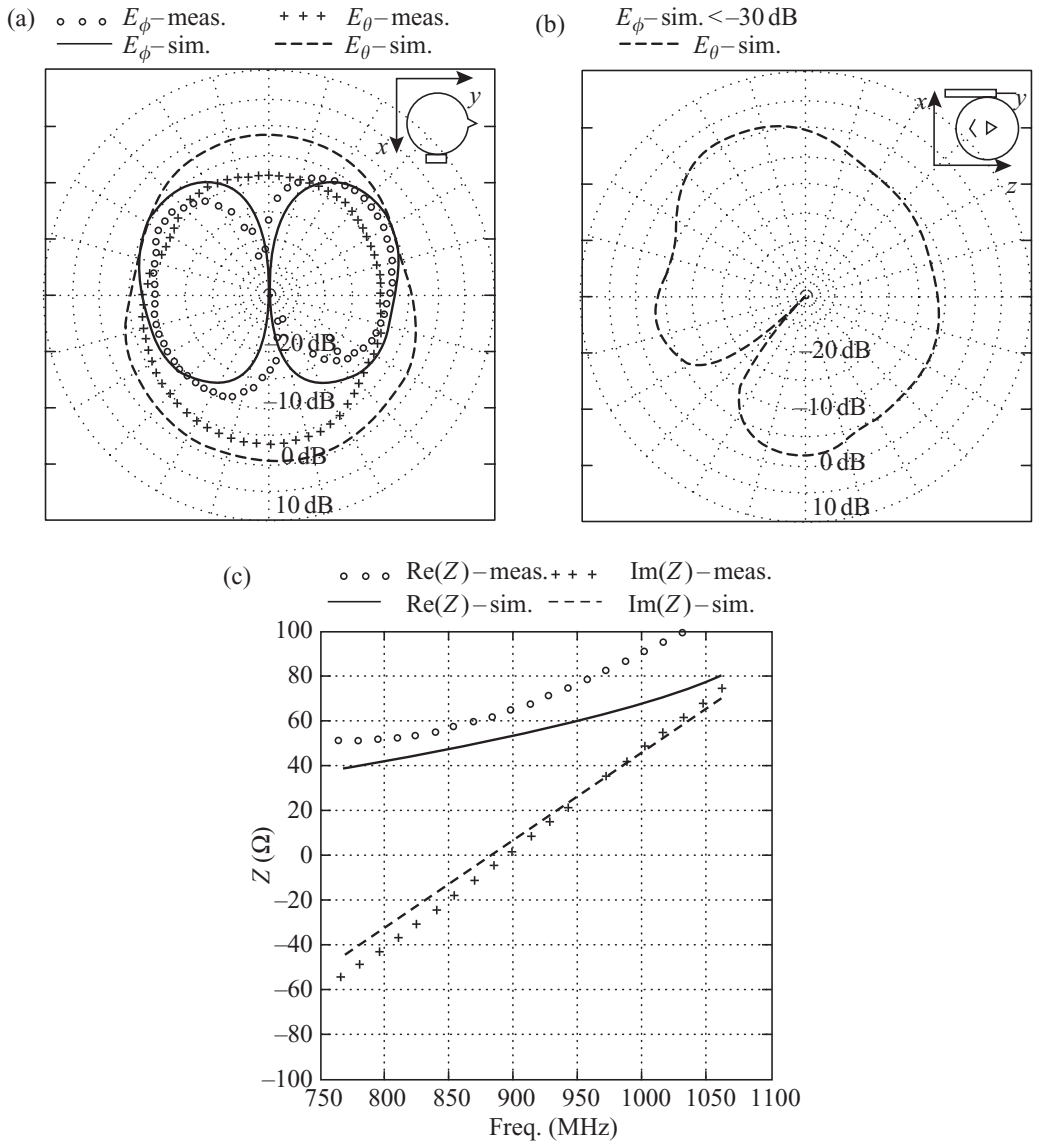


*Figure 8.2.24 Simulated and measured results for a λ/4 monopole mounted on a box (15*6*2.5 cm) at 900 MHz (Toftgård et al.). (a) Azimuth and (b) elevation patterns; (c) impedance. Person included*

and similarly for P_2 for the ϕ-polarisation. The antenna gains are normalised in the usual manner, such that

$$\int (G_\theta(\Omega) + G_\phi(\Omega))\, \mathrm{d}\Omega = 4\pi. \tag{8.2.67}$$

By also normalising the powers, Taga (1990) arrived at a very useful expression, coined the Mean Effective Gain, *MEG*,

$$MEG_{\text{random}} = \int \left(\frac{XPD}{1+XPD}\, P_\theta(\Omega) G_\theta(\Omega) + \frac{1}{1+XPD} P_\phi(\Omega)\, G_\phi(\Omega) \right) \mathrm{d}\Omega \tag{8.2.68}$$

where the ratio between the two average polarisation powers, or cross-polar discrimination, $XPD = P_1/P_2$, has been used.

The *MEG* is thus a normalised gain measure in a random environment, equal to $\frac{1}{2}$ for isotropic antennas. It expresses clearly that uncorrelated signals are added, since they are added (integrated) on a power basis. Note also that P and G only appear as a product.

An interesting question is the choice of antenna pattern to maximise the received power for a one-port antenna with a fixed pattern. In classical antenna theory there is only one incident wave at a time, and it is easy to maximise the gain in that direction. In a random environment it is not so simple. Consider an environment with a Gaussian distribution of power as a function of angle and a Gaussian antenna beam. We assume only one polarisation, and the patterns are sufficiently narrow so that we can integrate to infinity in a theoretical 2D situation. The incident power distribution and the power pattern have functions

$$P_\phi = \frac{1}{\sqrt{2\pi}\Delta_P}\, \mathrm{e}^{-\phi^2/(2\Delta_P^2)}, \quad G_\phi = \frac{1}{\sqrt{2\pi}\Delta_G}\, \mathrm{e}^{-\phi^2/(2\Delta_G^2)}, \tag{8.2.69}$$

where Δ_P is the angular spread of the incident power, and Δ_G is a measure of the antenna beamwidth. The result of the integration for the *MEG* is a normalised power of

$$MEG = \sqrt{\frac{\Delta_P^2}{\Delta_P^2 + \Delta_G^2}} \tag{8.2.70}$$

so the maximum gain is achieved when the antenna beamwidth tends to zero. The real significance, however, is that it does not pay to have a beamwidth much smaller than the angular spread of the environment. For $\Delta_G = \Delta_P/2$, only 0.5 dB is lost through a reduction in the distributed directive gain, and so the available gain of narrowing the beam further is almost insignificant. It should be emphasised that this is for a one-port antenna, or fixed beam. If the antenna is adaptive, an array with many elements, the situation is quite different and large gains may be achieved, as discussed in Chapter 9.

The interaction between the antenna and environment is further emphasised by studying the case of an antenna in a completely random environment, characterised by $XPD = 1$, and $P_\theta = P_\phi = 1/(4\pi)$. The *MEG* is

$$MEG_{\text{random}} = \int \left(\frac{1}{1+1}\frac{1}{4\pi}\, G_\theta(\Omega) + \frac{1}{1+1}\frac{1}{4\pi}\, G_\phi(\Omega) \right) \mathrm{d}\Omega = \frac{1}{2} \tag{8.2.71}$$

so the received power is completely independent of the antenna gain, or expressed differently, the effective gain is unity. The antenna designer should have an easy task! It also shows that the effect of shadowing (but not absorption) by the person, incorporated in the gain, is insignificant in a totally random environment.

These limiting results are a result of the scenario model of a continuous power distribution arriving from all directions. In practice the environment is not uniform and a variation in the measured mean effective gain will occur when the person and antenna are rotated. This is shown in Figure 8.2.25 (Pedersen and Andersen, 1999)

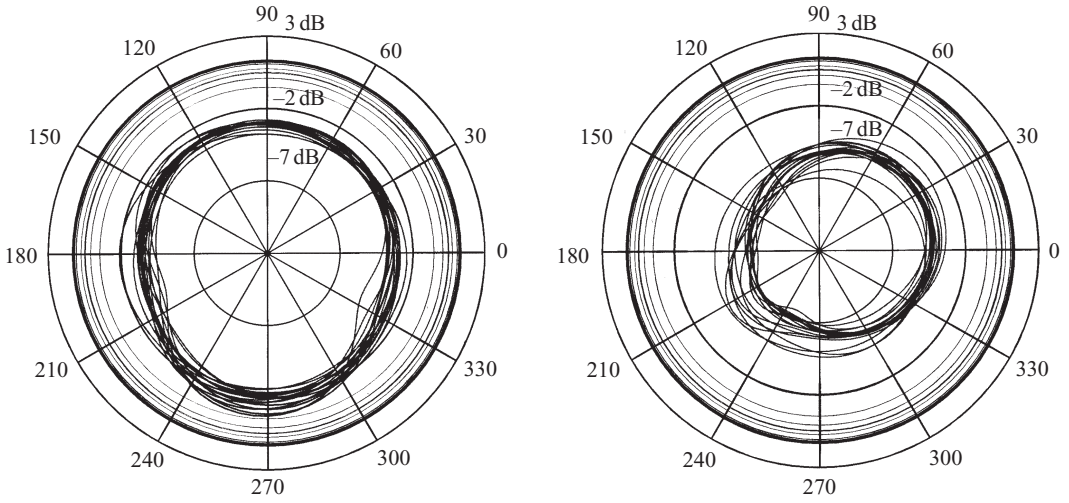


Figure 8.2.25 Combined experimental and numerical MEG values for a dipole antenna on a handheld phone, when the handheld is rotated 360° *in the horizontal plane for 18 different locations. The handheld is indoor and the base station is outdoor. The handset has an inclination angle from vertical of 60°. The outermost thin circles are MEG values for a vertical reference dipole in the same environments. To the left without a user and to the right with a model of a user's head (Pedersen and Andersen, 1999)*

where experimentally obtained values of the incident power distribution in 18 different environments are integrated with numerical results for an antenna (in this case a dipole on a box) to obtain the estimate of *MEG*. The left plot is without a person. It is noted that the gain is rather insensitive to the environments and to rotation of the antenna.

8.3 Antennas with circular polarisation

8.3.1 Introduction

Most satellite communications systems use circular polarisation. The reason is that Faraday polarisation rotation in the ionosphere creates a changing orientation of the linear polarisations. Fixed, linearly polarised antennas cannot maintain good polarisation efficiency in such a situation, but for circular polarisations, the rotation does not affect this efficiency. Circularly polarised antennas are generally more complicated to design and build than linearly polarised antennas. For portable terminals, performance is traded for construction simplicity and so linearly polarised antennas are often used, despite a 3 dB polarisation mismatch. Similarly, the ideal pattern is traded off by what is practicable, and in doing so, the distributed directivity is further degraded from the ideal. In this section, the basic concepts of circular polarisation and circularly polarised mobile satellite antennas are reviewed.

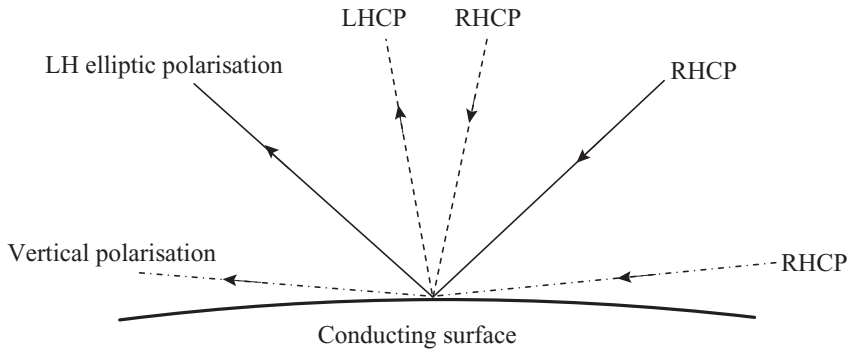


Figure 8.3.1 Circular polarisation alters after a reflection. Here a smooth conductive surface reflects right-hand circular polarisation (RHCP) to left-hand (LHCP) at broadside, and at grazing angles only the vertical polarisation emerges

8.3.1.1 Polarisation of reflections from a smooth surface

Even in the multipath environment of land–mobile communications, circular polarisation, in principle, offers some advantage, based on the polarisation purity of reflections from smooth surfaces. An incident right-hand polarised wave reflects as a left-handed wave for broadside reflection; see Chapter 3. For grazing angles along a conducting surface, the polarisation is linear, and for angles in between, the polarisation is left-hand elliptic, as depicted in Figure 8.3.1. So for purely circular-polarised antennas, the received power from first reflections is lower than for the case of linearly polarised antennas. For a line-of-sight situation where there is also some multipath, a circularly polarised link may perform better, since the power in the first reflections of the multipath becomes suppressed. Some early experimental results showing the decreased fading from using circular polarisation indoors are given by Rappaport and Hawbaker (1992).

8.3.1.2 Mobile satellite links

Most mobile satellite links, despite having a line-of-sight in most operating environments, still experience some Rician fading. This is because of the multipath from ground reflections and elevated scatterers such as trees and power-lines. This can be minimised by employing directive antennas, but the motion of the mobile terminal (and thence any fixed beam from the mobile) can then create fading as the beam strays about the line-of-sight direction, owing to the motion of the terminal. For vehicular-mounted satellite antennas, some solutions employ array antennas which accept pointing direction data from other sensors in the vehicle, while other array solutions adaptively track the line-of-sight direction. The result of these approaches is high-performance, high-cost antennas. Pencil-beam-type directivity requires large antennas, or at least large antenna aperture, and generally in mobile communications, small, low-cost antennas are required. Fixed beam antennas have the obvious advantages of simplicity and low cost.

For geostationary satellites, the elevation angle as seen by the mobile terminal changes with the terminal's latitude. For a local region on earth, the latitude is essentially fixed (the satellite moves around its nominal orbital position only by a small amount), and so the elevation angle, in geographical coordinates, to the satellite, is essentially fixed. However, the motion of the mobile acts to spread this elevation angle in terms of the coordinates of the mobile antenna. For vehicles travelling significant distances in latitude, the geographic elevation angle also changes significantly. Some antennas have been designed which can have their beam, which is otherwise fixed, adjusted as a mobile changes its latitude. Pattern examples are given below.

If the satellite link is to work independently of the orientation of the vehicle (direction of travel in geographical coordinates), then the required beams are rotationally symmetric. 'Conical' beams therefore provide the best gain independent of the orientation in the horizontal plane of the mobile, and with finite tolerance (because of the finite elevation beamwidth of the conical pattern) to elevation angle changes.

A vertical, cylindrical antenna offers a practicable aperture for realising such a pattern, and the structure is typically a scanning-mode helix. An advantage of the scanning-mode helix is that, unlike monopoles, its vertical structure offers patterns that are essentially independent (in the main lobe region) of the mounting structure, such as a rooftop or gutter mount on vehicles, or even a handheld terminal. An alternative antenna shape is a low profile arrangement, normally comprising patch-type antennas, which can be arranged to produce circular polarisation. Synthesising a circularly polarised, ideal beam shape for geostationary satellite links is difficult with patch antennas; however, the patch solutions are practical candidates for low-earth orbit (LEO) systems.

Some satellite systems operate with a constellation, for example the global positioning satellite (GPS) system, used for navigation and other applications; and communications systems such as Iridium. Iridium-type systems have low earth orbit satellites, so-called because of their extremely low height above earth. They require many tens of satellites for global coverage. (The Iridium concept started with 77 satellites – hence its periodic table based name – and was later realised with 66.) In constellation systems for communications, the idea is that the terrestrial mobile links to whichever satellite is the most visible, i.e. has the strongest signal strength through the least path loss. A GPS receiver tunes into several satellites at once, typically processing up to six satellite signals. In the GPS system, there are fewer satellites and a much higher orbit.

8.3.2 Circular polarisation parameters and relations

Circular polarisation comprises equal amounts of the linear polarisations in quadrature. When the two linear polarisations are not in quadrature, or their amplitudes are different, or both, then the resulting polarisation is elliptic. If the linear components are in phase, the polarisation is slant linear (see Section 3.1).

Any polarisation comprises a linear combination of any orthogonal pair of polarisations. In any orthogonal pair, including the linear or the circular pairs, the wanted

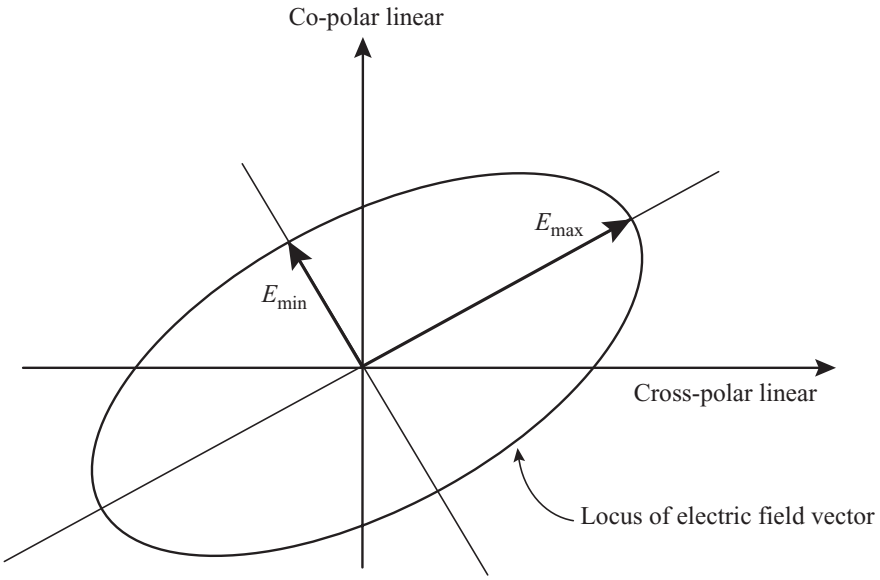


Figure 8.3.2 The polarisation ellipse showing the axial ratio

polarisation is called the *co-polar* component and the other is called the *cross-polar* component. The basic relations between the circular and linear polarisations, for antenna patterns, are as follows. The subscript 'R' is for right-hand circular polarisation, or RHCP, and similarly 'L' for left.

$$E_{\mathrm{R}} = \frac{1}{\sqrt{2}}(E_\theta - \mathrm{j}E_\phi), \quad E_{\mathrm{L}} = \frac{1}{\sqrt{2}}(E_\theta + \mathrm{j}E_\phi) \tag{8.3.1}$$

$$E_\theta = \frac{1}{\sqrt{2}}(E_{\mathrm{L}} + E_{\mathrm{R}}), \quad E_\phi = \frac{-\mathrm{j}}{\sqrt{2}}(E_{\mathrm{L}} - E_{\mathrm{R}}). \tag{8.3.2}$$

The condition for circular polarisation in a given direction, or set of directions, follows as $E_\theta = \pm \mathrm{j}E_\phi$, where the plus sign is for RHCP and the minus sign is for LHCP.

The basic parameters for polarisation amplitude ratios are the *axial ratio* and the *circular polarisation ratio*. The axial ratio is the ratio of the maximum-amplitude linearly polarised component, and the minimum-amplitude linearly polarised component, and these are orthogonal to each other. The polarisation ellipse is shown in Figure 8.3.2. By going to amplitude ratios, note that the phase information of the polarisations is lost.

The axial ratio is defined by

$$A = \frac{|E_{\max}|}{|E_{\min}|} \tag{8.3.3}$$

and the circular polarisation ratio is

$$C = \frac{|E_{\mathrm{R}}|}{|E_{\mathrm{L}}|}. \tag{8.3.4}$$

These equations show that the degree of circular polarisation can be also measured by using power measurements from two oppositely polarised CP antennas. The relationship between the ratios is

$$C = \frac{A+1}{A-1}, \quad A = \frac{C+1}{C-1}. \tag{8.3.5}$$

The polarisation purity is the ratio of powers in each polarisation, for example for LHCP,

$$\eta_{\mathrm{L}} = \frac{|E_{\mathrm{L}}|^2}{|E_{\mathrm{L}}|^2 + |E_{\mathrm{R}}|^2} = \frac{1}{1+C^2}. \tag{8.3.6}$$

These quantities are used in measurements and antenna specifications. For example, a spinning dipole has been traditionally used to measure both polarisations in a pattern measurement, and when such a measurement is expressed as a dB pattern, the axial ratio in dB can be read directly from the amplitude of the oscillations in the pattern caused by the spinning dipole (see below). The advantage of such an axial ratio measurement is that the antennas need not be aligned in a polarisation sense and the convenience with which the pattern parameters can be viewed, but has the disadvantage that polarisation phase angle information may not be recorded. Another way to characterise circularly polarised patterns is to measure directly with co-polar and cross-polar circular polarisations. Examples of both pattern types are given below.

The axial ratio is plotted against the circular polarisation ratio in Figure 8.3.3. For example, an axial ratio of 0.5 dB means that the polarisation is very pure, with the circular cross-polarisation component over 30 dB below the co-polar power. Most antenna realisations cannot attain such polarisation purity over a typical working relative bandwidth of several percent.

A common problem is to confuse the two circular polarisations, a situation that is not helped by different definitions from the various physics and engineering disciplines. Recall that the antenna polarisation is defined from the transmitting pattern (see Section 2.1), and this polarisation is that seen by the antenna as it propagates away from the antenna.

An RHCP wave travelling in the positive z direction (field dependence $\exp(\mathrm{j}\omega t - \mathrm{j}k_0 z)$) has its electric field vector rotating in a clockwise direction, as seen from the transmitting antenna. The locus of the vector traces out a right-hand (screw thread sense) helix. Similarly, the electric field vector of an LHCP wave rotates counterclockwise as seen from the transmitter. At the receiving antenna, the propagating wave is seen from its arrival propagation direction, so the transmitted RHCP is seen as rotating counterclockwise. Table 8.12 summarises these properties.

A related and very similar measure to the polarisation purity is the *polarisation efficiency*, which is a measure of the match between an antenna and an incident wave. It is defined using an incident wave polarisation, $\hat{p}_{\mathrm{inc}}$, and the receiving antenna (transmit) polarisation, $\hat{p}_{\mathrm{tx}}$, which is related to the receiving pattern polarisation, $\hat{p}_{\mathrm{rx}}$, by $\hat{p}_{\mathrm{rx}} = \hat{p^*}_{\mathrm{tx}}$. With the normalisation made explicit here, the polarisation efficiency is

$$\eta_{\mathrm{pol}} = \frac{|\hat{p}_{\mathrm{inc}} \cdot \hat{p}_{\mathrm{rx}}|^2}{|\hat{p}_{\mathrm{inc}}|^2\,|\hat{p}_{\mathrm{rx}}|^2} = \frac{|\hat{p}_{\mathrm{inc}} \cdot \hat{p^*}_{\mathrm{tx}}|^2}{|\hat{p}_{\mathrm{inc}}|^2\,|\hat{p}^*_{\mathrm{tx}}|^2}. \tag{8.3.7}$$

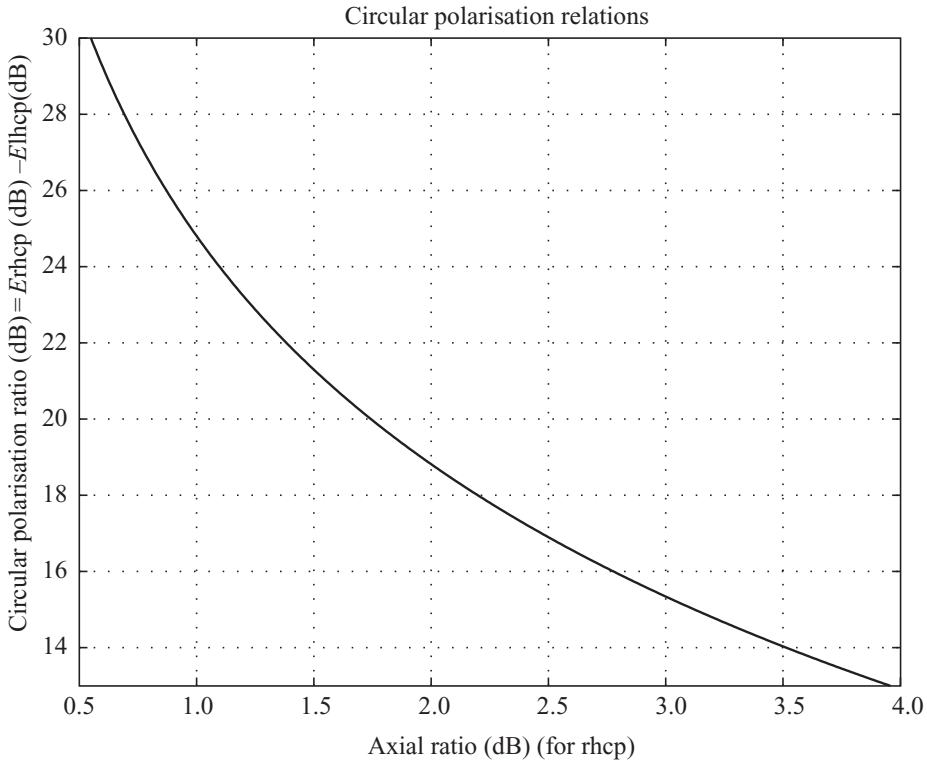


Figure 8.3.3 The circular polarisation ratio in terms of the axial ratio

Table 8.12 Definitions of senses for circular polarisations

RHCP	Rotation of electric field vector seen from transmitter	clockwise
	Rotation seen from receiver	counterclockwise
	Transmit pattern polarisation vector for antenna	$\frac{1}{\sqrt{2}}(\hat{\theta} - \mathrm{j}\hat{\phi})$
	Receive pattern polarisation vector and Incident wave polarisation vector	$\frac{1}{\sqrt{2}}(\hat{\theta} + \mathrm{j}\hat{\phi})$
LHCP	Rotation of electric field vector seen from transmitter	counterclockwise
	Rotation seen from receiver	clockwise
	Transmit pattern polarisation vector	$\frac{1}{\sqrt{2}}(\hat{\theta} + \mathrm{j}\hat{\phi})$
	Receive pattern polarisation vector and Incident wave polarisation vector	$\frac{1}{\sqrt{2}}(\hat{\theta} - \mathrm{j}\hat{\phi})$

For circular polarisation, the RHCP vector conjugates to the LHCP vector, and vice versa. A circularly polarised wave received by a linearly polarised antenna, or vice versa, means a polarisation efficiency of 0.5, i.e. a contribution to the path loss of 3 dB. The polarisation efficiency is normally used for fixed point-to-point links,

where the path loss budget is deterministic, but can take a stochastic interpretation for mobile communications, as noted in Section 6.7.

For circular-polarisation testing antennas, a helix operating in the axial mode is a practical option. This has a diameter of about one wavelength for air-cored helix, and needs only a single wire (monofilar helix) with pitch angle about 12° (Kraus, 1950). A right-hand helix (the wires follow a right-hand screw) transmits predominantly right-hand polarisation in the axial direction away from the feed end. A left-hand helix is also required for measurements of the LHCP. In practice, the polarisation from axial mode helices (or most other CP antennas) will not be perfectly circular, but elliptic. Care should be taken in precision measurements that the polarisations of the testing antennas are properly characterised. Elliptical polarisations are orthogonal when their circular polarisation ratios are the reciprocal of each other and their ellipse tilt angles are at right angles to each other.

In a line-of-sight communications link, with a transmitting right-hand helix, the receiving antenna should also be a right-hand helix, which of course has a left-hand receiving pattern. The link situation can be imagined by considering a single axial mode helix and cutting it in half and separating these halves as the transmitting and receiving antennas.

A scanning mode helix radiates the opposite hand of polarisation to its structure, i.e. a right-handed scanning mode helix radiates LHCP.

Note that the confusion between polarisations does not arise with linear polarisations and antennas. In a commercial world where the customer is more correct than the polarisation conventions, the antenna designer asked for a specific circularly polarised antenna is well advised to make both LHCP and RHCP versions!

8.3.3 Patterns from space loss function for satellite links

For a terrestrial observer at radius r_e, the distance to the satellite d is a function of the zenith angle referenced at the antenna, θ, and the mean orbital radius of the satellite, r_s, depicted in Section 6.7.

The optimum terrestrial pattern depends on the space loss, any pattern shaping of the satellite antennas, and some statistical weighting from the density of satellite constellation. The space loss is recalled from Section 6.7 as

$$\left[d(\theta)\right]^{-2} = \left[r_s^2 + r_e^2 - 2r_s r_e \cos(\beta(\theta))\right]^{-1}, \tag{8.3.8}$$

where

$$\beta(\theta) = \frac{r_e \ \sin^2\theta + \cos\theta \sqrt{r_s^2 - r_e^2 \sin^2\theta}}{r_s}, \quad |\theta| < 90°. \tag{8.3.9}$$

If the power pattern from the satellite is shaped to compensate for this space loss variation, then this should be accounted for at this stage of the design. For an omni satellite antenna pattern, the required directive gain function, to maintain constant average received power, is the reciprocal of the space loss. The space loss alone, relative to the minimum value (when the satellite is directly overhead) is plotted in

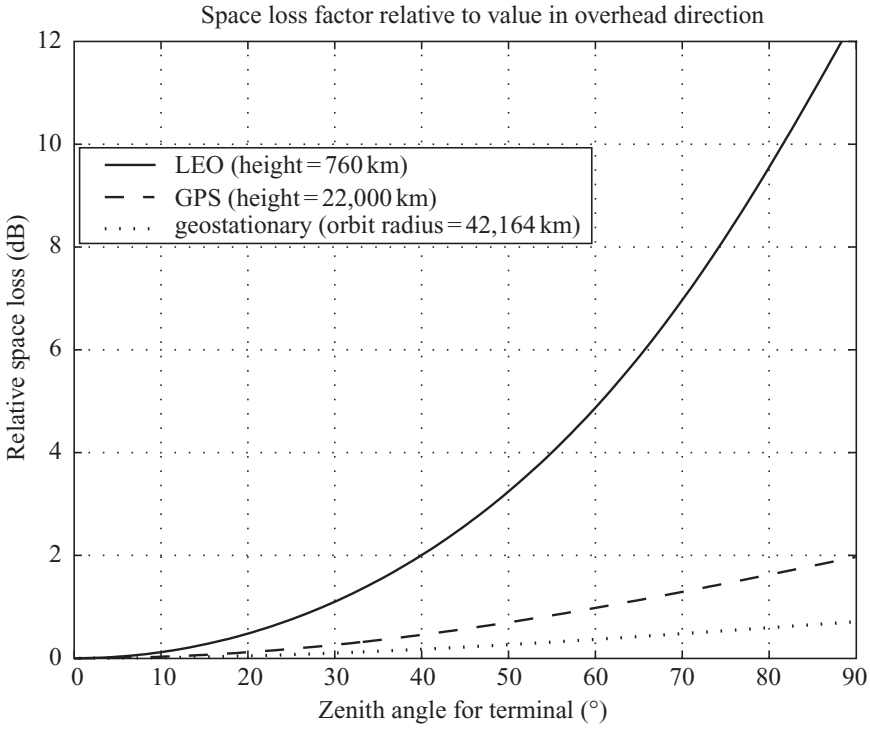


Figure 8.3.4 The space loss of a satellite link normalised by the value at zenith, against the zenith angle. The LEO orbit has a different space loss 'taper' to the GPS and geostationary orbits

Figure 8.3.4. The earth radius is 6371 km, and for a low-earth orbit (LEO) satellite, the height above earth is taken as 760 km; for the global positioning constellation, the height is 22 200 km, and the geostationary orbit is 43 164 km.

The space loss does not vary much with zenith angle for the GPS and geostationary orbits. The LEO space loss is some 12 dB greater at the horizon than at zenith. So without satellite pattern shaping or any statistical weighting for the satellite positional density, the terrestrial antenna should be some 12 dB higher at the horizon than at zenith, in order to receive a constant signal strength from visible satellites. Such ideal patterns are shown in Figure 8.3.5. Consideration must be given to any pattern component near and below the horizon, since ground illumination will contribute to the antenna noise and will also pick up any co-polarised component of scattering from the environment. Because the antennas should be compact and will have relatively smooth patterns, there is a trade-off between higher gain near the horizon and low gain at and below the horizon.

These pattern shapes, with circular polarisation, can be well approximated using short multifilar helices, such as quadrifilar (Kilgus, 1968) operating in the scanning mode; or orthogonal, phased (by 90°) dipoles. The dipoles can be 'drooped' for improved pattern shaping. Low-profile aperture antennas can be realised using planar spirals and patch antennas. Of these, the patch is an easily fed, matched and manufactured option. A short, multifilar helix is the most compact, although appropriately phased feeding and impedance matching is required.

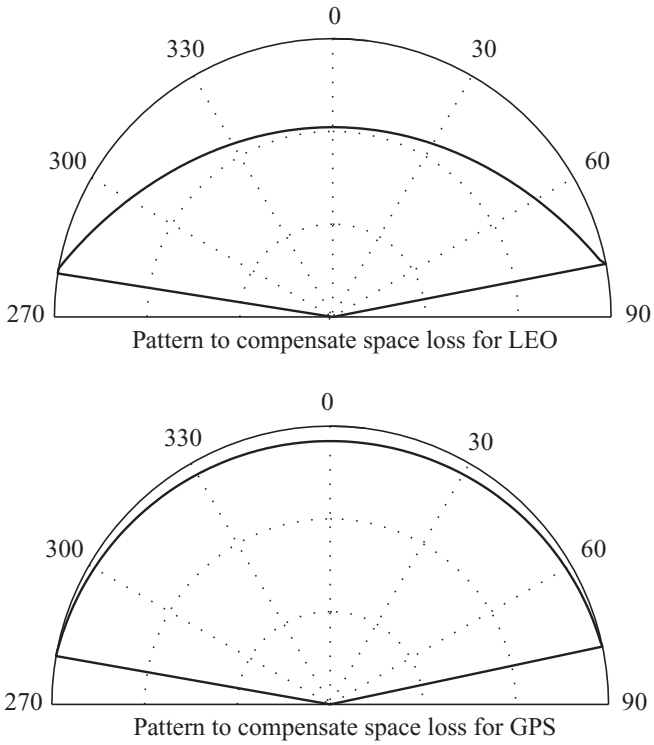


Figure 8.3.5 The directional function required of a mobile antenna for a satellite link, taken directly from the relative space loss. The patterns are truncated at 10° *above the horizon. The upper pattern is for LEO, the lower pattern for a GPS satellite. The polar scale is 10 dB per division*

8.3.4 Circularly polarised patch elements

Patches can be made to be circularly polarised by exciting similar modes in quadrature. The basic circular patch excited in the dominant TE mode is the starting point. Square patches will perform similarly, and the design process is similar. The mode of interest is the $n = 1$ mode in the modes which have field dependence $\exp(\mathrm{j}n\phi)$.

Using a convenient high-dielectric substrate, such as alumina ($\epsilon_\mathrm{R} = 9.9$), the patch approximate diameter, $2a$, is found from the first solution of

$$J_n'(k_\mathrm{S}a) = 0 \tag{8.3.10}$$

where k_S is the wavenumber in the substrate, to be less than a quarter wavelength. Although this is a conveniently small patch, the groundplane needs to be larger than this, extending past the patch by at least four to five times the height of the patch for the $n = \pm 1$ mode. (The $n = \pm 1$ modes are the only ones that can radiate at broadside; see below, and Section 8.4.) A feed position around a radial location of about $0.3a$ (recall that a is the radius of the patch) can give a 50 Ω impedance, as discussed in Section 8.1.

The height of the patch influences the bandwidth and has a small bearing on the centre frequency. As a narrow band antenna, the cavity structure also acts as a filter which assists in reducing the received noise, so there is an interest in keeping the

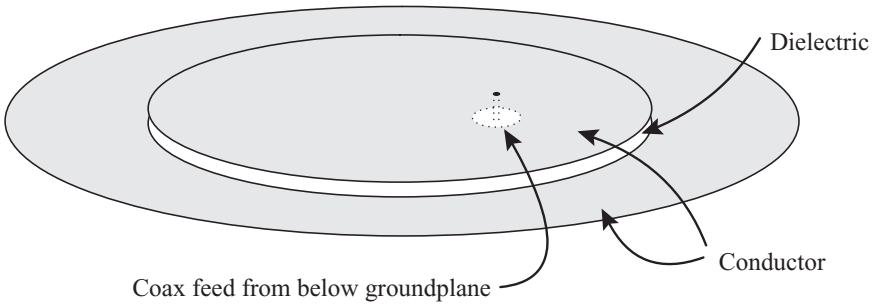


Figure 8.3.6 Basic circular patch antenna, here with substrate only under the conducting patch and a circular groundplane

patch low from the electrical, as well as the physical, viewpoint. Figure 8.3.6 shows the geometry, where a coax feed is used. The substrate can terminate at the edge of the patch, or extend as far as the edge of the groundplane. These alterations tend to shift the centre frequency because of the changed field fringing at the edge of the patch. Similarly, if the groundplane diameter is less than about a half free-space wavelength, then the feed impedance will be sensitive to variations in the size of the groundplane.

In practice, the final patch dimension and feed position are found with a few experimental iterations, normally required because the dielectric constant of the substrate is not known exactly, and because a small groundplane is difficult to design for. Sometimes a shorting pin is placed through the centre of the patch to suppress the $n = 0$ mode, but this is not required for narrowband design. For higher mode design ($n > 1$, see Section 8.4), the shorting pin is normally thicker and this is helpful for modal purity.

For circular polarisation, another $\cos\phi$ mode is generated in quadrature and orthogonal to the first mode. In a perfectly circular patch, a single feed located at $\phi = \phi_0$ will excite a mode with field dependence $\cos(\phi - \phi_0)$. Another feed at $\phi = \phi_0 + \pi/2$ produces the same mode, but displaced in azimuth by 90°, i.e. the field dependence is $\sin(\phi - \phi_0)$. By putting a $\pi/4$ phase shift in this feed, and adding the two feeds together, the wanted mode is produced; $\cos\phi + \exp(\mathrm{j}\pi/4)\sin(\phi) = \exp(\mathrm{j}\phi)$.

For the other hand of circular polarisation, either the phase shift is changed to $-90°$ or the azimuthal feed position of the second feed is shifted to a position of $+90°$ relative to the first feed. In practice, a dual feed system is unnecessary if the antenna is for narrowband operation. The second mode comprising $\sin(\phi - \phi_0)$ (azimuthally displaced by 90° from the $\cos(\phi - \phi_0)$ mode) can be produced in the correct amplitude and phase by perturbing the patch shape. This technique was pioneered by Haneishi (Chapter 4 in James and Hall, 1989) using extra strips of conductor at the patch edges, and alternatively strips of missing conductor on the patch edges.

Figure 8.3.7 shows a pattern measured from a circular patch. Co- and cross-polarisations have been measured, showing the cross-polar level to be over 20 dB down at zenith (corresponding to an axial ratio of less than 1.5 dB) and becoming decreasingly purely polarised towards the horizontal directions. So at zenith the power lost due to polarisation impurity (the loss is measured by the polarisation efficiency,

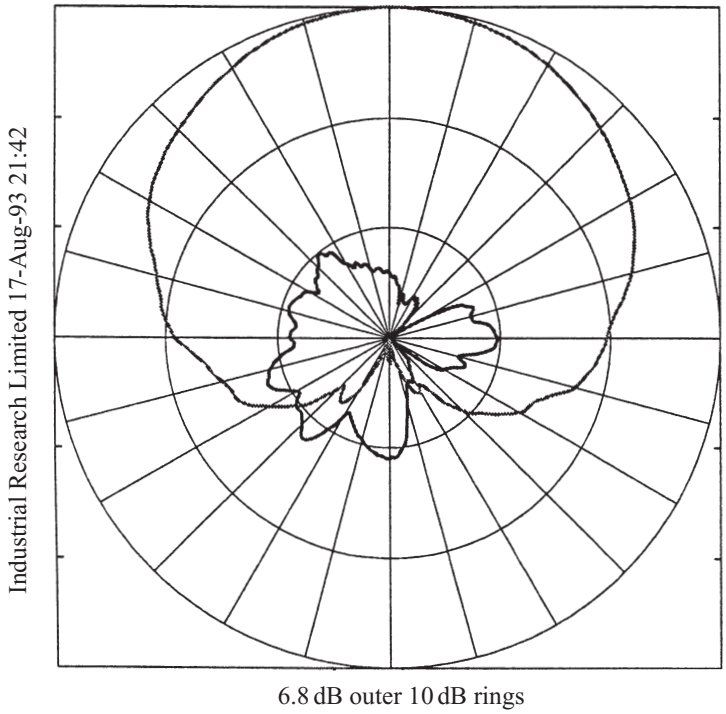


Figure 8.3.7 Measured co- and cross-polar patterns for a compact resonant patch antenna for GPS reception. There is a single feed to an elliptic patch, positioned at 45° with respect to the elliptic axes, and at a radius of $0.026\lambda_0$,giving a VSWR of less than 1.2 at the GPS frequencies. The patch has major axes $0.1812\lambda_0$ and $0.1784\lambda_0$, respectively, and height $0.0013\lambda_0$. The zenith gain is 6.8 dB, and the full scale is 30 dB. The two patterns are measured using RHCP and LHCP testing antennas. The circular polarisation ratio is over 20 dB at zenith and the co-polar gain is 6.8 dB

above) is negligible. In this design, the patch has single feed with an elliptic shape for the excitation of both modes with sufficient accuracy for good polarisation purity.

8.3.5 Distributed directivity for mobile-to-geostationary satellites

Fixed beams from antennas on a vehicle for geostationary satellites ideally require a conical shaped beam which should of course be circularly polarised. An idealised pattern form is shown in Figure 8.3.8. The finite beamwidth is required because of the pointing variations from the motion of the mobile terminal. However, such a pattern can only be approximated owing to the finite aperture of any practical antenna (Vaughan and Scott, 1993).

The directivity limitations on such a fixed beam are easily found from the conical truncation on a sphere. The rotationally symmetric directivity over a beamwidth of

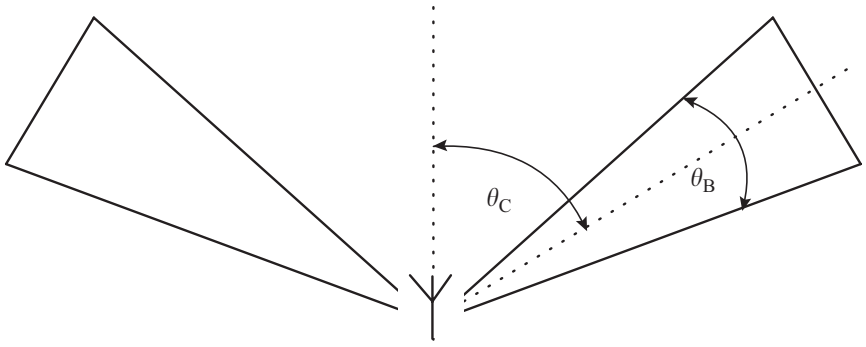


Figure 8.3.8 Cross section of a rotationally symmetric conical pattern – the idealised pattern for a mobile antenna for linking to a geostationary satellite

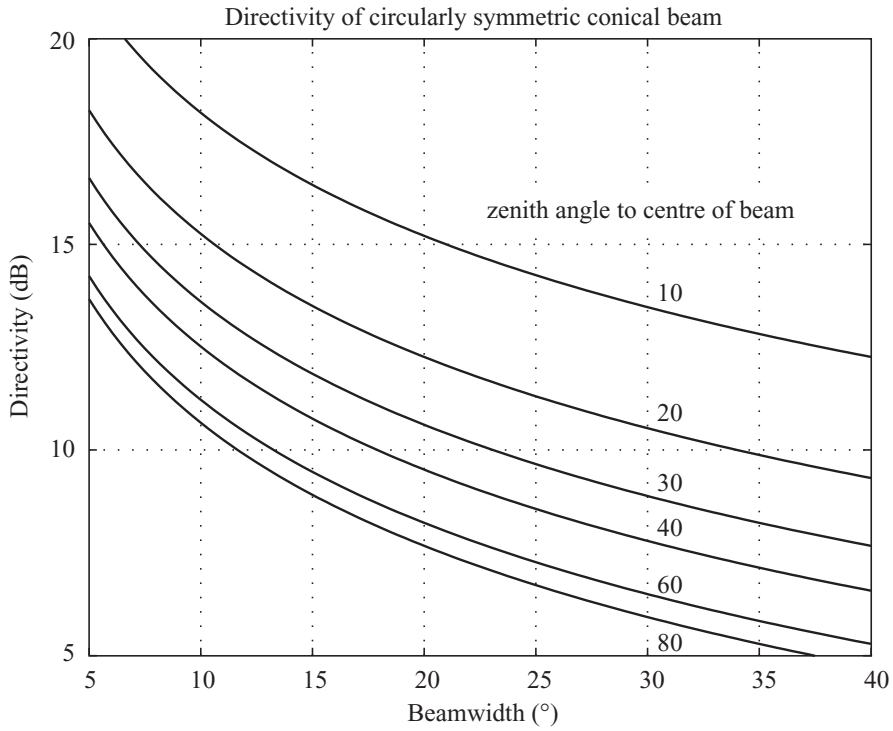


Figure 8.3.9 The directionality of the conical beam for mobile to geostationary satellite link

angle θ_B, centred at zenith angle θ_C, is found from the relations in Section 8.1 as

$$D(\theta_C, \theta_B) = \operatorname{cosec}\theta_C \operatorname{cosec}\left(\frac{\theta_B}{2}\right). \tag{8.3.11}$$

This is plotted in Figure 8.3.9. In this case, the directivity, uniformly covering the cylindrical cone area, is the same as the single-direction directivity. The directivity obviously increases with decreasing θ_B and increases with increasing θ_C. For example,

a beamwidth of 35° centred at 70° elevation (zenith angle $\theta_C = 20°$), has the constant directivity of 10 dBi. (The i here refers to a full sphere, isotropic reference.) This represents a limit on the directivity irrespective of the antenna implementation.

8.3.6 Circularly polarised, scanning-mode helix

In practice, the directivity of an antenna will be much lower than the theoretical limits above, which would need an infinitely large antenna. The effect of using a finite vertical cylindrical aperture for the antenna can be treated independently of the antenna construction details. A scanning mode helix can have a vertical, relatively thin cylindrical profile, and by correct construction and feeding, can have excellent circular polarisation purity. The mode of interest for the fields on the cylinder is the $\exp(j\phi)$ mode. It is a surface-wave antenna (Collin and Zucker, 1969) and the surface wave approach to the helix (Vaughan and Andersen, 1983) provides a far-field expression in terms of the antenna length L and surface-wave slowness factor τ;

$$A(\theta, \tau, L) = \frac{1 - e^{jk_0\ (\cos\theta - \tau)L}}{jk_0\ (\tau - \cos\theta)} \tag{8.3.12}$$

which is effectively an array factor. The slowness factor corresponds to the inverse of the speed of the axial direction surface wave on the structure, relative to the speed of light, $\tau = c/v_{\text{axial}}$.

The scanning mode corresponds to a fast wave, and the slowness factor is in the order of 0.7. It is controlled by the helix design. Losses on the helix structure can be simulated by introducing a finite imaginary component to the slowness factor. This decreases the gain slightly and fills in the pattern nulls considerably. The diameter of the structure ($2a$) has little influence over the pattern, but practical designs have a diameter of about one tenth of the wavelength.

The analogous element factors for the linear polarisations for the $n = 1$ mode are given by

$$F_\theta\ (\theta) = Z_0 H_z \cos\theta (J_0 + J_2) + jE_z\ (J_2 - J_0) - 2Z_0 H_\phi \sin\theta J_1 \tag{8.3.13}$$

$$F_\phi(\theta) = jE_z \cos\theta (J_0 + J_2) - jZ_0 H_z (J_2 - J_0) - 2E_\phi \sin\theta J_1 \tag{8.3.14}$$

where $J_n = J_n(k_0 a \sin\theta), ka$ is the electrical circumference of the helix, E_z, E_ϕ, H_z, H_ϕ are the tangential fields on the helix cylindrical aperture, and Z_0 is the free space impedance, $120\pi\ \Omega$. For circular polarisation everywhere, the helix design is such that the surface fields have the relationship $\boldsymbol{E} = \pm jZ_0\boldsymbol{H}$ (Rumsey, 1966). The element pattern is the same in the forward and backward direction, and normally has somewhat higher gain in these axial directions than at broadside.

These array and element factors do not rigorously correspond to the physical array of several single turn helical elements. They come about due to a separation of the radiation integral into length dependent and length independent factors. Whereas the array factor is a function of the antenna length and surface wave speed, the element factor is a function of the fields on the surface. The slowness factor and the surface

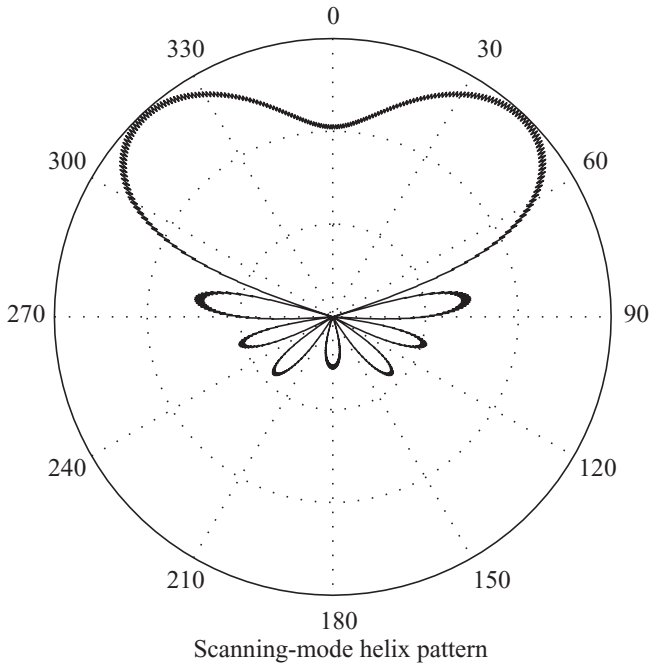


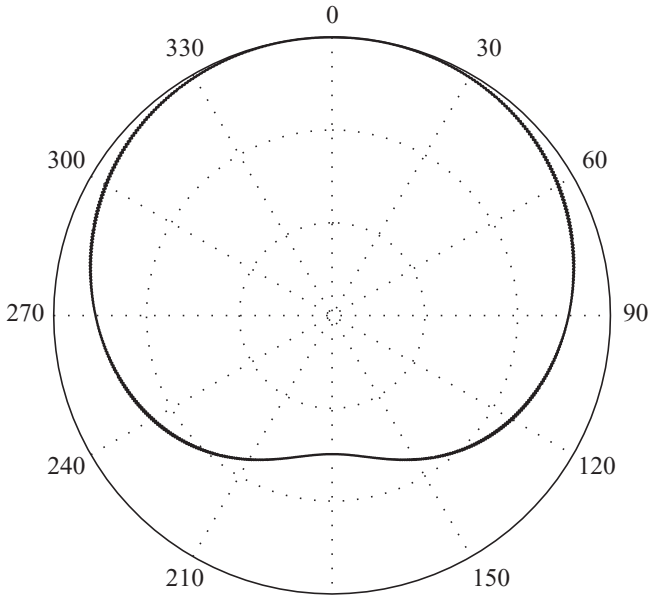
Figure 8.3.10 A pattern calculated for a vertically oriented, 2.5 wavelength long, scanning mode helix, suitable for mobile geostationary satellite links

fields are controlled by the helix design and excitation. The element factor gives the polarisation, so the polarisation is independent of the length of the antenna according to this approach (Vaughan and Andersen, 1983).

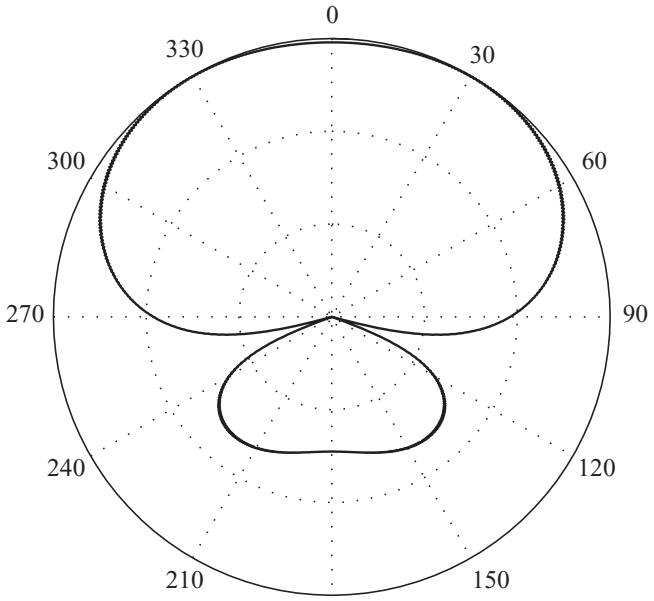
An example of a pattern is given in Figure 8.3.10. The length is 2.5 wavelengths, and the slowness factor is $\tau = 0.7 + \mathrm{j}0$, and the surface field relationships have been made imperfect ($E_z \neq \pm \mathrm{j} Z_0 H_z$, $E_\phi \neq \pm \mathrm{j} Z_0 H_\phi$) so that there is some cross-polarisation. The pattern simulates that measured by a spinning dipole, and so the axial ratio, which here is about 1 dB, indicates that the cross-polarisation is some 25 dB down.

Patterns calculated using these equations agree very well with those found from a moment method approach for a thin wire helix. Using a straight wire formulation, the directivity calculation converges using eight (or more) straight filaments per turn (octagonal helix).

For optimum performance, there is an optimum length for a given distributed directivity. From a moment method analysis, an example is for a beam centred at a 56° zenith angle, and for beamwidths of (8°, 10°, 12°, 14°, 16°), the optimum lengths in wavelengths are about (6.8, 6.0, 5.0, 4.0, 3.5) respectively. For lengths shorter than a wavelength or so, not much pattern control can be exercised, since the beam now has low directivity. Figure 8.3.11 shows a couple of examples for low-gain array patterns. Figures 8.3.10 and 8.3.11 give a feel for matching the practicable patterns to ideal patterns.



Perfectly polarised, length = 0.5 wavelengths



Perfectly polarised, length = 1 wavelength

Figure 8.3.11 Array patterns for scanning helices of length 0.5 wavelengths (top) and 1 wavelength (bottom)

8.3.6.1 Radiation mechanism of scanning-mode helix

A note is in order regarding the radiation mechanism of the scanning-mode helix. The radiation is from a leaky wave and is a backfire radiation, meaning that the radiation is in a direction backwards from the propagation of the surface wave. So for the radiation to be in the direction of the upper hemisphere (with a vertically mounted antenna) the

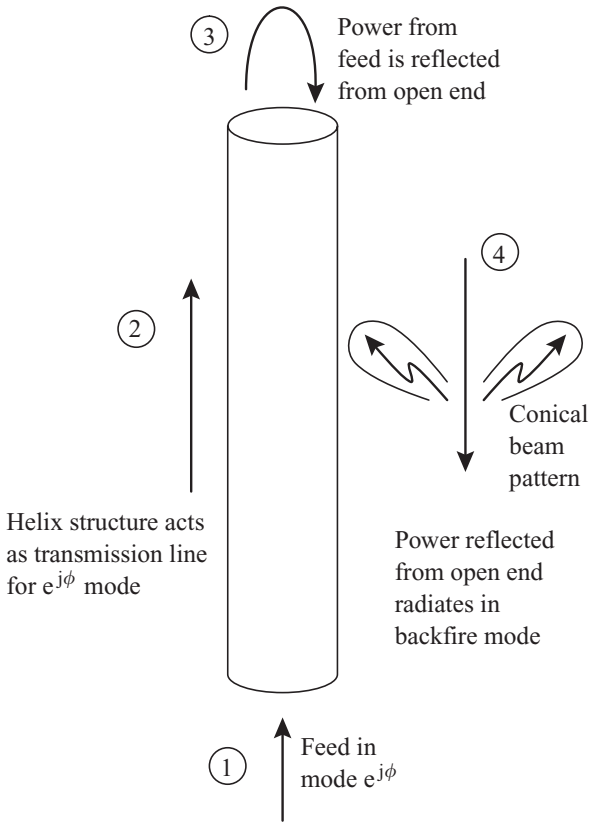


Figure 8.3.12 The mechanism for radiation from a bottom-fed scanning-mode helix antenna. The cylinder is the aperture on which the helix is implemented

leaky wave is travelling towards the feed end (bottom). The radiation is therefore not strongly dependent on the feed details such as whether there is a groundplane or not.

The mechanism is depicted in Figure 8.3.11 (Vaughan *et al.*, 1994). At the feed end the surface wave is launched and the structure of the helix supports a bound, non-radiating wave. The transmission line action of the helix is akin to a wave on twisted pair. At the open end of the surface wave antenna, the helix conductor currents are reflected. The surface wave returning towards the feed end 'sees' a different structure in the sense that the $n = 1$ mode is no longer supported as a bound, non-radiating wave. Instead, the wave is a leaky wave and the radiation occurs in the backfire direction. Further reflections from each end are quite small and most of the radiation is from the first pass of the leaky wave. For the helix structure to support the $n = 1$ mode in this way, there should be three or more conductors, i.e. a trifilar or quadrafilar helices, for example. The bifilar helix presents the same structure to the wave travelling in either direction along the antenna in the sense that the structure cannot discriminate between the $n = +1$ and $n = -1$ modes. The radiation patterns from the bifilar structure are similar but the radiation occurs in both the forward and backward

directions and symmetrical 'butterfly' patterns result in the elevation pattern cuts. The gain is halved as a result of this.

The scanning-mode helix can also be fed from the top by passing a feed line through the centre of the cylindrical structure or by using the conductors themselves (implemented using coax line for example) in an infinite balun structure.

An interesting beam steering possibility arises from exciting the $n = +1$ mode or $n = -1$ mode at the feed end. These will radiate in different hemispheres. So by phasing a trifilar helix conductor for example as 0°, 120°, 240° ($n = 1$ mode), the helix radiates in the upper hemisphere and if excited as 0°, 240°, 120° ($n = -1$ mode), the radiation pattern is a similar shape but is in the lower hemisphere.

Some examples of patterns are given in Figure 8.3.13 for an adjustable helix. These demonstrate the radiation direction control by altering the helix structure, for example by having the conductors wound using a dielectric sheet 'roller blind', or wound on a flexible dielectric (Vaughan *et al.* 1992, 1994). The aperture length is $4.4\lambda_C$ and the diameter is $0.055\lambda_C$. A series of patterns at different frequencies demonstrates the frequency-scanning behaviour of the antenna, and the adjustable main beam directions. With these patterns, the cross polar component is about 15 dB below the co-polar component in the main lobe region.

A problem with scanning-mode operation is that, as the name suggests, the beam scans as the frequency changes, shown in the above plots. For short antennas, it is clear that the beamwidth degenerates into a broad, single lobe, and the beam scanning is of little consequence. For longer antennas with directive beams, however, it turns out that across a typical bandwidth requirement, the position of the beam can move away from the target direction. A structural technique for mitigating the scanning action is described in du Toit *et al.* (1999) in which the frequency scanning is reduced from 8.6° to 1° over a 6.3 per cent bandwidth.

8.3.7 Fan patterns for vehicular mobile satellite

An alternative to the fixed, rotationally symmetric patterns above, is to create a 'fan' pattern, which is not rotationally symmetric. It is fixed in its elevation beam, but can track by pattern rotation in azimuth as the vehicle turns. The advantage of doing away with the rotational symmetry is that the directivity can be increased. However, the complication of a beam steering requirement is introduced. In one example, the steering is implemented mechanically and the otherwise fixed antenna can be implemented using a Yagi-type antenna with patch elements (Huang and Densmore, 1991).

8.4 Diversity antennas

8.4.1 Introduction

The use of antenna diversity for fading channels has the advantage of improving channel quality for a given radio spectral allocation, relative to that available from a single-port antenna. In this section, the improvement of receive diversity is couched

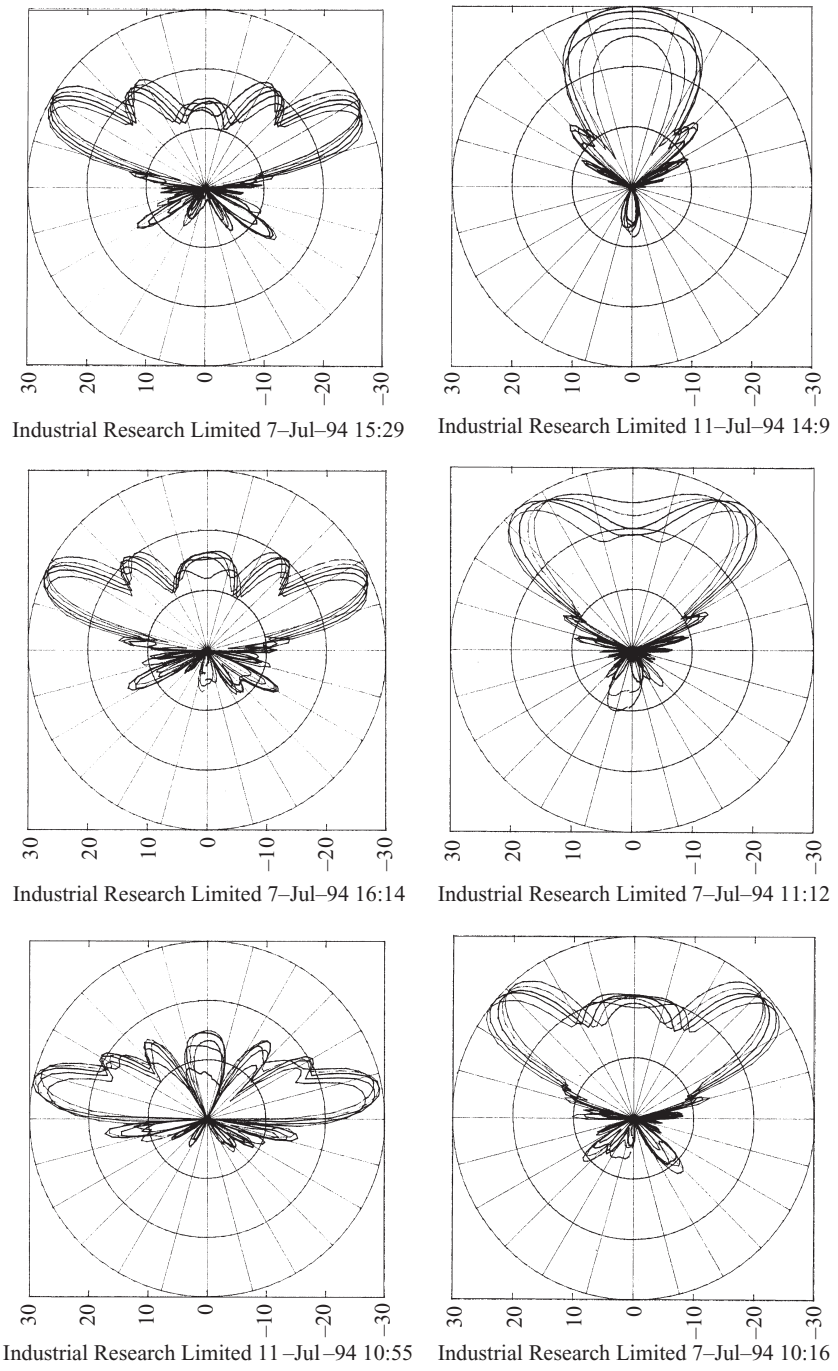


Figure 8.3.13 *Measured co-polar patterns from a scanning mode helix showing the beam direction control by varying the pitch angle of the conductors. The patterns are elevation angle cuts and are rotationally symmetric (not shown). From top left: number of turns of helix: 6, 7, 7.5, 8, 9, 9.5. The directivity for the dual lobe patterns is about 9 dB. Each plot has a set of patterns with frequencies:* $0.97 f_C$; $0.985 f_C$; f_C; $1.015 f_C$ *and* $1.03 f_C$. *The patterns scan towards the broadside direction as the frequency increases. The scale is 10 dB per division*

in terms of the improvement in signal-to-noise ratio (SNR) relative to the best single branch performance. The improvements in system spectral efficiency follow from this, but can be complicated to estimate because they also require statistical models of the communications system, multi-user traffic, outage criterion and grade-of-service parameters, etc. Transmit diversity can have similar benefits and joint transmit–receive diversity as well. Transmit–receive diversity and the capacity improvements are looked at in Chapter 9. The basic requirement for efficient diversity action is to have uncorrelated antennas with equal gain. How these requirements can be implemented is covered here after a review of basic combining principles.

8.4.2 Signal combining in diversity

8.4.2.1 Diversity gain

The *diversity gain* is defined here as the improvement in the time-averaged SNR from the combined signals of a set of diversity elements, relative to the SNR from the best single antenna element. The single element is sometimes a reference antenna, such as a dipole, but usually it is the best of the branches in the diversity set. The definition of diversity gain is conditioned by a probability that the SNR is above a reference level. There is no standard for the probability value, but it is often taken as 0.5 per cent.

Unless otherwise stated, the diversity discussed here is to counter the short-term, Rayleigh-like fading, sometimes called micro-diversity. Macro-diversity is the term for antenna diversity applied for countering the long-term, shadow fading. Here the antennas are spaced so far apart that the shadow fading is different at each branch. The short-term fading will also be uncorrelated for macro-diversity antennas.

Mathematically, the diversity gain can be written as

$$\left[\frac{\gamma_C}{\Gamma_C}\ (\text{dB}) - \frac{\gamma_1}{\Gamma_1}\ (\text{dB})\right]_{\text{Probability given}} \tag{8.4.1}$$

where γ_C is the instantaneous signal-to-noise ratio (SNR) of the diversity-combined signal, $\Gamma_C = \langle\gamma_C\rangle$ is the mean signal-to-noise ratio of the combined signal, γ_1 is the highest SNR of the diversity branch signals, and $\Gamma_1 = \langle\gamma_1\rangle$.

The diversity gain can be illustrated as the change in SNR on a time-series signal trace, and on a signal statistics plot. The following two diagrams present these cases.

Figure 8.4.1 illustrates the diversity effect with an example of combining eight uncorrelated, equal power Rayleigh channels. The combined signal is that of selection, equal gain, and maximum ratio combining. The 5 per cent diversity gain (illustrated with the bold arrow), estimated from this finite data is about 18.5 dB, close to the true value, or statistical expectation of 18 dB, given below.

Figure 8.4.2 gives the cumulative density function for the eight-branch selection-combined signal, and that of one of the branches. The diversity gain is read off the abscissa. This diversity gain is the true, or statistical mean value. A similar plot for maximum ratio combining is given, in Figures 8.4.6 and 8.4.7.

In multi-user systems, the capacity is often interference-limited, rather than noise-limited. In such systems, it is more appropriate to define a diversity gain as the

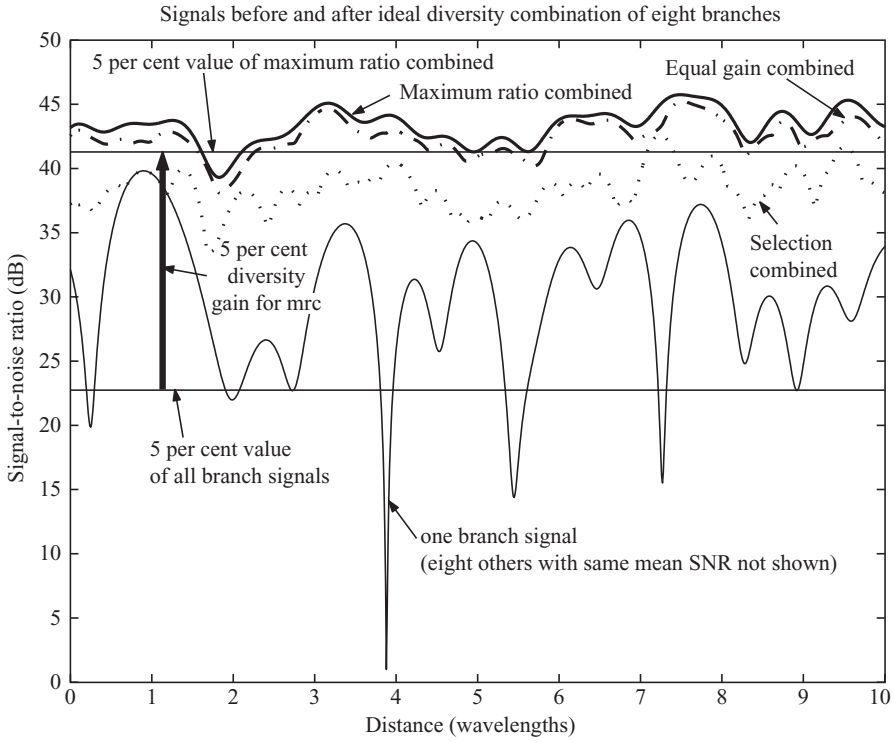


Figure 8.4.1 The 0.05 probability diversity gain is the difference between the 5 per cent probability levels of the single branch signal and the combined signal. The signal combining is ideal with only one of the eight diversity branch signals shown

improvement in signal-to-noise-plus-interference ratio (SNIR), or simply the signal-to-interference ratio, when the noise can be neglected. By optimising reception of the wanted signal power during combination, the SNIR is improved anyway, since the fading of the interferers will be uncorrelated with that of the wanted signal. However, further improvement is available using optimum combining, in which the wanted signal and the interferers are detected and processed. Optimum combining is expensive in hardware and power consumption to implement, but offers improvements that single-signal receivers cannot attain (see Chapter 9).

8.4.2.2 Diversity algorithm performance from distributions of the branch signals

Mathematically, the starting point for diversity gain analysis is a joint distribution between the branch signals. As noted in the previous chapters, the traditional analysis of antenna diversity relates to the signal envelopes.

For two branch signals, for example, with complex signals s_1, s_2, the joint distribution between the envelopes, $r_1 = |s_1|, r_2 = |s_2|$, is written $p_{r_1,r_2}(r_1, r_2)$. For most situations, we strive for independent branch signals, stated as $p_{r_1,r_2}(r_1, r_2) = p_{r_1}(r_1)\,p_{r_2}(r_2)$, which normally simplifies the analysis and gives the best diversity gain. A combining algorithm mixes the branch signals so that the resultant envelope

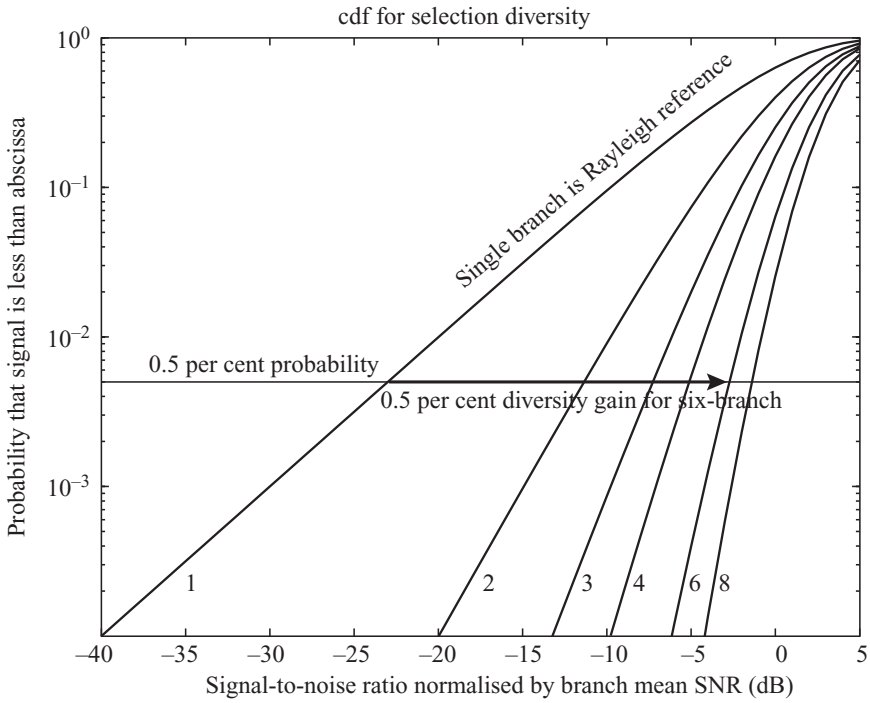


Figure 8.4.2 The cumulative density function for the signal-to-noise ratio for a single Rayleigh branch and up to eight-branch selection-combined signal. The diversity gain can be read off the abscissa for a given probability of signal level, the example here shown for the 0.5 per cent level and six branches. The branch mean powers are the same

is some function of the components, $r = f(r_1, r_2)$, and the probability density of the combined signal, $p_r(r)$, is found from function transformation (Appendix B).

8.4.2.3 Basic combining algorithms

The basic combining algorithms are:

- switching;
- selection;
- equal gain combining;
- maximum ratio combining.

The predetection maximum ratio and equal gain combining techniques are depicted in Figure 8.4.3. These, and switching combinations are discussed more below. The linear combining structure may be implemented as rf processes, or may be realised as digital processors.

The idealised algorithms, when applied to the signal envelopes, are expressed as follows. Here the M branches all have the same mean branch noise power and the combined signal envelopes are scaled so that the noise of the combined signal is the same as that of a single branch:

- equal gain combining: $r_C = r_1 + r_2 + \cdots + r_M/\sqrt{M}$;

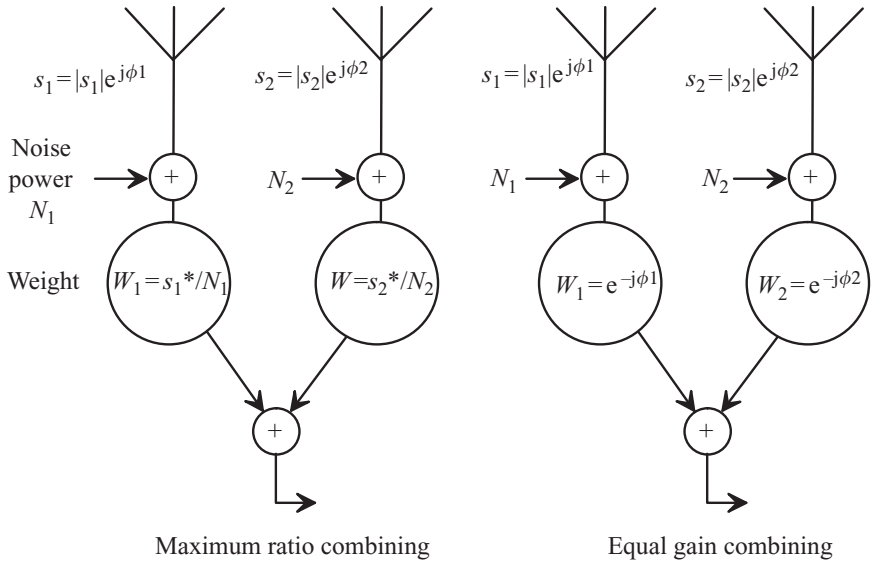


Figure 8.4.3 The equal gain and maximum ratio combining algorithms shown for predetection combining. In maximum ratio combining (left) the signals are phase aligned and have an amplitude weighting proportional to the ratio signal voltage over the noise power. In equal gain combining (right) the phases are aligned and the amplitude weights are all the same

- maximum ratio combining: $r_C = \sqrt{\sum_{m=1}^{M} r_m^2}$;
- selection: $r_C = \max(r_1, r_2, \ldots, r_M)$.

In implementing these algorithms in a real-world situation, the signal combination is always less than perfect. This is especially the case in predetection combining (see below) where the signals must be phase-aligned before combination, so a phase estimate is required in each branch. In maximum ratio combining, the weighting relies on an estimate of the signal magnitude as well. In practice, the phase and amplitude estimates suffer inaccuracies because they are based on a finite number of measured samples of the available signal which comprises the wanted signal, the noise and interference.

For statistical analysis it is usually more convenient to deal directly with the real and imaginary components of the complex signal, or the power signals, rather than the envelopes. Use is also made of the characteristic functions in the process of finding the distributions of combined signals, as shown below.

With antenna diversity for mitigating the short-term fading, we deal with Rayleigh distributed envelopes for the branch signals, i.e, exponentially distributed signal powers. Other fading distributions can be used, for example log-normal for macro-diversity.

The joint distribution between Rayleigh envelopes, $p_{r_1,r_2}(r_1, r_2)$, or exponential powers, $p_{\gamma_1,\gamma_2}(\gamma_1, \gamma_2)$, can be found from the multivariate Gaussian form, given in Appendix C, and is also discussed in Section 7.6. For independent Rayleigh envelopes,

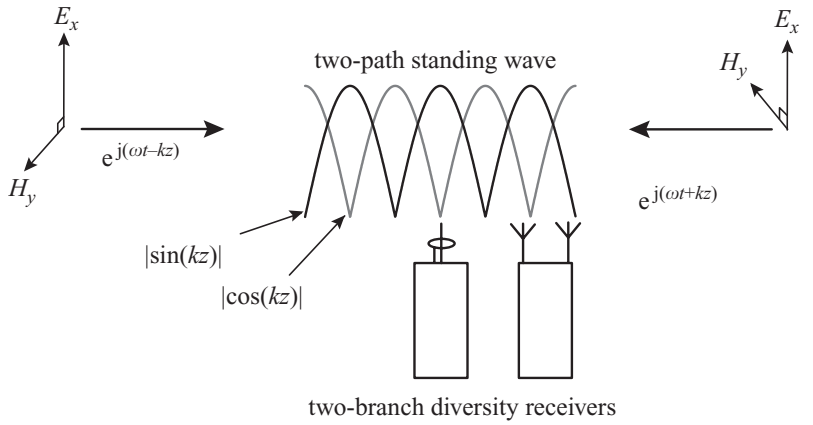


Figure 8.4.4 The special case of two incident waves from opposite directions gives a standing wave pattern that features negative correlation values. The standing wave pattern can be received by diversity antennas and combined, in this case, to give no fading. This type of fading is unlikely to occur in reality over significant areas

all the cross-correlation coefficients are zero. It is also convenient to have equal mean branch powers, since the analysis further simplifies, and the diversity gain is maximised since we are not having one branch (which has the higher SNR and provides the reference performance) dominating the combined signal.

Recapping the situation for Rayleigh distributed envelopes (see Appendix C), the branch signals are written as $s = x + \mathrm{j}y$, where x and y are independent (i.e. uncorrelated) Gaussians, the relation between the correlation coefficients for the power signals (envelope squared), the envelopes, and the in-phase/quadrature components, is, for the autocorrelations,

$$|\rho_s|^2 = \left|\tfrac{1}{2}\left(\rho_x + \rho_y + \mathrm{j}\left(\rho_{xy} - \rho_{yx}\right)\right)\right|^2 = \rho_x^2 = \rho_{|s|^2} \approx \rho_{|s|} \tag{8.4.2}$$

and similarly for the cross correlations,

$$|\rho_{s_1 s_2}|^2 = \left|\tfrac{1}{2}\left(\rho_{x_1 x_2} + \rho_{y_1 y_2} + \mathrm{j}\left(\rho_{x_1 y_2} - \rho_{y_1, x_2}\right)\right)\right|^2 = \rho_{x_1 x_2}^2 = \rho_{|s_1|^2 |s_2|^2} \approx \rho_{|s_1|\,|s_2|}. \tag{8.4.3}$$

For making correlation estimates from finite samples, it is better to calculate using complex numbers if the complex signal is available, rather than calculate using the envelopes only; see Section 7.1. Further below, the correlation coefficient is approximated using antenna parameters, in particular the patterns, and under certain conditions, the mutual coupling.

8.4.2.4 Negatively correlated signals

In some situations, it is possible to use correlated branch signals – where the correlation coefficients are negative – for larger diversity gains than occurring for

uncorrelated signals. In this situation, as one branch signal fades, the other is more likely to be 'unfaded' than for the case of uncorrelated branches.

In practice, envelope and power correlation coefficients are often estimated from real-world signals to have negative values. Assuming that the estimation procedure is sound (appropriate number of samples for the required accuracy of the estimate, etc.), the negative values indicate that the envelopes are not truly Rayleigh distributed, since the minimum (of the true value) envelope correlation for Rayleigh signals is zero.

As an example to illustrate the negative correlations, consider the situation where there are dominant effective sources in front and behind a moving terminal. This can happen for a vehicular terminal in a street of high-rise buildings where the dominant propagation directions tend to be along the street, or it could be more artificially arranged using a directional (front and back directions only) mobile antenna. To demonstrate the effect, we exaggerate the situation by neglecting other effective scatterers and making the incoming waves with equal amplitudes. So the signal at the mobile is deterministic with form $s(z) = \exp(\mathrm{j}k_0 z) + \exp(-\mathrm{j}k_0 z)$ (here z is the forward/backwards direction), and the standing wave pattern, or the envelope of the received signal, is $|s(z)| = |\sin k_0 z|$. The situation is shown in Figure 8.4.4, where the standing wave is shown for the vertical electric and horizontal magnetic field.

By spacing antennas along the direction of the periodic standing waves at a separation distance of a quarter wavelength, the received power signals are $|s_1|^2 = \sin^2 k_0 z$, $|s_2|^2 = \cos^2 k_0 z$, the power correlation coefficient is $\rho_{|s_1|^2|s_2|^2} = -1$ (the envelope correlation can be calculated to be 0.92), and the summed power from these two branches is a constant. So by properly combining just two deep-fading branches, a diversity signal combination can be created that has no fading! It is emphasised that this is the result of the deterministic, correlated branch signals of this example, and the situation is unlikely to occur consistently in real-world signals.

8.4.2.5 Selection and switched combining techniques

In *selection combining*, the branch with the highest quality signal is chosen from the candidate branches, and this branch is then used for a period of reception processing. All the branch signals must be monitored in some way for this type of algorithm to be implemented, although the samples need not be taken from the branches simultaneously. This branch quality can be derived from a measure of signal power, *SNR*, or *BER*, etc., depending on receiver system details. A common method is to choose based on the signal power, which in practice is sampled from the signal-plus-noise. This is inexpensive to implement, and approximates the *SNR* except for low *SNR*s. Typically, the branch monitoring is undertaken between synchronised (with the transmitter) bursts of communication using the previous best choice of branch. At the end of the burst, the receiver stops the signal detection action, and commutates through the available diversity branches, sampling and storing samples of the channel quality, and comparing these. For the following burst period of communication, the best branch from this sample set is used. This commutation and selection technique is also called *scanning diversity*. It is depicted in Figure 8.4.5.

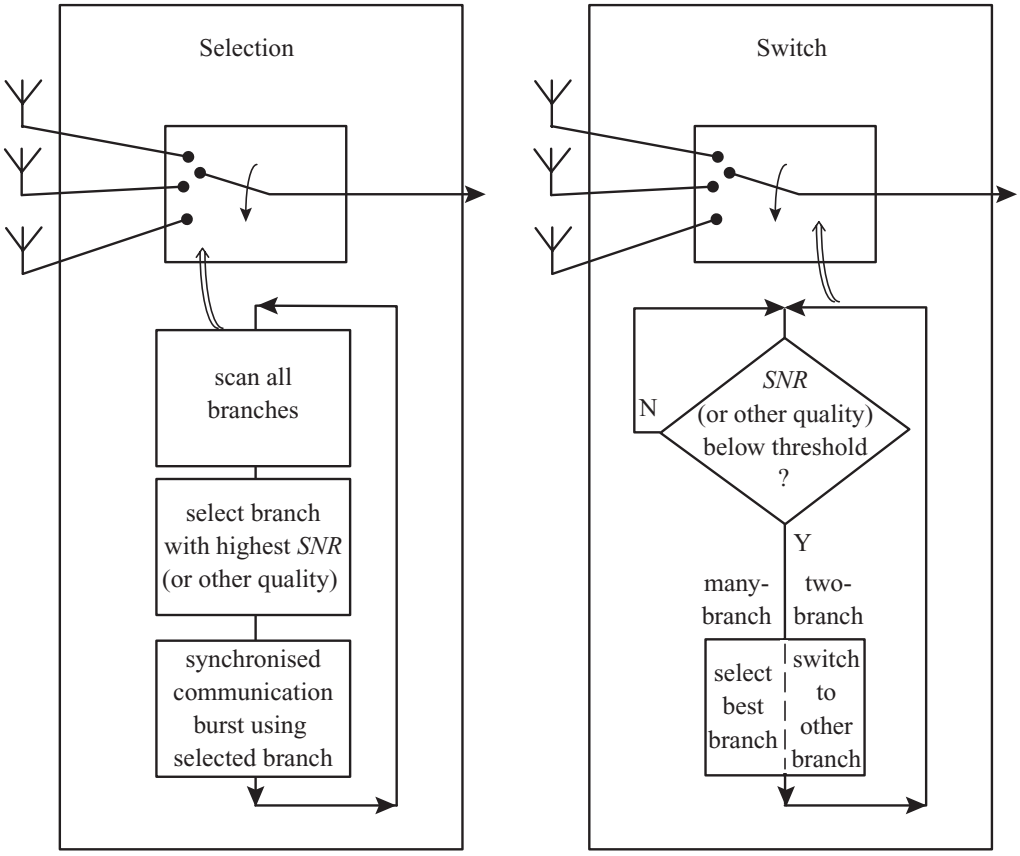


Figure 8.4.5 Selection (left) and switched (right) combining methods for diversity. In the selection algorithm, all the branches are monitored regularly to allow a choice of antenna branch, and the switch is made either with a synchronised scheme or asynchronously as required. In switched combining, a channel quality threshold is used to trigger a change of antenna branch, and for more than two-branch diversity, the current channel quality of available branches is sampled in order to select the best one

For unsynchronised branch selection, a change of branch may have to be undertaken during the communications reception/detection process, and so will normally feature a phase jump in the combined signal. This may create synchronisation problems and cause, for example, a bit- or frame-error, although this bit can be tagged as unreliable within the receiver/decoding scheme since the receiver knows that a branch selection change has been made at this instant. Soft switching (a gradual change of mix of the signals from one branch to the other) for averting the sudden phase transition will assist this problem. Another technique is to co-phase the signals before selection combination. However, arranging the co-phasing action may account for most of the processing required for branch combination, and once the branch signals are co-phased, simple addition of the signals, or equal-gain combining, is simpler than selection, and gives a better diversity gain. In unsynchronised selection diversity, the problem also arises of how to decide when to sample and select from the other

branches without missing communications data. However, in several current systems, in particular TDMA types (e.g. GSM, UMTS) with bursts of framed data and pilots, there is time to select or switch between the used time slots and phase synchronise.

A compromise for implementing selection-type combination implementation is *switched combining*, where selection of a new branch occurs when the current branch quality falls below a threshold. One switching strategy is the same as described above: to scan and compare the available channels and choose the best channel; see Figure 8.4.5. The performance is described below.

8.4.2.6 Predetection and postdetection in non-switched combining: equal gain

In non-switched type combiners, the signals from all the branches are used simultaneously. A basic combining classification is whether the action is predetection or postdetection. In predetection combining, the carrier phases of the branches are aligned before linear combination of the branch signals. So the signals, which have aligned phases, add in voltage, and the noise, which is assumed uncorrelated between the branches, adds as powers. For postdetection combining, the phase is removed from the branch signals in the detection process, and then the signals are linearly combined. It follows that the powers add for both the signal and the noise in this case. This can be illustrated for simple addition, also called equal gain combining. Two branch signals, s_1, s_2 with $\langle|s_1|^2\rangle = \langle|s_2|^2\rangle$ and having noise powers $\langle|n_1|^2\rangle = \langle|n_2|^2\rangle$ combine to give *SNR*s estimated by averaging (indicated by $\langle\cdot\rangle$) as (Lee, 1982):

$$\begin{aligned} S\hat{N}R &= \frac{\langle|s_1+s_2|^2\rangle}{\langle|n_1|^2+|n_2|^2\rangle} = 2\frac{\langle|s_1|^2\rangle}{\langle|n_1|^2\rangle} \quad \text{(predetection)} \\ S\hat{N}R &= \frac{\langle|s_1|^2+|s_2|^2\rangle}{\langle|n_1|^2+|n_2|^2\rangle} = \frac{\langle|s_1|^2\rangle}{\langle|n_1|^2\rangle} \quad \text{(postdetection)} \end{aligned} \tag{8.4.4}$$

and it is assumed that processes such as the detection and phase alignment have not contributed significant noise to the receiver. So ideal predetection equal gain combining provides a 3 dB *SNR* advantage over postdetection equal gain combining for two similar branch signals, and similarly 4.7 dB for three branches, and so on. This is similar to the array directive gain (with no mutual coupling), discussed in Chapter 9, and see also Figure 8.4.20.

Postdetection equal gain combining offers an interesting benchmark performance because the combiner is very simply implemented – independent receivers with added baseband signals. Any predetection combination requires co-phasing of the carriers in each branch, and this normally represents a major aspect of the process, whether implemented as a carrier operation such as in a feedforward technique (Jakes, 1974) or as an operation on digitised complex baseband signals such as in the GSM and UMTS systems.

The combining algorithm of main interest for analysis is *maximum ratio combining*, because much theoretical progress for the combiner performance is possible and because it gives the best *SNR* performance from linear combination. For uncorrelated

branch signals, the combined signal *SNR* is simply the sum of the branch *SNR*s, and this makes performance quick and simple to evaluate.

8.4.3 Signal statistics of combined signals

This section reviews the cumulative distributions of combined signals from Rayleigh fading branches (see also Appendix C).

8.4.3.1 Maximum ratio

The combining algorithm is to multiply each branch signal by the conjugate of its instantaneous signal-to-noise ratio. The conjugate signal part of the weight acts to align the phase of the branch signal to a common reference between all the branches, and the amplitude weighting acts to emphasise the branches with the largest SNR. The noise voltage, n_i, is taken as having constant average power which is the same for all channels, $N_0 = \langle |n_i|^2 \rangle = E\{|n_i|^2\}$. For most implementations, the dominant noise is from the thermal contribution of front end electronics, and these are separate, identical devices between branches, so this assumption is sound.

In maximum ratio combining, the ith branch receives an instantaneous complex signal voltage of s_i, then the weight vector is s_i^*/N_0 (Figure 8.4.3), and the weighted branch signal is written $|s_i|^2/N_0 = \gamma_i$. For uncorrelated branches, a summation of these branch signals gives a combined signal which has an instantaneous SNR which is of course the sum of the instantaneous branch SNRs. Similarly, the average SNR of the combined signal is the sum of the average SNRs.

Even for maximum ratio combining, there is no simple expression for the combined signal SNR for the general case of many-branch, arbitrarily correlated and unequal mean branch powers. However, some special case results are available, and these are given below.

8.4.3.2 Uncorrelated branches with different mean powers

For uncorrelated branches, but with different mean powers, the correlation matrix of the maximum ratio weighted branch signals has eigenvalues which correspond to the mean powers in the branches, denoted $\Gamma_k = \langle \gamma_k \rangle = E\{\gamma_k\}$ for the kth branch. This correlation matrix refers to an expectation (estimated in the combining algorithm through averaging) over the fading signals of the diversity branches, and the correlation measure is with respect to this fading.

Using characteristic function results from Kac and Siegert (1947) and Turin (1960) for multivariate Gaussian variables, Stein (1966) gives a general expression for the probability density function of the combined signal-to-noise ratio in terms of the Laplace characteristic function. For M uncorrelated branches, the characteristic function is

$$\Phi(s) = \frac{1}{\prod_{k=1}^{M} (1 + s\Gamma_k)} = \frac{1}{\det(\boldsymbol{I} + s\boldsymbol{R})} \tag{8.4.5}$$

where here s is the Laplace-space variable (note s is used for denoting a signal, above), $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{R}$ is the $M \times M$ correlation matrix of the branch signals.

(For Fourier transforms, set $s = \mathrm{j}\nu$, and ν is the Fourier-space variable.) The Laplace transform gives a probability density distribution,

$$p_{\mathrm{MR}}(\gamma) = \frac{1}{2\pi\mathrm{j}} \int_{c-\mathrm{j}\infty}^{c+\mathrm{j}\infty} \frac{\mathrm{e}^{s\gamma}}{\prod_{k=1}^{M}(1+s\Gamma_k)} \mathrm{d}s, \quad c \geq 0, \tag{8.4.6}$$

and here c is real and sufficiently positive to allow the integration contour to close to the right of the poles. The integral is solved by the method of residues,

$$p_{\mathrm{MR}}(\gamma) = \frac{1}{2\pi\mathrm{j}} \sum \text{ residues of poles.} \tag{8.4.7}$$

For simple poles only (unequal mean branch powers), the result is

$$p_{\mathrm{MR}}(\gamma) = \sum_{n=1}^{M} \frac{\mathrm{e}^{-\gamma/\Gamma_n}}{\left(\prod_{k=1}^{M}\Gamma_k\right) \prod_{\substack{m=1 \\ m\neq n}}^{M} ((1/\Gamma_m) - (1/\Gamma_n))} \tag{8.4.8}$$

and this can be written as

$$p_{\mathrm{MR}}(\gamma) = \sum_{n=1}^{M} \frac{1}{\epsilon_n} \mathrm{e}^{-\gamma/\Gamma_n}, \tag{8.4.9}$$

where

$$\epsilon_n = \Gamma_n \prod_{\substack{m=1 \\ m\neq n}}^{M} \left(1 - \frac{\Gamma_m}{\Gamma_n}\right). \tag{8.4.10}$$

The cumulative probability density follows as

$$P_{\mathrm{MR}}(\gamma \leq x) = \sum_{n=1}^{M} \frac{1}{\epsilon_n} \left(1 - \mathrm{e}^{-x/\Gamma_n}\right) \quad \text{(unequal branch powers).} \tag{8.4.11}$$

These results do not hold for equal, or numerically close branch *SNR*s as can be seen from the equations. In the case of common SNRs, $\Gamma_k = \Gamma$, the residues of higher-order poles must be evaluated.

8.4.3.3 Uncorrelated branches with common mean branch powers

For the case where all the mean branch powers are the same, $\Gamma_k = \Gamma, \forall k$; the solution is available from the characteristic function of a chi-squared random variable with $2M$ degrees of freedom (see Appendix C, which includes the Rician case). This characteristic function can also be seen from Stein's general form above, as $\Psi(s) = (1 - s\Gamma)^{-M}$. The probability density function follows as

$$p_{\mathrm{MR}}(\gamma) = \frac{\gamma^{M-1}\mathrm{e}^{-\gamma/\Gamma}}{\Gamma^{M}(M-1)!} \tag{8.4.12}$$

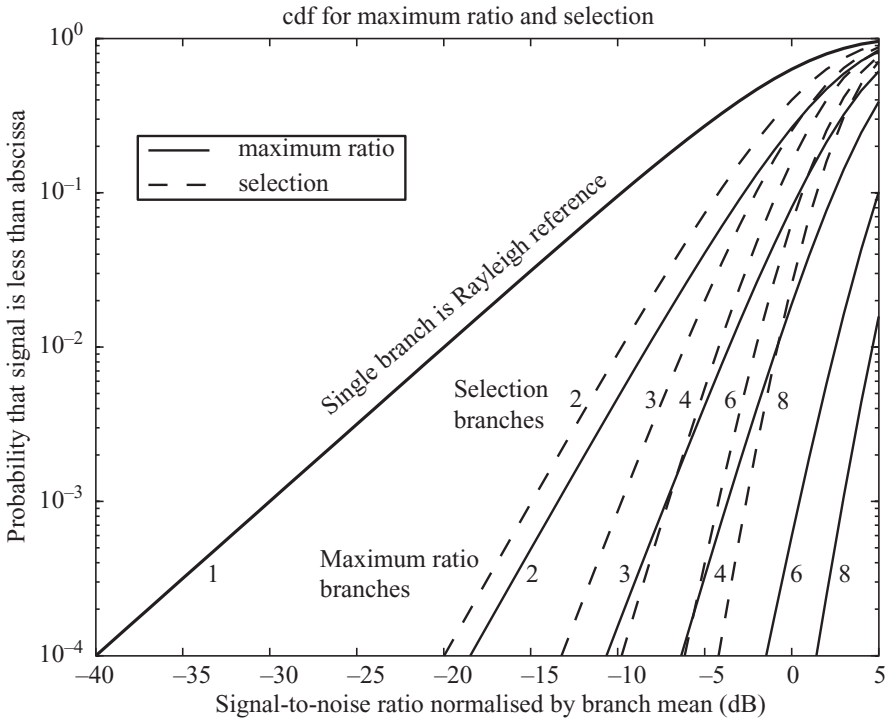


Figure 8.4.6 The cumulative density function for maximum ratio combining, with the values for selection combining (see Figure 8.4.2) for comparison. The difference between the diversity gains for selection and maximum ratio is approximately constant for a given number of branches

and this reduces to the exponential distribution for $M = 1$, a single branch system, as expected. The cumulative density follows by integration as

$$P_{\mathrm{MR}}\ (\gamma \leq x) = 1 - \mathrm{e}^{-x/\Gamma} \sum_{k=1}^{M} \frac{(x/\Gamma)^{k-1}}{(k-1)!} \quad \text{(equal branch powers)}, \tag{8.4.13}$$

and this is given in Figure 8.4.6 (cf. Figure 8.4.2 for selection combining) and some diversity gains calculated for maximum ratio combining are shown in Figure 8.4.7.

8.4.3.4 Selection combining

For ideal selection combining, the best branch is selected all the time. In practice, this may not be the case because once a selection is made, the channels may change, and the communications burst period may occupy a time when the branch is no longer the best choice. However, for a selection process which is invoked often relative to the rate of change of the branch channels, the assumption can hold.

From the cumulative probability of a single Rayleigh branch, given by

$$P_1\ (x) = \mathrm{Prob}\ (\gamma \leq x) = \int_0^x p_\gamma\ (\gamma)\, \mathrm{d}\gamma = 1 - \mathrm{e}^{-x/\Gamma_1}, \tag{8.4.14}$$

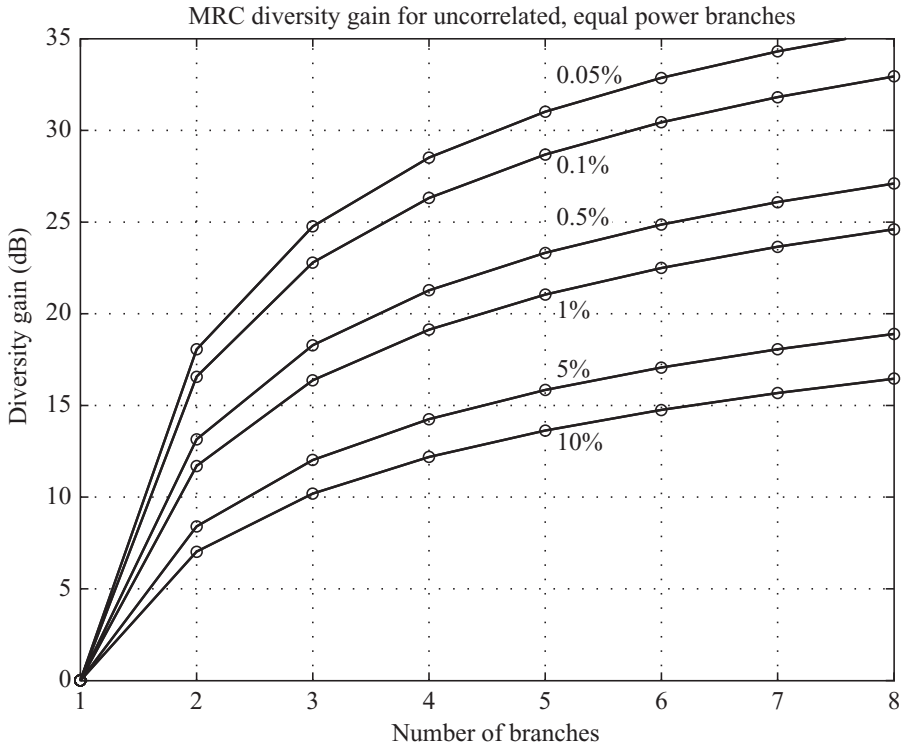


Figure 8.4.7 The diversity gain, shown here for maximum ratio combining, becomes large as the reference probability becomes small

the cumulative probability of M branches all being below some value x is

$$P_S(x) = \text{Prob}(\gamma_1, \ldots, \gamma_M \leq x) = \prod_{k=1}^{M} (1 - \mathrm{e}^{-x/\Gamma_k}) \\ = (1 - \mathrm{e}^{-x/\Gamma})^M \qquad \text{(equal branch powers).} \tag{8.4.15}$$

This is shown above, in Figure 8.4.2. The probability density function is found by differentiation, here for the equal mean branch power case, as

$$p_S(\gamma) = \frac{M}{\Gamma}(1 - \mathrm{e}^{\gamma/\Gamma})^{M-1}\mathrm{e}^{-\gamma/\Gamma} \quad \text{(equal branch powers).} \tag{8.4.16}$$

8.4.3.5 Diversity gain relation between selection and maximum ratio

The diversity gain is usually of interest for low levels of combined signal *SNR*, relative to the common branch mean; the relation $\exp(-\alpha) \approx 1 - \alpha, \alpha \ll 1$ allows the approximation for the cdf as

$$P_S(\gamma < x) \approx \frac{x^M}{\prod_{k=1}^{M} \Gamma_k}, \quad \gamma \ll \Gamma. \tag{8.4.17}$$

Similarly, for maximum ratio, here for equal mean branch powers,

$$P_{\mathrm{MR}}\ (\gamma < x) \approx \frac{1}{M!}\frac{x^M}{\Gamma^M}, \quad \gamma \ll \Gamma \quad \text{(equal branch powers)}. \tag{8.4.18}$$

For equal mean branch powers, the diversity gains of the selection and maximum ratio combinations can be equated

$$\left(\frac{x_{\mathrm{S}}}{\Gamma}\right)^M = \frac{1}{M!}\left(\frac{x_{\mathrm{MR}}}{\Gamma}\right)^M, \quad \gamma \ll \Gamma \quad \text{(equal branch powers)}, \tag{8.4.19}$$

so the ratio of the asymptotic forms of the diversity gains is

$$\frac{x_{\mathrm{S}}}{x_{\mathrm{MR}}} \approx \sqrt[M]{M!}, \quad \gamma \ll \Gamma \quad \text{(equal branch powers)}. \tag{8.4.20}$$

This provides a rule of thumb for the difference in diversity gain between selection combining and maximum ratio combining. For example, for maximum ratio diversity gain of 25 dB (with eight branches at 1 per cent probability; three branches at 0.05 per cent probability), the diversity gain using selection combining is about (5.7 dB; 2.6 dB, respectively) lower than using maxiumum ratio combining.

8.4.3.6 Equal gain combining

In (predetection) equal gain combining, the signals are co-phased and added; see Figure 8.4.3. The difference in diversity gain between the equal gain and maximum ratio turns out to be rather small. So equal gain combining offers most of the performance returns without having to amplitude-weight the signals.

There is no general form for the general case distribution of the equal gain combined SNR, although Lee (1982) gives approximate expressions for the small-value region of the pdf by introducing multiplicative factors which appropriately reduce the maximum ratio pdf. Equal gain falls between the MRC performance and the selection combiner performance. For example, for $M = 6$ uncorrelated branches with common average SNRs, the diversity gain at the 0.05 per cent level for perfectly implemented combiners is approximately: maximum ratio: 30 dB ; equal gain: 29 dB; selection: 25 dB.

8.4.3.7 Special cases of combining with correlated branches: two branch

For the two-branch case, with Gaussian correlation coefficient $\rho = \rho_{x1x2} = \rho_{y1y2}$, the correlation matrix is expressed in a form of signal-to-noise ratios

$$\boldsymbol{R} = \begin{bmatrix} \Gamma_1 & \rho\sqrt{\Gamma_1\Gamma_2} \\ \rho^*\sqrt{\Gamma_1\Gamma_2} & \Gamma_2 \end{bmatrix} \tag{8.4.21}$$

(note $\rho^* = \rho$) and the eigenvalues of this are

$$\begin{aligned} \lambda_1 &= \frac{\Gamma_2}{2}\left((1+\chi) - \sqrt{(1+\chi)^2 - 4\chi(1-|\rho|^2)}\right) \\ \lambda_2 &= \frac{\Gamma_2}{2}\left((1+\chi) + \sqrt{(1+\chi)^2 - 4\chi(1-|\rho|^2)}\right), \end{aligned} \tag{8.4.22}$$

where $\chi = \Gamma_1/\Gamma_2$. These eigenvalues can be used in the general expression for the pdf and integrated to get the cdf. For common mean branch powers,

$$\lambda_1 = \Gamma(1 - |\rho|), \quad \lambda_2 = \Gamma(1 + |\rho|) \tag{8.4.23}$$

and the cdf is, for $\rho \neq 0$

$$P_{MR}^{(2)}(\gamma < x) = 1 - \frac{1}{2|\rho|}\left((1 + |\rho|)\mathrm{e}^{-x/\Gamma(1+|\rho|)} - (1 - |\rho|)\,\mathrm{e}^{-x/\Gamma(1-|\rho|)}\right). \tag{8.4.24}$$

This expression can be used to investigate the performance trade-offs for having a finite correlation coefficient and/or the different branch powers. Such compromises from the ideal situation are normally necessary for realising compact diversity antennas. Stein (1966) makes observations from the asymptotic behaviour of the two branch function, as follows.

8.4.3.8 Equivalent equal branch powers for two correlated branches

For correlation coefficients which are not too close to unity, the equal branch powers, the equivalent uncorrelated two-branch mean power level is

$$\Gamma_{\text{eff}} \approx \Gamma\sqrt{1 - |\rho|^2}, \tag{8.4.25}$$

and $|\rho|^2$ is recalled as the envelope correlation coefficient. From the geometric mean of the branch powers above, the equivalent mean power of two, equal mean power, uncorrelated branches is simply

$$\Gamma_{\text{eff}} \approx \sqrt{\Gamma_1\Gamma_2}. \tag{8.4.26}$$

Combining these results gives

$$\frac{\Gamma_2}{\Gamma_1} \approx 1 - |\rho|^2 \approx 1 - \rho_r. \tag{8.4.27}$$

This offers a rule-of-thumb for the trade-off between branch correlations and unequal branch powers. For example, for equi-power branches with an envelope correlation coefficient of $\rho_r = 0.7$, a similar diversity gain is available from uncorrelated branches with a power difference of about 5 dB.

8.4.3.9 Symmetric three-branch case

The special case for three branches is also possible, where the correlation coefficient is the same between all three branches. This occurs in a symmetrical three element circular array. The result is (Vaughan and Andersen, 1987a)

$$P_{\text{MR}}^{(3)}(\gamma < x) = 1 - \frac{1}{(3|\rho|)^2}\left(P_1\,\mathrm{e}^{-x/(\Gamma(1+2|\rho|))} - P_2\,\mathrm{e}^{-x/(\Gamma(1-|\rho|))}\right) \tag{8.4.28}$$

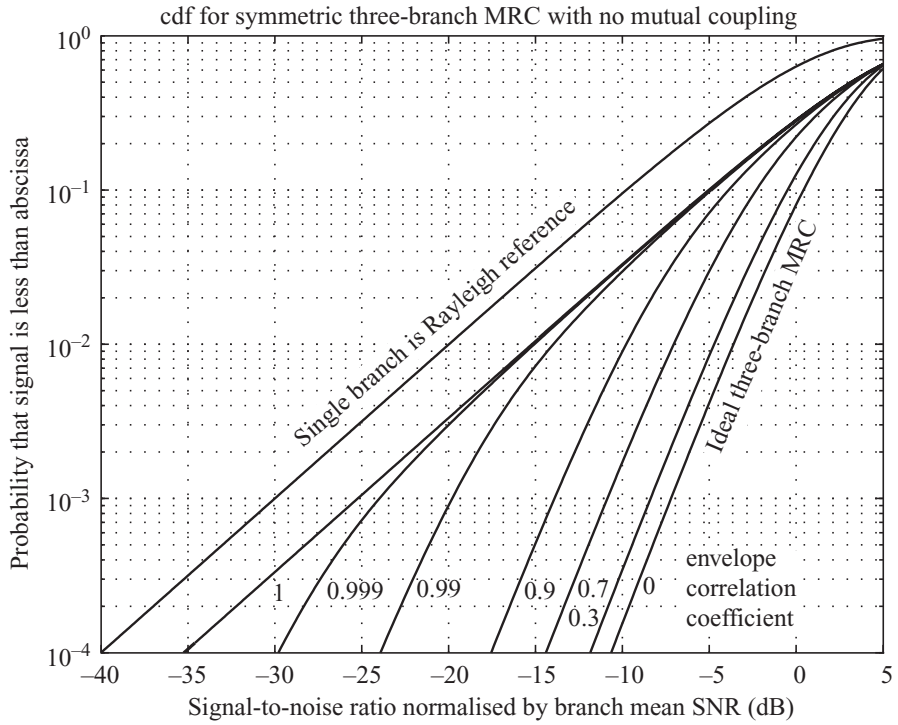


Figure 8.4.8 The cumulative density function for three symmetric branches with common mean powers and maximum ratio combining. There is a diversity gain of three (4.7 dB) for fully correlated branches, although this is reduced in the mobile situation if mutual coupling is accounted for

where

$$\begin{aligned} P_1 &= (1+2|\rho|)^2\,, \\ P_2 &= 3|\rho|\left(\frac{x}{\Gamma}+1-|\rho|\right)+(1+2|\rho|)(1-|\rho|). \end{aligned} \tag{8.4.29}$$

This result is used below for an example of a circular array of sloping monopoles. The cdf is plotted in Figure 8.4.8.

The limit for fully correlated signals is

$$P_{\mathrm{MR}}^{(3)}\ (\gamma < x)|_{\rho\to 1} = 1 - \mathrm{e}^{-x/(3\Gamma)} \tag{8.4.30}$$

So even if the branch signals are fully correlated, there is a diversity gain of three relative to a single Rayleigh branch, independent of the probability. Similarly for the fully correlated two branch case, there is a diversity gain of two. This situation is like a purely point-to-point situation and adding another receiving antenna. The signals are fully correlated which means the fading is the same for all branches: there is multipath but there is no appreciable angular spread so that each of the antenna elements sees an identical incoming signal. The gain increase results from increasing the receiving array size. However, in the mobile situation, where there is large angular

spread, this cannot be the situation because to have fully correlated antennas means that the omni elements, in an omni scenario, must merge together to form a single antenna in order to have fully correlated signals. However, the single antenna cannot have a non-unity diversity gain. So a factor needs to be included in the diversity gain definition which caters for this situation. This factor is discussed in Section 8.4.11.2.

8.4.3.10 Switched diversity

Switched combining does not perform as well as selection combining, and the diversity gain depends on the choice of threshold signal-to-noise ratio, denoted A. Above the threshold level, the performance is better than single branch, since the combiner switches, more often than not, to a higher signal level once it fades to the threshold level. Below the threshold, the behaviour depends on the switching strategy. If the switch is to another branch, independent of that branch's level ('blind switching'), then the performance below threshold is worse than selection, since some of the time, the switch will be to a worse channel which cannot occur with selection. If the switching is based on a selection strategy, then the performance below threshold is the same as selection diversity.

For two-branch switching, with uncorrelated branches that have the same mean powers the cdf is calculated from (Rustako *et al*., 1973; Lee, 1982):

$$P_{\mathrm{SW}}^{(2)}(\gamma < x) = (\mathrm{Prob}\,(\gamma < x)|\gamma = \gamma_1)\ \mathrm{OR}\ (\mathrm{Prob}\,(\gamma < x)|\gamma = \gamma_2), \qquad (8.4.31)$$

which can be stated, for blind switching strategy,

$$\begin{aligned}
&\text{below threshold } (x \geq A):\\
&P_{\mathrm{SW}}^{(2)}(\gamma < x) = \mathrm{Prob}\,(A < \gamma_1 < x)\ \mathrm{OR}\ (\mathrm{Prob}\,(\gamma_1 < A)\ \mathrm{AND}\ \mathrm{Prob}\,(\gamma_1 < x))\\
&\text{above threshold } (x > A):\\
&P_{\mathrm{SW}}^{(2)}(\gamma < x) = (\mathrm{Prob}\,(\gamma_1 < A)\ \mathrm{AND}\ \mathrm{Prob}\,(\gamma_1 < x)),
\end{aligned} \qquad (8.4.32)$$

and the probabilities are known directly from the single-branch cdf, above, and the OR and AND correspond to addition and multiplication, respectively.

Figure 8.4.9 shows the two-branch switched combination performance, where the switching occurs independently of the comparison of the levels. The threshold relative to the mean *SNR* is -10 dB. Figure 8.4.10 shows the performance for the three-branch case where the switching strategy is that of selection combining. In this case the threshold is -8 dB. The difference between the diversity gains is clear from the two figures, and it is evident that the switching threshold can affect the diversity gain, depending on what probability level is selected for its definition.

8.4.4 Effect of correlated noise (interference)

A basic assumption in the above discussion of combining is that the noise in the branches is mutually uncorrelated. Antenna noise is thermal noise that is gathered by

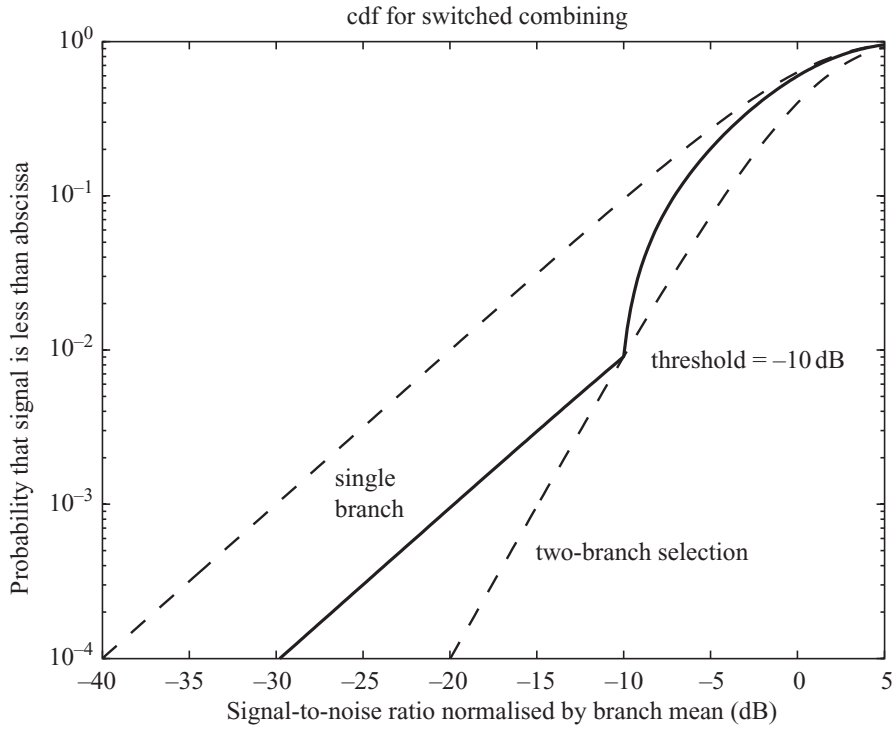


Figure 8.4.9 Two-branch switch combining for uncorrelated, equal power branches. The switch action occurs every time the currently used branch passes the threshold of −10 dB in the negative direction

the antenna elements, and although it is the same thermal source for each antenna, this is unlikely to be correlated between branches if the antennas are themselves uncorrelated with respect to the received, fading signals. This is because the source of thermal noise is distributed, probably at least as much as the wanted signal. It is straightforward to introduce the effect of correlated noise into the diversity gain, although in practice, such noise is most likely to be classed as interference. Other correlated noise sources may include noise entering through power supplies or grounding systems which are common to the multiple receivers, and noise such as phase noise entering through common local oscillator electronics.

The correlated noise is denoted N_{I}, the uncorrelated noise N_{T}, and the wanted signal power is S. The situation is shown in Figure 8.4.11, although the combining weighting is not explicit.

For M equal powers in the branches for each of the quantities, the combined signal-to-noise ratio is

$$SNR = \frac{M^2 S_{\mathrm{I}}}{M^2 N_{\mathrm{I}} + M N_{\mathrm{T}}} = \frac{M\,(S_{\mathrm{I}}/N_{\mathrm{T}})}{M\,(N_{\mathrm{I}}/N_{\mathrm{T}}) + 1} \tag{8.4.33}$$

and this is shown in Figure 8.4.12. The correlated noise acts to decrease the signal-to-(total) noise ratio, cutting in strongly once the correlated noise comes within about

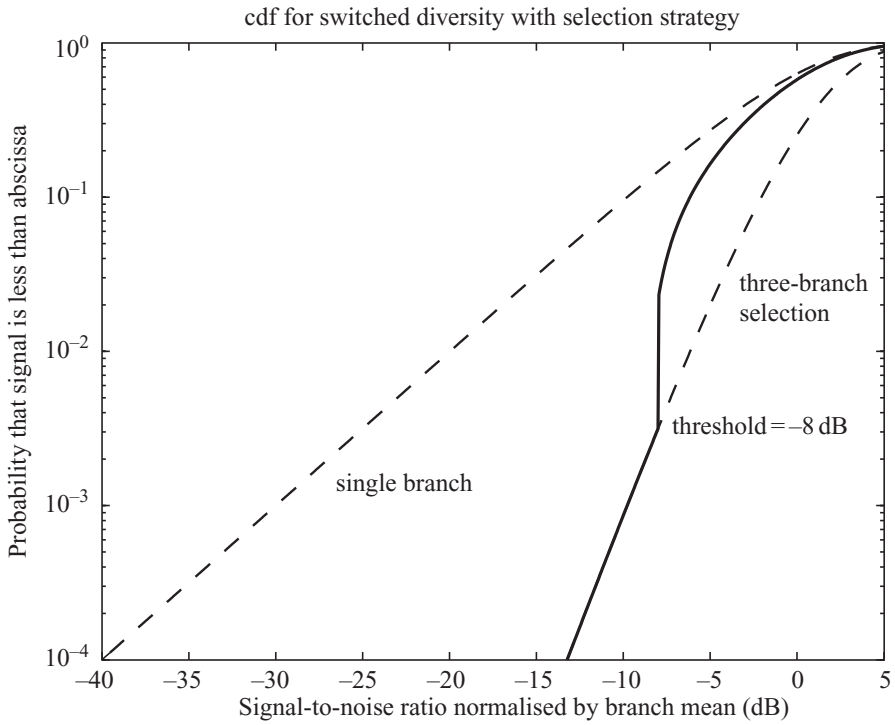


Figure 8.4.10 Three-branch switch combining for a threshold of −8 dB, and using selection combining algorithm strategy at switching

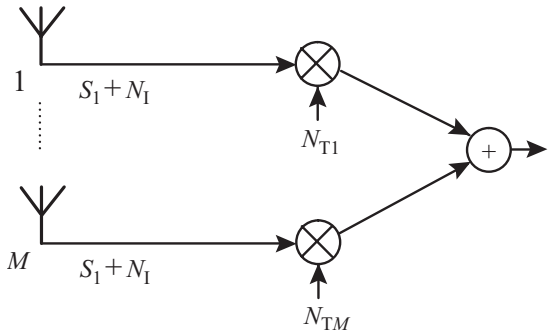


Figure 8.4.11 Correlated noise (interference) arriving through the antennas, and additive, uncorrelated noise from the thermal contribution of the front-end electronics

10 dB of the uncorrelated noise level. The effect and processing of interference is covered further in Chapter 9.

8.4.5 Incident fields and antenna conditions for diversity

The basic ideal conditions for diversity action are that the received signals:

- are uncorrelated (experience different fading); and
- have the same mean power.

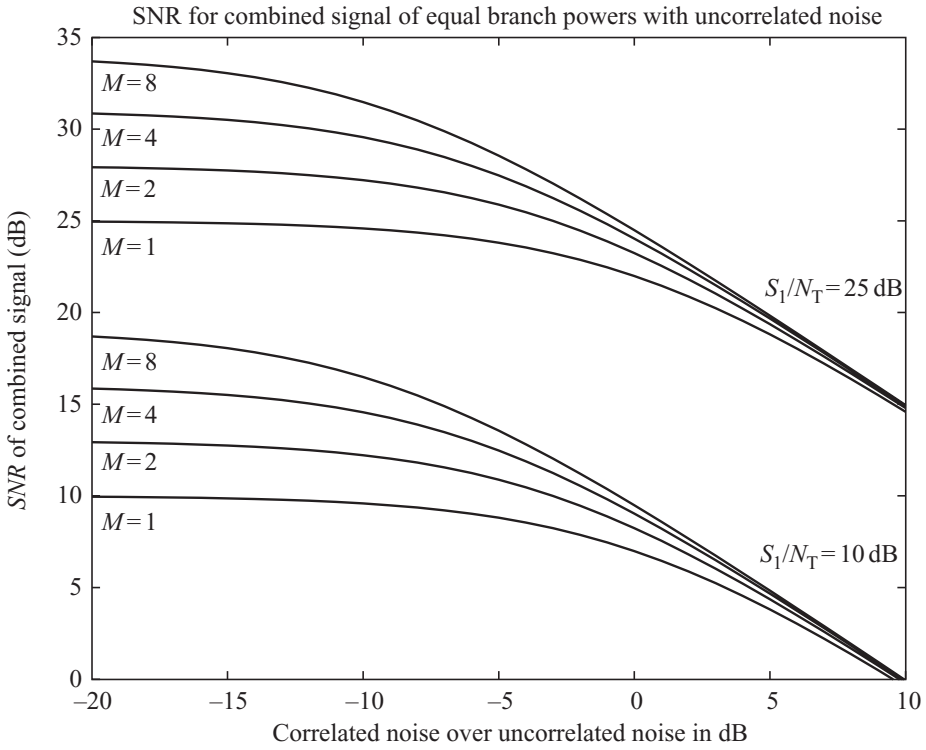


Figure 8.4.12 *The effect of correlated noise on the average SNR, which can be interpreted as the effect of interference on the maximum ratio combined resultant SNIR. Other sources of correlated noise at the receivers are noise through power supplies and noise such as phase noise from local oscillators which are common to the receivers*

Diversity gain can be realised when such signals are combined. Practical antennas normally depart from having these ideal conditions, with finite correlations and dissimilar mean power levels compromising the performance, but still offering diversity gain.

For the design of diversity antennas, it is of interest to formalise the conditions required of the antennas for a given incident field scenario. A general expression containing the polarisation information can be found by defining polarisation matrices for both the incident wave distribution and the antenna patterns (Collin and Zucker, 1969; Vaughan and Andersen, 1987a,b). Referring to antenna coordinates, the incident field (in V (m^{-1})steradian^{-1}), is denoted (see Chapter 2)

$$\boldsymbol{E}\ (\theta, \phi, t) = E_\theta\ (\theta, \phi, t)\hat{\theta} + E_\phi\ (\theta, \phi, t)\hat{\phi} \tag{8.4.34}$$

and the antenna diversity element pattern is, for example for the kth antenna,

$$\boldsymbol{h}_k\ (\Omega) = h_{\theta k}\ (\Omega)\hat{\theta} + h_{\phi k}\ (\Omega)\hat{\phi}. \tag{8.4.35}$$

8.4.5.1 Incident fields and diversity antenna polarisation matrices

A polarisation matrix for the incident field is

$$\boldsymbol{\Gamma}'\,(\theta_1, \phi_1; \theta_2, \phi_2) = \begin{bmatrix} \Gamma'_{\theta\theta} & \Gamma'_{\theta\phi} \\ \Gamma'_{\phi\theta} & \Gamma'_{\phi\phi} \end{bmatrix} \quad \text{(incident fields)} \tag{8.4.36}$$

where the matrix elements are, for example,

$$\Gamma'_{\theta\phi}\,(\theta_1, \phi_1; \theta_2, \phi_2) = \langle E_\theta\,(\theta_1, \phi_1, t) \cdot E^*_\phi\,(\theta_2, \phi_2, t)\rangle \tag{8.4.37}$$

and the averaging is over the time or space, as discussed in Section 6.7. By averaging over space, or rather many different spaces, the averaging is equivalent to the expectation over many different local, or snapshot scenarios. By keeping the mobile terminal within a local area, or local space, when the space and time can be interchanged, the averaging is within the same physical area only, so this corresponds typically to a single sample of the scenarios in which the terminal operates.

Similarly, a polarisation matrix for the jth and kth antenna elements is denoted here using the solid angle $\Omega = (\theta, \phi)$, as

$$\boldsymbol{\Gamma}_{j,k}\,(\Omega_1, \Omega_2) = \begin{bmatrix} \Gamma_{\theta\theta jk} & \Gamma_{\theta\varphi jk} \\ \Gamma_{\varphi\theta jk} & \Gamma_{\varphi\varphi jk} \end{bmatrix} \quad \text{(antenna elements)} \tag{8.4.38}$$

where a matrix element is given in terms of the antenna element patterns such as

$$\Gamma_{\theta\phi jk}\,(\Omega_1; \Omega_2) = h_{\theta j}\,(\Omega_1) \cdot h^*_{\phi k}\,(\Omega_2). \tag{8.4.39}$$

The open circuit voltage correlation coefficient is now

$$\begin{aligned} \rho_{0jk} &= \int\int (\Gamma_{\theta\theta jk}\Gamma'_{\theta\theta} + \Gamma_{\phi\phi jk}\Gamma'_{\phi\phi} + \Gamma_{\theta\phi jk}\Gamma'_{\theta\phi} + \Gamma_{\phi\theta jk}\Gamma'_{\phi\theta})\,\mathrm{d}\Omega_1\,\mathrm{d}\Omega_2 \\ &= \int\int \mathrm{tr}\,(\boldsymbol{\Gamma}_{jk}\boldsymbol{\Gamma}')\,\mathrm{d}\Omega_1\,\mathrm{d}\Omega_2. \end{aligned} \tag{8.4.40}$$

So the condition for uncorrelated signals between the antennas (here, the jth and kth elements) is that the source (incident fields) and sink (receiving antenna patterns) polarisation matrices be orthogonal in the sense of the inner product between the polarisation matrices of the antenna patterns and the incident fields: $\int\int \mathrm{tr}\,(\boldsymbol{\Gamma}_{jk}\boldsymbol{\Gamma}')\,\mathrm{d}\Omega_1\mathrm{d}\Omega_2 = 0$. This result is general. In the reporting of experimental results, it is emphasised that care needs to be taken with the interpretation of the averaging estimates of the scenarios.

8.4.5.2 Correlation from the incident power distribution and antenna patterns

The major simplifications for scenario modelling (see Section 6.7) are that the polarisations are mutually uncorrelated and also uncorrelated in space, and have the same incident angular power density form, $P_S(\theta, \phi)$. Then

$$\boldsymbol{\Gamma}'\,(\theta_1, \phi_1; \theta_2, \phi_2) = P_S\,(\theta, \phi)\delta(\theta_1 - \theta_2)\delta(\phi_1 - \phi_2)\begin{bmatrix} XPD & 0 \\ 0 & 1 \end{bmatrix} \tag{8.4.41}$$

and the cross-polar discrimination, XPD, is recalled from Section 6.7 as the directional ratio of powers in each polarisation, and is here constant since the incident power density for each polarisation is taken as the same. These simplifications result in the open circuit correlation coefficient being expressed as the dot product of the antenna pattern vectors and weighted by the incident power density function; here for normalised patterns $\int |\boldsymbol{h}\ (\Omega)|^2 \mathrm{d}\Omega = 4\pi$, and for the incident power as a probability density function, $p_{\mathrm{S}}\ (\Omega)$,

$$\rho_{0jk} = \int p_{\mathrm{S}}\ (\Omega) \boldsymbol{h}_j\ (\Omega) \cdot \boldsymbol{h}_k^*\ (\Omega)\, \mathrm{d}\Omega. \tag{8.4.42}$$

So the condition for uncorrelated open-circuit antenna signals is that the patterns are orthogonal over the incident power distribution (Vaughan and Andersen, 1983, 1987a,b).

While the above formulation is compact, more interpretation is available if the polarisations of the patterns and the incident fields are made explicit. The polarisations need not have the same distributions in general. Using the (θ, ϕ) polarisations, for example, the probability density functions for the polarised components of the incident waves are denoted $p_\theta\ (\Omega)$ and $p_\theta\ (\phi)$. The correlation coefficient can now be written out as

$$\rho_{0jk} = \frac{H_{jk}}{\sqrt{H_j H_k}}, \tag{8.4.43}$$

where

$$H_{jk} = \int (p_\theta\ (\Omega) h_{j\theta}\ (\Omega) h_{k\theta}^*\ (\Omega) + p_\phi\ (\Omega) h_{j\phi}\ (\Omega) h_{k\phi}^*\ (\Omega)) \mathrm{d}\Omega \tag{8.4.44}$$

and

$$H_j = \int (XPD(\Omega) p_\theta\ (\Omega) |h_{j\theta}\ (\Omega)|^2 + p_\phi\ (\Omega) |h_{j\phi}\ (\Omega)|^2)\, \mathrm{d}\Omega \tag{8.4.45}$$

and similarly for H_k. Here, the cross-polar discrimination is directional, as in Section 6.7: $XPD(\Omega) = \Gamma'_{\theta\theta}\ (\Omega) / \Gamma'_{\phi\phi}\ (\Omega)$.

8.4.5.3 Mean power from the incident power distribution and antenna pattern

To keep the branch powers the same, the gains of the elements into the incident power distribution should be the same, here stated using the normalised patterns and the distributed directivities,

$$P_{\mathrm{rec}}^{(j)} \propto \int p_{\mathrm{S}}\ (\Omega) |\boldsymbol{h}_j\ (\Omega)|^2 \mathrm{d}\Omega = \int p_{\mathrm{S}}\ (\Omega) |\boldsymbol{h}_k\ (\Omega)|^2 \mathrm{d}\Omega \propto P_{\mathrm{rec}}^{(k)} \tag{8.4.46}$$

and these should of course be kept as high as possible in order to maximise the mean branch power. By separating the polarisations, and introducing the powers received by perfectly polarised, omnidirectional reference antennas, $P_1 = E\{|E_\theta\ (\theta, \phi)|^2\} = \Gamma'_{\theta\theta}$

for the θ-polarisation and similarly P_2 for the ϕ-polarisation, the received power is (Jakes, 1974)

$$P_{\text{rec}} = \int (P_1 p_\theta\ (\Omega)|h_\theta\ (\Omega)|^2 + P_2 p_\phi\ (\Omega)|h_\phi\ (\Omega)|^2)\,\mathrm{d}\Omega \tag{8.4.47}$$

and by normalising with the received power of both polarisations, $P_1 + P_2$, the *mean effective gain* becomes written as

$$MEG = \int \left(\frac{XPD}{1+XPD} p_\theta\ (\Omega)|h_\theta\ (\Omega)|^2 + \frac{1}{1+XPD} p_\phi\ (\Omega)|h_\phi\ (\Omega)|^2 \right) \mathrm{d}\Omega \tag{8.4.48}$$

as discussed in Section 8.2.8. Here $XPD = P_1/P_2$ is the ratio of powers received by the polarised omnidirectional antennas in the scenario.

As with the correlation coefficient formula, care should be taken regarding interpretation of the statistical expectation in the probability density function. If a snapshot scenario is considered (so that $p_\theta\ (\Omega)$ and $p_\phi\ (\Omega)$ are interpreted as simply a density function for a given location) , then the mean effective gain is likely to be dependent on the orientation of the antenna within the scenario, since a typical snapshot scenario is directional. Here, the 'mean' is over the directionality of the local scenario. When the expectation of the incident power density is taken, so that the distributions are pdfs, this directionality is typically removed, or modelled as omnidirectional, at least in the azimuthal direction: see Section 6.7. In this situation, the mean effective gain is the expected gain over many different scenarios.

From the previous sections, the correlation coefficient need not be zero; a value of 0.7 for the envelope correlation, or $\rho_0 \approx \sqrt{\rho_e} \approx 0.8$, is sufficiently low, for equal branch powers, to get most of the diversity gain, i.e.

$$\int p_S\ (\Omega)\boldsymbol{h}_j\ (\Omega) \cdot \boldsymbol{h}_k^*\ (\Omega)\,\mathrm{d}\Omega \lesssim 0.8 \tag{8.4.49}$$

showing that the patterns need not be truly orthogonal. Likewise, the branch powers can be somewhat different, a few dB difference for uncorrelated branches.

8.4.6 Antenna patterns for diversity

For designing diversity antennas, we seek patterns that are sufficiently different to give a low correlation coefficient. The complex vector nature of the pattern, $\boldsymbol{h}$, offers three independent methods for changing one pattern from another: amplitude, phase, and polarisation, each a specialised form of pattern diversity. Using each of these alone, in antenna coordinates, results in:

- space diversity: identical patterns, but displaced in space (by distance $\boldsymbol{d}$)

$$\boldsymbol{h}_2(\Omega) = \boldsymbol{h}_1(\Omega)\mathrm{e}^{-\mathbf{j}\boldsymbol{k_0}\cdot\boldsymbol{d}}; \tag{8.4.50}$$

- polarisation diversity: identical patterns, but with different polarisations $(\hat{p}_1, \hat{p}_2)$

$$\boldsymbol{h}_2(\Omega) \text{ with } \hat{p}_1, \boldsymbol{h}_2(\Omega) \text{ with } \hat{p}_2 : \quad |\boldsymbol{h}_1\ (\Omega)|^2 = |\boldsymbol{h}_2\ (\Omega)|^2; \tag{8.4.51}$$

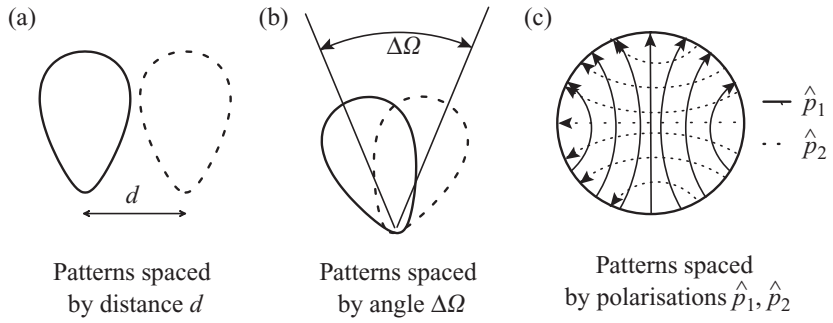


Figure 8.4.13 Three degrees of freedom for realising pattern diversity: (a) space, (b) angle, and (c) polarisation

- angle (amplitude) diversity: identical patterns in polarisation and phase, but having different power patterns, the idealised special case being *angle diversity*, where the patterns are displaced in the pointing angle (by $\Delta\Omega$),

$$\boldsymbol{h}_2(\Omega) = \boldsymbol{h}_1(\Omega + \Delta\Omega). \tag{8.4.52}$$

These are depicted in Figure 8.4.13.

8.4.7 Space diversity

When the incident wave distribution covers large angles, i.e. has a large angular spread (or *Doppler spread*, see Sections 6.3 and 6.4), even a small change of phase between patterns has a strong effect on decorrelating the received signals. The classic space diversity relations for omni antennas in the 2D Clarke scenario demonstrate this effect. A vertically polarised pattern in polar (instead of spherical) coordinates can be written $\boldsymbol{h}_1\ (\theta = \pi/2, \phi) = h(\phi)\hat{\theta}$, and a second pattern is identical, but translated by a distance d in direction ϕ_0 with respect to the reference axis, and the second pattern can be written $\boldsymbol{h}_2(\theta = \pi/2, \phi) = h(\phi)\mathrm{e}^{j2\pi d/\lambda \cos\phi_0}\hat{\theta}$. For a uniform, 2D scenario (Clarke model), and omnidirectional elements, $h(\phi) = 1$, and the open circuit voltage correlation coefficient function is the well-known result (see also Section 6.4)

$$\begin{aligned}\rho_0(d) &= \int_0^{2\pi} (\hat{\theta}) \cdot (\hat{\theta}\mathrm{e}^{\mathrm{j}2\pi d/\lambda \cos\phi})\mathrm{d}\phi \\ &= \int_0^{2\pi} \mathrm{e}^{\mathrm{j}2\pi d/\lambda \cos\phi}\mathrm{d}\phi = J_0(2\pi d/\lambda).\end{aligned} \tag{8.4.53}$$

If the scenario covers the full sphere, and the patterns are 3D omnidirectional (not easily realised!), then the spatial correlation coefficient function does not change much, as far as pure spacing requirements are concerned, becoming $\rho_0(d) = \sin kd/(kd)$.

For a zero correlation coefficient, these scenarios give a spacing of $d(\rho_0 = 0) = 0.38\lambda$ (2D) and $d(\rho_0 = 0) = 0.5\lambda$ (3D). The various correlation coefficient functions are shown in Figure 8.4.14. For an envelope correlation coefficient of 0.7, which still gives most of the diversity gain, the spacings reduce to $d(\rho_\mathrm{e} = 0.7) \approx 0.13\lambda$ (2D)

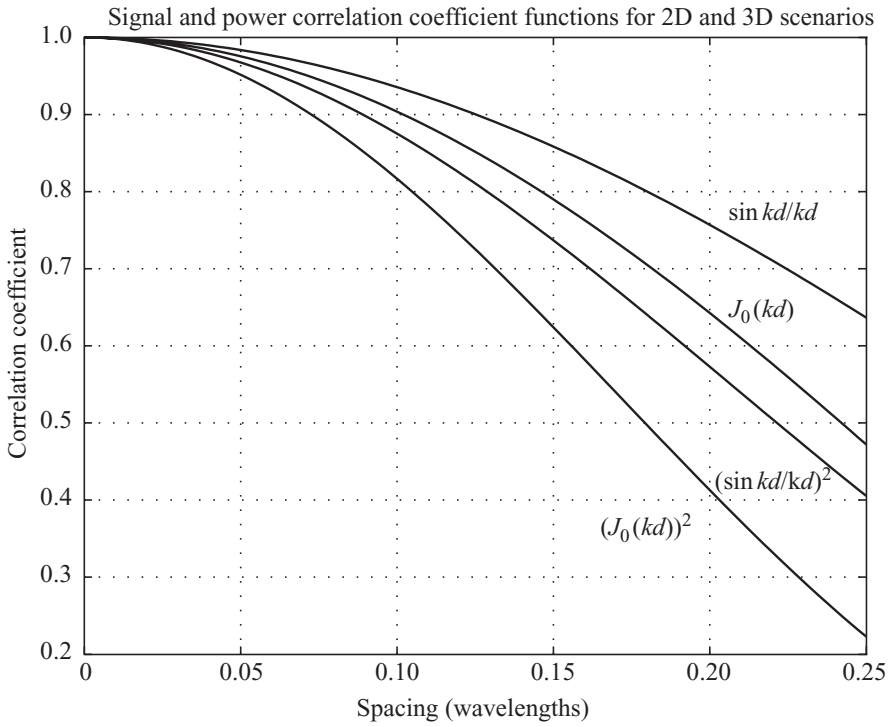


Figure 8.4.14 The spatial correlation coefficient functions for the uniform 2D Clarke scenario, and for the uniform full-sphere scenario

and $d(\rho_e = 0.7) \approx 0.16\lambda$ (3D) , so that under the assumed scenario, effective space diversity can be realised with very small antenna spacings. Further below, it is shown that for loaded antennas (which is the usual situation), the mutual coupling allows even closer spacings than those indicated here by the fields correlation coefficient function.

8.4.8 Angle diversity

The need for compact antennas at the mobile terminal constrains the antenna directivity to be low, with currently used carrier frequencies. At base stations, however, where there is more space available for larger aperture antennas, medium and high directivities are possible. Also, at outdoor base stations, there can be a directive scenario, meaning that there is often a cluster of scatterers about a mobile user that are illuminating the base station. As carrier frequencies used in cellular systems continue to increase, the electrical size of the aperture available on personal terminals will allow higher directivities. Space diversity for directive patterns, or directive scenarios, was discussed in Section 8.4.7. Here, angle diversity for both low- and high-gain antennas is discussed.

For low directivity, the patterns can be modelled conveniently for correlation analysis with circular functions, e.g. $g(\theta) = \cos^n(\theta/2)$, and for high directivities, a Gaussian function is convenient, $g(\theta) = e^{-\theta/2\sigma_g^2}$, as discussed in Section 8.1. The combined effect of a directional scenario with a directional antenna is now treated.

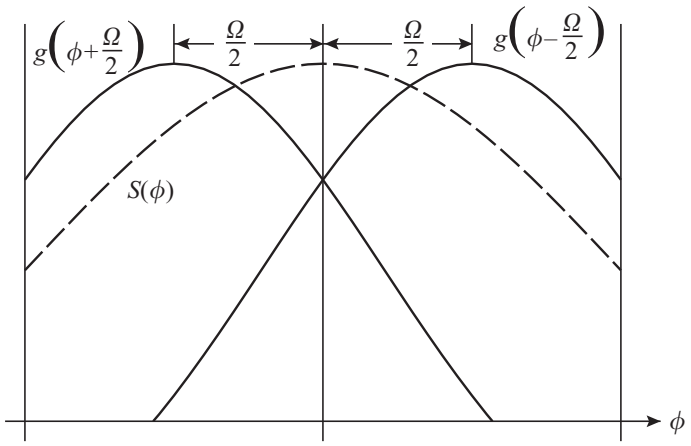


Figure 8.4.15 Angular spacing of identical beams about the direction of maximum power from a directional scenario. The scenario power density is Gaussian, as are the antenna beams

The scenario is taken as a 1D (single pattern cut) case, with the incident power density Gaussian in the azimuthal coordinate denoted

$$S_G(\phi) = p_\phi(\phi) = K_S \mathrm{e}^{-\phi^2/2\sigma_S^2}, \quad -\pi \le \phi \le \pi, \tag{8.4.54}$$

and σ_S is the spread of the angular power distribution. The function is normalised so that it can be used as a pdf using (see Appendix C)

$$K_S^{-1} = \sqrt{2\pi}\sigma_S \operatorname{erf}\left(\frac{\pi}{\sqrt{2}\sigma_S}\right). \tag{8.4.55}$$

The antenna *amplitude* pattern is also Gaussian but is bound by the support of the circular coordinate, written

$$g(\phi) = \mathrm{e}^{-\phi^2/2\sigma^2}, \quad -\pi \le \phi \le \pi. \tag{8.4.56}$$

So the correlation between a pair of beams displaced by an angle Ω is (here showing the formulation for complex beams)

$$\begin{aligned} R_{\tilde{g}_1\tilde{g}_2}(\Omega) &= \int_{-\infty}^{\infty} S(\phi)\tilde{g}\left(\phi+\frac{\Omega}{2}\right)\tilde{g}^*\left(\phi-\frac{\Omega}{2}\right)\mathrm{d}\phi \\ &\simeq \int_{-\pi}^{\pi} S(\phi)\tilde{g}\left(\phi+\frac{\Omega}{2}\right)\tilde{g}^*\left(\phi-\frac{\Omega}{2}\right)\mathrm{d}\phi. \end{aligned} \tag{8.4.57}$$

The symbol $\simeq$ is used here to indicate the effect of the truncation of the Gaussian beam functions by the circular coordinate support. The situation is depicted in Figure 8.4.15, with the correlation displacement shown as a symmetric split about the direction of maximum power in the directional scenario.

8.4.8.1 Rule of thumb for Gaussian beams in circular coordinate system

A conservative, simple rule-of-thumb for the limit of the spread of the beam, σ_{g}, is obtained by maintaining the support γ_{B} as about $5\sigma_g$ radians (from $\pm 2.5\ \sigma_{\mathrm{g}}$ radians about the centre of the beam), corresponding to 92 per cent (from erf (2.5/2)) of the energy of a centralised Gaussian beam being within the support, and allowing the maximum correlation lag to equal twice the HPBW ($\approx 2^{*}0.6\sigma_{\mathrm{g}}$ radians). The bound follows as $5\sigma_g + 1.2\sigma_g < 2\pi$, giving a maximum standard deviation parameter for the beam function of $\sigma_{\mathrm{g}} \approx 1$ radian, or a maximum HPBW of about 95°.

Note that this rule-of-thumb addresses the accuracy of an approximation within the mathematical formulation, and does not account for inaccuracies of the Gaussian beam modelling a real-world beam. For comparison, a |sinc|-shaped beam (typical for a uniformly illuminated linear array) has about 90 per cent of its total energy in the sense of having support of $\pm\infty$ within the support of its main lobe. For the beam function being well within the support, i.e. $\sigma_{\mathrm{g}} \leq\sim 1$ radian, corresponding to moderate and high directivities, there results $\sigma_{\mathrm{c}}^2 \approx \sigma_g^2/2$ and $K_{\mathrm{c}} \approx \sqrt{\pi}\sigma_g$ (K_{c} is a normalisation constant.) The above rule-of-thumb is for a uniform scenario; for directive scenarios, the beams can be less directive. Using beams with lower directivity ($\sigma_g <\sim 1$ radian) results in the angular correlation coefficient function values increasing, especially for its smaller values. The correlation angle (at which the signals from each beam become effectively decorrelated) increases as a result.

8.4.8.2 Correlation function for angularly displaced beams: Gaussian beams for high directivity

The open circuit voltage correlation coefficient function for the beams, in an uncorrelated scenario, is

$$\rho(\Omega) = \frac{\langle V_{O1} V_{O2}^{*} \rangle}{\sqrt{\langle |V_{O1}|^2 \rangle \langle |V_{O2}|^2 \rangle}} = \frac{R_{g_1 g_2}(\Omega)}{\sqrt{R_{g_1 g_1}(\Omega) R_{g_2 g_2}(\Omega)}}, \tag{8.4.58}$$

and we have

$$R_{g_1 g_2}(\Omega) \simeq K_{\mathrm{S}} K_{\mathrm{c}} \mathrm{e}^{-\Omega^2/(4\sigma_g^2)} \tag{8.4.59}$$

where

$$K_{\mathrm{c}} = \sqrt{2\pi}\sigma_{\mathrm{c}}\, \mathrm{erf}\left(\frac{\pi}{\sqrt{2}\sigma_{\mathrm{c}}}\right), \tag{8.4.60}$$

and

$$\sigma_{\mathrm{c}}^2 = \frac{\sigma_{\mathrm{s}}^2 \sigma_g^2}{2\sigma_{\mathrm{s}}^2 + \sigma_g^2}. \tag{8.4.61}$$

From the symmetry of the configuration, $R_{g_1 g_1}(\Omega) = R_{g_2 g_2}(\Omega)$, so the normalising factor contains

$$R_{g_1 g_1}(\Omega) \simeq K_{\mathrm{S}} K_k \mathrm{e}^{-\Omega^2/4\left(\sigma_g^2 + 2\sigma_{\mathrm{S}}^2\right)} \tag{8.4.62}$$

where

$$K_k = \sqrt{2\pi}\frac{\sigma_c}{2}\left(\text{erf}\left(\frac{\pi - k\Omega}{\sqrt{2}\sigma_c}\right) - \text{erf}\left(\frac{-\pi - k\Omega}{\sqrt{2}\sigma_c}\right)\right) \tag{8.4.63}$$

and here

$$k = \frac{\sigma_S^2}{\sigma_g^2 + 2\sigma_S^2} = \frac{\sigma_c^2}{\sigma_g^2}. \tag{8.4.64}$$

If the two beams $g(\phi \pm \Omega/2)$ are well within the support of the scenario, $S_G(\phi)$, then the correlation coefficient function is approximately Gaussian, namely

$$\rho(\Omega) \approx \text{e}^{-\Omega^2/(2\sigma_\rho^2)}, \tag{8.4.65}$$

where

$$\sigma_\rho^2 = \sigma_g^2\left(\frac{\sigma_g^2}{\sigma_S^2} + 2\right). \tag{8.4.66}$$

This equation relates the spread of the correlation coefficient function to that of the beam and the scenario. The power correlation coefficient, which is similar to the envelope correlation coefficient, namely $|\rho(\Omega)|^2 \sim \rho_e(\Omega)$, is shown in Figure 8.4.16 with the scenario spread, σ_S, as a parameter. As noted above, the HPBW has little influence as long as it is less than about 90°.

The limiting cases for the scenario are as follows. For a single incident wave, $\sigma_S = 0$, which gives $\sigma_\rho = \infty$, and the correlation coefficient function is always unity, independent of the beam, as long as it has non-zero gain. This result can also be seen directly from the defining equation with $S(\phi) = \delta(\phi)$. The result comes from the fact that for a single incident wave, the open circuit voltages of the two beams are simply scaled versions of the same signal.

For the case of the scenario having the same spread as the beams, $\sigma_S = \sigma_g$, then $\sigma_\rho = \sqrt{3}\sigma_g$, i.e. the correlation coefficient function has $1/\sqrt{3}$ of the spread of the beams. Similarly, for the spread of the scenario being the same as the spread of the power pattern, $\sigma_S = \sigma_g/\sqrt{2}$, then $\sigma_\rho = 2\sigma_g$.

For a uniform scenario, $\sigma_s \to \infty$, giving $\sigma_S \,\text{erf}\,(\pi/\sqrt{2}\sigma_s) \to \sqrt{2\pi}$, and $S(\phi) = S_C = 1/2\pi$. This results in $\sigma_\rho = \sqrt{2}\sigma_g$, i.e. the correlation coefficient function has a spread that is $\sqrt{2}$ times the spread of the amplitude beam.

The 0.7-correlation angle, Ω_d, can be defined as the angle for which $\rho_e = 0.7 \approx 1/\sqrt{2}$, and is found from

$$\rho_e(\Omega_d) \approx \text{e}^{-\Omega_d^2/\sigma_\rho^2} \tag{8.4.67}$$

$$\Rightarrow \Omega_d \approx \sigma_g\sqrt{\ln\sqrt{2}((\sigma_g/\sigma_S)^2 + 2)} = HPBW\sqrt{\tfrac{1}{8}(\sigma_g/\sigma_S)^2 + \tfrac{1}{4}}. \tag{8.4.68}$$

When $\sigma_g/\sigma_S >\sim 3$ and in particular for the uniform scenario,

$$\Omega_d = \frac{HPBW}{2} \quad \text{(uniform scenario)}, \tag{8.4.69}$$

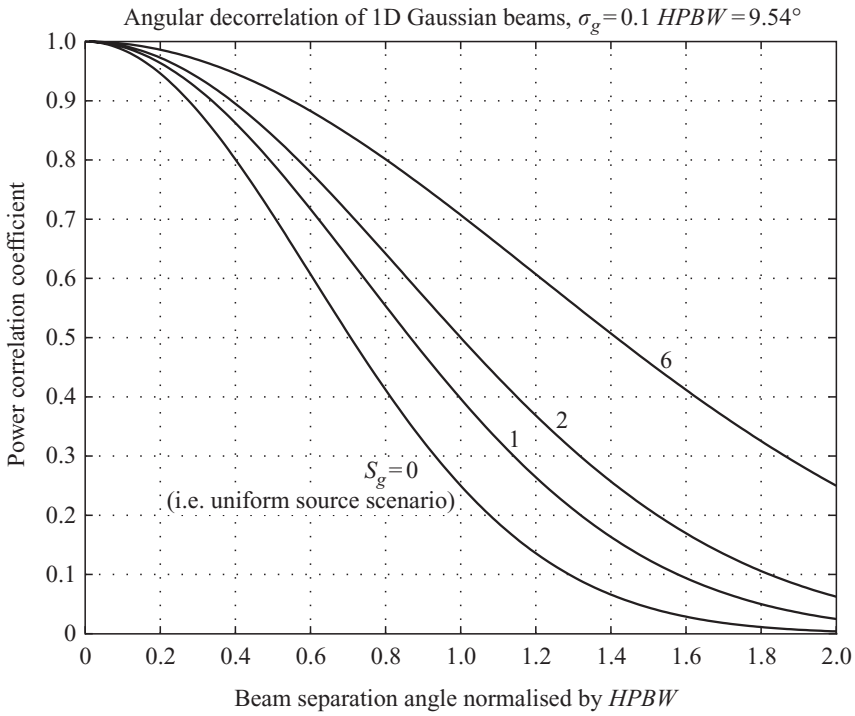


Figure 8.4.16 *The power correlation coefficient (similar to the envelope correlation coefficient) of the signals received from two angularly spaced directive beams spaced symmetrically in a directive scenario. The parameter is $S_g = \sigma_g^2/\sigma_S^2$, where σ_S is the spread of the scenario power distribution, and σ_g is the spread of each of the antennas' beam amplitudes. Reproduced with permission of IEEE (Vaughan, 1988)*

i.e. the 0.7-correlation angle is half of the half power beamwidth of the antenna pattern (Vaughan, 1998).

This simple result is independent of the pattern beamwidth, as long as the beam is well within the circular support. In practice, the sidelobe structure and non-uniform phase of a real-world pattern will make the correlation angle smaller (see below), so the above rule-of-thumb is interpreted as an upper limit for the minimum correlation angle for a uniform scenario. Figure 8.4.16 offers a feel for the effect of the spread of the scenario. For example, in the case where the power pattern has the same spread as the scenario, i.e. $(\sigma_S/\sigma_g)^2 = 2$, the 0.7-correlation beam separation increases from 0.5 to $1/\sqrt{2} = 0.7$ of the HPBW.

8.4.8.3 Correlation function for angularly displaced beams: Circular function beams for low directivity

For low directivities, the Gaussian beam becomes increasingly truncated by the circular support and this affects the results. Circular functions are therefore better pattern models. Using a beam along the x-axis, $g(\phi) = \cos^n(\phi/2)$, and an omnidirectional

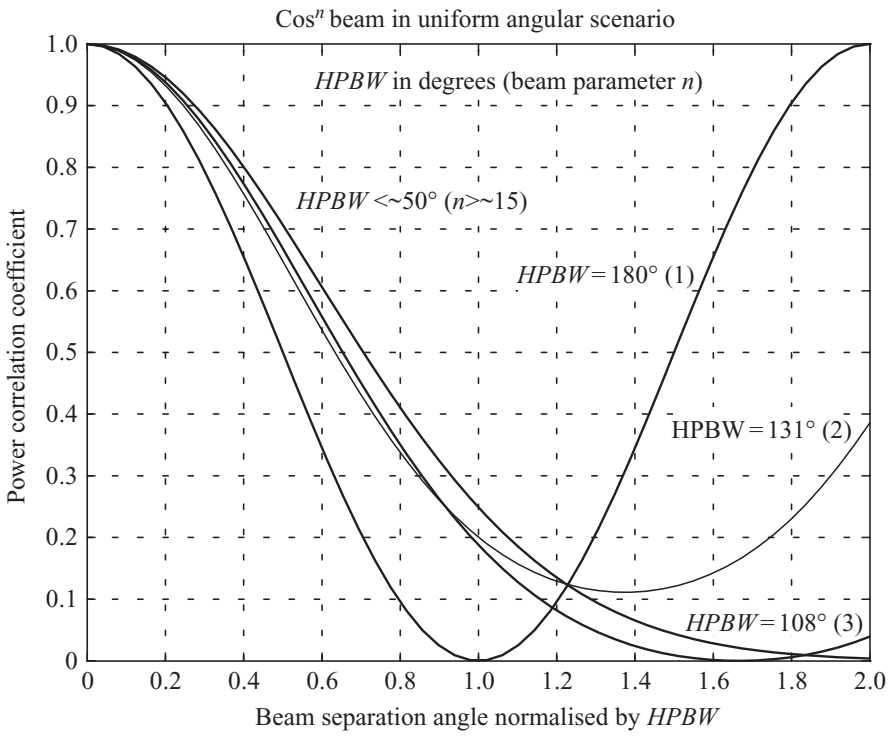


Figure 8.4.17 The power correlation coefficient for low directivity beams, the examples here being HPBW = 180°, 131°, 108°

scenario, $S(\theta, \phi) = S_{\mathrm{U}} = 1/4\pi$, the correlation coefficient is (Vaughan, 1988)

$$\rho(\Omega) = \frac{(2n)!!}{(2n-1)!!} \cdot \frac{1}{2^{2n-1}} \cdot F_n\left(n, \frac{\Omega}{2}\right), \tag{8.4.70}$$

where

$$\begin{aligned} F_n(n, \gamma) &= \sum_{k=0}^{(n-1)/2} \binom{n}{k}^2 \cos(n-2k)\gamma, \quad n \in I, n \text{ even}; \\ &= \sum_{k=0}^{n/2} \binom{n}{k}^2 \cos(n-2k)\gamma \cdot \epsilon(n-2k), \quad n \in I,\ n \text{ odd}; \end{aligned} \tag{8.4.71}$$

where $\epsilon(n) = 1/2$ for $n = 0$, and $\epsilon = 0$ otherwise.

The ϕ-dependence drops out in the normalisation, as expected from the discussion of the previous section. The power correlation coefficient function, $|\rho_g(\Omega)|^2 \approx \rho_{\mathrm{e}}(\Omega)$, is plotted in Figure 8.4.17 (Vaughan, 1988) for *HPBW*s of 180° ($n = 1$), 131° ($n = 2$), 108° ($n = 3$), and *HPBW* $\lesssim 50°$ ($n \gtrsim 15$).

For the latter case of moderate and high directivities, the same rule-of-thumb that was found for the Gaussian beams is confirmed: the $\rho_{\mathrm{e}} = 0.7$-correlation angle is half of the HPBW of the beam. For example the beam with an HPBW value of

180° has a 0.7-correlation angle of 0.38 times the HPBW, i.e. about 67°; and for the 108° beamwidth, we get about 50° for the 0.7-correlation angle. The broader the low directivity cosine beam, the less the correlation angle relative to the HPBW becomes. Increasingly directive scenarios serve to increase the correlation angle in the same way as with the Gaussian beams.

8.4.8.4 Realistic patterns in an ideal uniform scenario

For comparison with the results using mathematically modelled patterns, some patterns that are more realistic, and that are complex rather than real, are investigated. The patterns are calculated from the moment method. A six-element Uda-yagi antenna with an HPBW value of 52° gives a $\rho_e = 0.7$-correlation angle of 23°, which fits the above rule-of-thumb of half the HPBW. The complex pattern gives an angle of 13°, i.e. a factor of half smaller. This is because the beam and in particular the sidelobe region features phase variations (there is a complete phase sign change at each null in the pattern). In terms of the corresponding pattern overlap, adjacent 'uncorrelated' beams overlap at the 0.8 dB points for the complex patterns and 3 dB for the real patterns. The pattern and the correlation functions are shown in Figure 8.4.18 (Vaughan, 1988). For low-directivity beams, there is less phase variation, and the correlation mechanism is dominated by the amplitude change only. An example for a lower-directivity pattern is given further below in Figure 8.4.46 and its angular correlation function in Figure 8.4.47. Another example is given by Nørklit *et al*. (1995) where a beam from a linear array is 'jittered' within an uncorrelated scenario, to demonstrate simulated diversity action from small beam movements.

8.4.9 Polarisation and field components

Identical patterns with orthogonal polarisations, obviously give uncorrelated signals if the incident waves are polarisation-uncorrelated. For example, for patterns with vertical and horizontal polarisations,

$$\int p_S\,(\Omega)(h(\Omega)\,\hat{\theta})\cdot(h(\Omega)\hat{\phi})\,\mathrm{d}\Omega = \int p_S(\Omega)|h(\Omega)|^2\hat{\theta}\cdot\hat{\phi}\,\mathrm{d}\Omega = 0 \qquad (8.4.72)$$

Finally, *field component diversity* is a term applied to the use of antennas whose patterns receive one only component of the fields at a common point in space. Such antennas can be realised using small dipoles and small loops. Their patterns have the same shapes, but are made orthogonal in the modelled incident wave scenario through either angle spacing (a 90° rotation) or polarisation change. These antennas effectively sample the energy density of the fields.

8.4.10 Energy density in the multipath field

If there is a 3D scenario of sources, such as a spherical scenario with the sources having mixed polarisations, then all field components will be present at the receiver, namely E_x, E_y, E_z, H_x, H_y, H_z. For a 2D scenario, only the components E_z, H_x, H_y are present. Gilbert (1965) introduced the energy density antenna, originally attributed

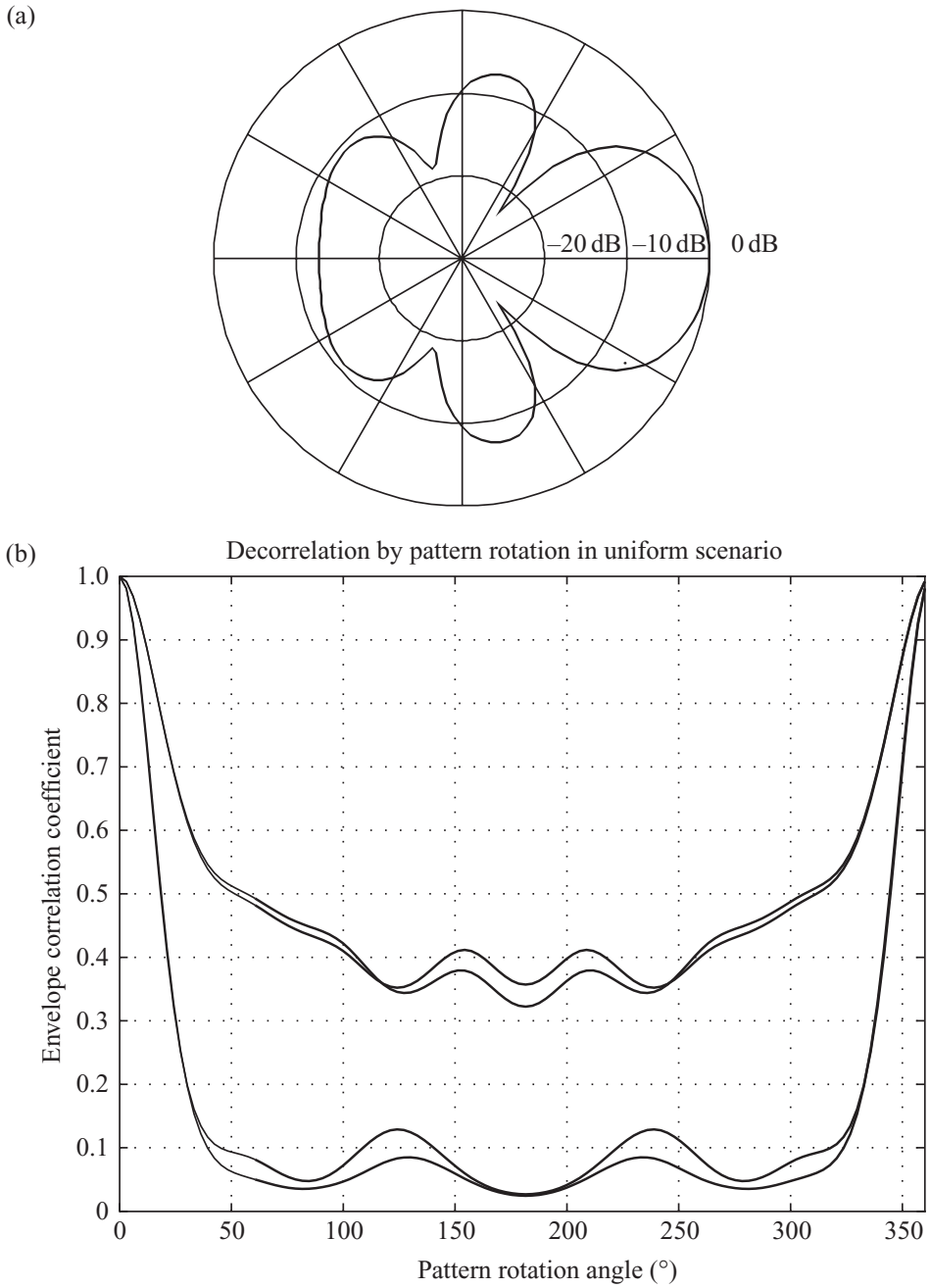


Figure 8.4.18 *(a) The power pattern and (b) the power correlation coefficient for the fading signals from two identical directive patterns in a uniform, uncorrelated scenario, as a function of their angular spacing. The correlation of complex patterns (lower traces) falls much more quickly than that for the amplitude patterns (upper traces), indicating that the pattern phase is a dominant decorrelation mechanism (cf. Figure 8.4.47 for a lower gain pattern)*

.

Table 8.13 Patterns for the energy density antenna made of electric and magnetic dipoles

Dipole current	Field component	$h_\theta(\theta, \phi)$ pattern	$h_\phi(\theta, \phi)$ pattern
J_x	E_x	$\cos\theta \cos\phi$	$-\sin\phi$
J_y	E_y	$\cos\theta \sin\phi$	$\cos\phi$
J_z	E_z	$-\sin\theta$	0
M_x	H_x	$\sin\phi$	$\cos\theta \cos\phi$
M_y	H_y	$\cos\phi$	$\cos\theta \sin\phi$
M_z	H_z	0	$\sin\theta$

to J.R. Pierce, for diversity, and this has been discussed more recently by Andrews *et al*. (2001). In the 3D scenario, six co-located dipoles (three orthogonal electric; and field component diversity has three orthogonal magnetic, built as wire loops) can, in principle, sense the six field components; see Figure 8.4.19.

The far-field pattern functions of these are given in Table 8.13. These patterns are all mutually orthogonal. If the combined antenna could be built, then a six-branch diversity antenna could be realised, occupying a common spatial position! The potential coverage of the antenna is perfectly omnidirectional in each polarisation, readily found from summing the power patterns.

Another way to look at the energy density antenna, is that the total energy, proportional to $|\boldsymbol{E}|^2 + |Z_0\boldsymbol{H}|^2$, even in a vanishingly small volume, does not fade much

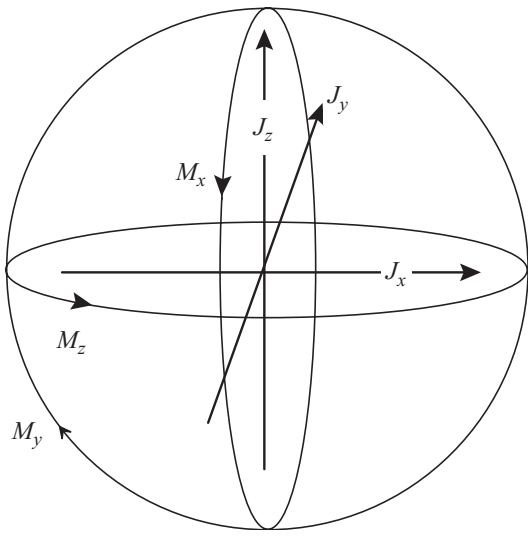


Figure 8.4.19 The energy density antenna, here depicted symbolically, uses small dipoles and loops to sense the different field components

with spatial position, the equivalent of up to six maximum ratio combined branches depending on the scenario. For a 2D Clark-type scenario on a groundplane, the equivalence is up to three branches (H_x, H_y, E_z) and for a vehicular mobile scenario (Section 6.7.6), up to about five branches (Vaughan and Andersen, 1987a).

8.4.11 Gain reduction caused by closely spaced elements

It is clear that the combination of M equal power, fully correlated branches ($\rho = 1$) can give a gain of M relative to a single branch, as seen in Figure 8.4.8. In such a case, the mutual coupling is ignored and the mean SNR increases but the relative range of fading in the combined signal is not improved. This is seen in a classical point-to-point situation with no angular spread of the incoming waves. Here, adding an extra, identical receiving antenna and combining, results in a 3 dB gain corresponding to the combination from two fully correlated branches. This gain corresponds exactly to the increase in the classical directive antenna gain. Any mutual coupling will act to reduce this gain. So mutual coupling is an important effect as antenna elements are drawn closer for making the antenna arrangement compact. In this section the role of mutual coupling is assessed against the array directivity and the diversity gain.

8.4.11.1 Directivity of linear array factor

Hansen (Chapter 5 in Collin and Zucker, 1969) gives the array directivity of an N element linear array as

$$D_{\mathrm{LA}}(\theta) = \frac{|F_{\mathrm{LA}}(\theta,\phi)|^2}{(1/4\pi)\int_0^{2\pi}\int_0^{\pi}|F_{\mathrm{LA}}(\theta,\phi)|^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi}$$

$$= \frac{|F_{\mathrm{LA}}(\theta)|^2}{1/N + (2/N^2)\sum_{n=1}^{N-1}((N-n)/(mk_0d))\sin(nk_0d)\ \cos\delta}, \quad (8.4.73)$$

where the 2D pattern is the array factor

$$|F_{\mathrm{LA}}(\theta,\phi)| = |F_{\mathrm{LA}}(\theta)| = \left|\frac{\sin(N\gamma/2)}{N\sin(\gamma/2)}\right|, \quad \gamma = \delta + k_0 d\cos\theta; \quad (8.4.74)$$

δ is the phase difference between successive elements and d is the physical spacing between adjacent elements. The broadside direction is $\theta = \pi/2, \delta = 0$; and the endfire direction is $\theta = 0, \delta = -k_0 d$.

Note that the formula is for the directivity of the array factor only; the element pattern is not included here (unless the elements are omnidirectional). For a general element pattern, there is no simple relation between the directivity of the array factor and the directivity of the array. The array directivity is sometimes called the *geometric directivity*, because it depends only on the geometry of the array including the linear phase weighting value. The form is useful for studying the loss of gain (directivity) as the elements merge, i.e. as $d \to 0$. The formula shows that if the spacing is any multiple of a half-wavelength, the denominator is $1/N$, independent of δ. Also, there is a direction (maybe more than one) for which the numerator is at its maximum

value of one, and so in this direction, the array directivity is equal to the number of elements.

Taking two elements only, the gain resulting from the array factor can be written

$$D_{\mathrm{LA}}(\theta = \pi/2) = \frac{1}{\frac{1}{2} + \frac{1}{2}\sin(k_0 d)/(k_0 d)} \quad \text{(broadside)}, \tag{8.4.75}$$

$$D_{\mathrm{LA}}(\theta = 0) = \frac{1}{\frac{1}{2} + \frac{1}{4}\sin(2k_0 d)/(k_0 d)} \quad \text{(endfire)}, \tag{8.4.76}$$

and these are given in Figure 8.4.20. The directivity value of two is evident for the half-wavelength spacings for both cases. The maxima of the directivity reduce, along with the magnitude of the oscillating values in general, as the spacing increases so that the directivity approaches $N = 2$ as the spacing becomes large.

The reduction in directivity as the elements merge is different for the endfire and broadside cases, and other directions fall in between these cases. Because the single direction nature of the directivity reduction is independent of the elements and is also direction-dependent in this formulation, there is no definitive behaviour from this formulation for the case of distributed directivity.

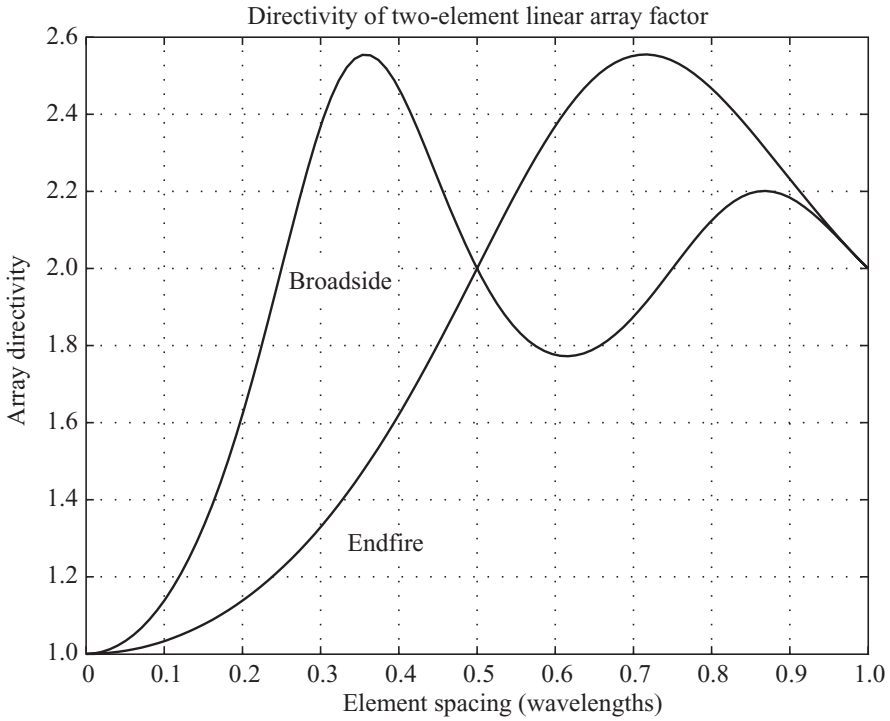


Figure 8.4.20 *The directivity of a linear array with omnidirectional elements. As the spacing approaches zero, the directivity approaches that of a single element*

8.4.11.2 Gain reduction from mutual coupling in a diversity array

In a diversity situation, the reduction in gain as the antenna becomes more compact can be found by formulating using the mutual coupling. As the antenna elements merge, the increasing mutual coupling acts to reduce the diversity gain, until the antennas merge into one and the gain has returned to that of a single element, as in the limit of the array directivity case.

At a mobile terminal, it is seen from above that two omni antennas can be spaced by only a fraction of a wavelength and still remain effectively uncorrelated (as far as the diversity gain is concerned) owing to the omni-type scenario. It is evident that as the antennas are moved closer together, the branches become increasingly correlated until the diversity gain reduces to 3 dB independent of the probability value, and also reduces further to 0 dB as the two antennas become one. The mutual coupling effects account for the changing diversity gain as the closely spaced diversity antennas merge.

8.4.11.3 Mutual impedance

The mutual impedance is defined for any M-port network, from the network relations

$$\begin{aligned} V_1 &= I_1 Z_{11} + I_2 Z_{12} + \cdots + I_M Z_{1M} \\ V_2 &= I_2 Z_{21} + I_2 Z_{22} + \cdots + I_M Z_{2M} \\ &\vdots \\ &\text{etc.,} \end{aligned} \tag{8.4.77}$$

and so a mutual impedance Z_{12} is defined directly by the voltage at port 1, the current at port 2, and all other ports open circuited:

$$Z_{12} = \left.\frac{V_1}{I_2}\right|_{I_1=0,\quad I_3=0,\ldots,I_M=0}. \tag{8.4.78}$$

The input impedance depends on the mutual coupling, for an M-element in an array,

$$Z_{\text{in}1} = Z_{11} + Z_{12}\left(\frac{I_2}{I_1}\right) + \cdots + Z_{1M}\left(\frac{I_M}{I_1}\right) \tag{8.4.79}$$

so any matching circuit has to account for the mutual coupling. In another view, coupled elements can be arranged to match an antenna (see Section 8.4.16). The network model for the mutual coupling between two antennas is shown in Figure 8.4.21.

If the measurement equipment includes a scattering parameter test set, then the measured scattering parameters, in a matrix $\boldsymbol{S}_{\text{m}}$, are converted directly to form the measured impedances matrix $\boldsymbol{Z}_{\text{m}}$ using

$$\boldsymbol{Z}_{\text{m}} = (\boldsymbol{I} - \boldsymbol{S}_{\text{m}})^{-1}(\boldsymbol{I} + \boldsymbol{S}_{\text{m}})\, Z_0 \tag{8.4.80}$$

where $\boldsymbol{I}$ is the identity matrix.

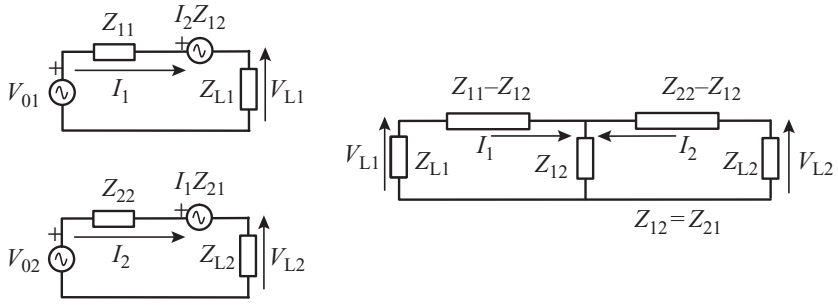


Figure 8.4.21 Network models for mutual coupling between two antennas. A condition for low correlation coefficient between antenna signals is to have low normalised mutual impedance between antenna elements, as seen from the load resistance (receiver)

8.4.11.4 Calculation of mutual impedances

For the calculation of mutual impedances for antennas, the physical situation is required of course. A technique which allows the calculation of both self and mutual impedances is the induced emf method (Section 8.1). A current is impressed at the feed of antenna 2, denoted $I_{2\text{in}}$. The electric field from antenna 2, as a function of position along the z-axis of antenna 1, is $\boldsymbol{E}_{12}(z)$, and the open circuit voltage in antenna 1 is $\boldsymbol{V}_1$. This voltage is given by the integral over the length of the antenna, of the incident electric field times the current distribution along the antenna,

$$V_1 = -\frac{1}{I_{1\text{in}}} \int_{-L_1/2}^{L_1/2} E_{12z}(z) I_{1z}(z)\, \mathrm{d}z \tag{8.4.81}$$

and the factor of $1/I_{1\text{in}}$ is the reference current at the input. The current distribution $I_{1z}(z)$ must be modelled, or calculated numerically, and the electric fields are calculated from source-field basic relations.

The mutual impedance follows as

$$Z_{12} = -\frac{1}{I_{1\text{in}} I_{2\text{in}}} \int_{-L_1/2}^{L_1/2} E_{12z}(z) I_{1z}(z)\, \mathrm{d}z. \tag{8.4.82}$$

For side-by-side dipoles, separated by distance d, and with assumed sinusoidal current distributions, the mutual impedance in free space is,

$$\begin{aligned} R_{12} &= 30\,(2\mathrm{Ci}(kd) - \mathrm{Ci}(kd_1) - \mathrm{Ci}(kd_2)) \\ X_{12} &= -30\,(2\mathrm{Si}(kd) - \mathrm{Si}(kd_1) - \mathrm{Si}(kd_2)) \end{aligned} \tag{8.4.83}$$

where Ci and Si are given in terms of common computing functions (equations (8.1.116) and (8.1.117)) in Section 8.1, and

$$d_1 = \sqrt{d^2 + L^2} + L, \quad d_2 = \sqrt{d^2 + L^2} - L. \tag{8.4.84}$$

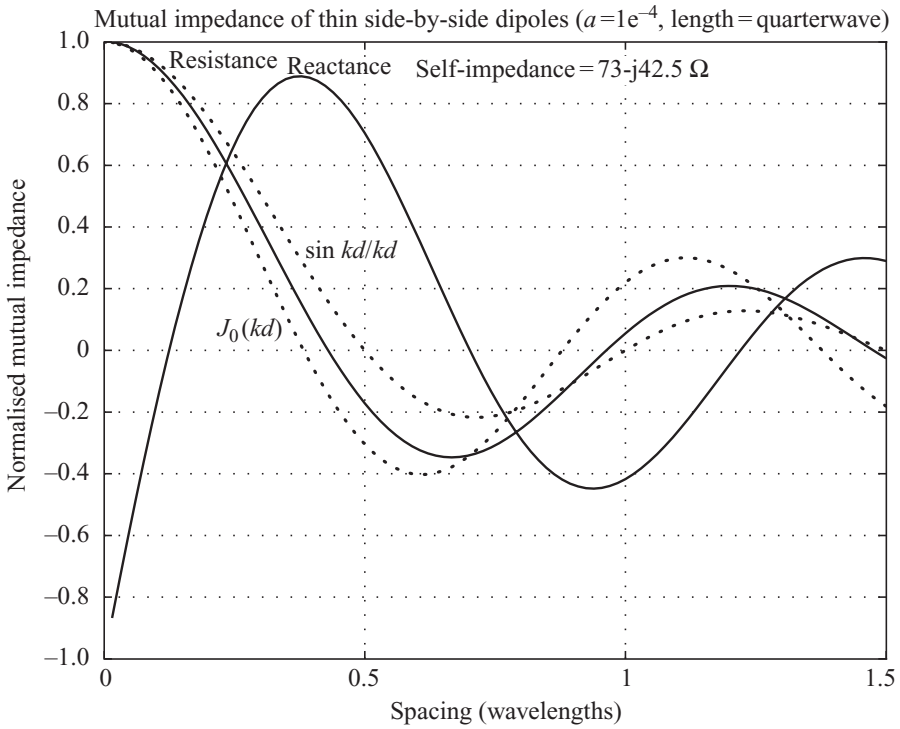


Figure 8.4.22 *The mutual impedance for side-by-side dipoles, calculated from the induced emf method, using a sinusoidal current distribution. The resistance and reactance are normalised by their respective self values. The dotted lines are open circuit voltage, or field correlation coefficient functions from scenario models, the J_0 corresponding to the Clarke 2D scenario and the* $\sin x/x$ *form corresponding to the 3D uniform scenario (see below, and also Section 6.7)*

Formulas are available (see Balanis, 1982) for collinearly spaced dipoles, and the general case of echelon-spaced dipoles, which is displacement in both sideways and vertical directions.

Figure 8.4.22 plots the above equations for spaced dipoles. The mutual impedance approaches the self-impedance as the dipoles merge. The mutual resistance is normalised to the self-resistance, $r_{12} = R_{12}/R_{11}$ and similarly for the mutual reactance, $x_{12} = -X_{12}/X_{11}$. The $J_0(kd)$ function, and also $\sin(kd)/(kd)$ are dotted onto the mutual resistance curve for discussion further below. Note that these curves are similar to the mutual resistance curve for small spacings.

8.4.11.5 Mutual impedance of monopoles

Wire monopoles are more common antenna elements than dipoles for mobile terminals. They are extremely simple to realise and as long as the groundplane is large (seldom the case) they have well-defined characteristics, mostly derived from the theory of dipoles. They are also much easier to measure and characterise experimentally than dipoles because of the coaxial feed, rather than balanced feed. There is departure

from the dipole characteristics because of the finite size of any practical groundplane. The mutual coupling between close monopoles is also slightly different to that of dipoles, owing to the feed details. Since measurements are usually with monopoles, this difference is important in ensuring that any mutual coupling measurements are correct. For self-impedances, the results from dipole analysis are accurate only if the groundplane is a reasonable size (see Section 8.2). For the mutual impedances, the approach by Taylor *et al*. (1970) correctly accounts for the impedance changes introduced by the coaxial feed arrangement, absent with the dipole analysis.

8.4.11.6 Mutual impedance from minimum scattering antennas patterns

Wasylkiwskyj and Kahn (1970) have shown that the mutual impedance between minimum scattering antennas can be expressed in terms of the far-field patterns of the antennas. Minimum scattering antennas (MSAs), discussed by Montgomery *et al*. (1948), and Kahn and Kurss (1965), comprise an idealised class of antenna with special properties. The basic property is that the antennas are electrically invisible when terminated in the appropriate reactance. If this reactance is infinite (open circuit), then the antennas are called canonical minimum scattering antennas. It follows from the basic property that the antennas must be lossless; when they are terminated in a perfectly matched load the scattered power and absorbed power is identical; when terminated in the appropriate reactance, their scattered fields are identical to their radiated fields; as long as the antenna is reciprocal, then the power pattern is symmetric in any plane through the origin; and finally, the mutual impedance can be written as a complex integral of the far-field patterns, as follows below. Most single mode antennas approximate MSAs.

The normalised mutual impedance between two MSAs spaced by $\boldsymbol{d}$, is (Wasylkiwskyj and Kahn, 1970)

$$z_{jk} = 2\int_{\phi=0}^{2\pi}\int_{C} \boldsymbol{h}_j(\theta,\phi)\cdot\boldsymbol{h}_k^*(\theta^*,\phi)\,\mathrm{e}^{j\boldsymbol{k}_0\cdot\boldsymbol{d}}\sin\theta^*\,\mathrm{d}\theta\,\mathrm{d}\phi, \tag{8.4.85}$$

where $\boldsymbol{k}_0$ is the incident propagation vector, and the patterns $\boldsymbol{h}_j$, $\boldsymbol{h}_k$, are referred to their own coordinate datums, and both have normalised total radiated power. The path of integration, C, in the complex θ-plane is from $-\pi/2 - \mathrm{j}\infty$ to $\pi/2 + \mathrm{j}\infty$.

The real part of the integration over real space will be the sole real contribution to the impedance and this corresponds to the mutual resistance. The imaginary part is of less interest since it can be tuned out at the load impedance, if necessary. Incorporating the pattern translation term into the second pattern giving a common coordinate datum for both antennas, the normalised mutual resistance is

$$r_{jk} = \frac{R_{jk}}{\sqrt{R_{jj}R_{kk}}} = \int \boldsymbol{h}_j(\Omega)\cdot\boldsymbol{h}_k^*(\Omega)\,\mathrm{d}\Omega \tag{8.4.86}$$

and the integral is over real space.

Recall that the correlation of the open circuit voltages of the antenna elements, from above, is the same as this mutual resistance equation except for the presence

of the scenario weighting function. To make ρ_{0jk} the same as r_{jk}, the patterns must be confined to the source region, and of course the assumptions regarding the source region must hold.

Most electrically small antennas such as short dipoles are considered good approximations to MSAs. So for simple, typically single mode resonant antennas, in a 3D uniform, uncorrelated scenario, or with their patterns confined to the uncorrelated source scenario,

$$r_{jk} = \rho_{0jk}, \tag{8.4.87}$$

i.e. under the above conditions, the zero- or low-correlation requirement for diversity antennas can now be stated as the requirement that the normalised mutual resistance is zero or low.

This result, that the mutual resistance can be ideally the same, and in practice similar, to the correlation coefficient (Vaughan and Andersen, 1985a,b, 1987a,b), is useful for the design and development of diversity antennas, since the mutual resistance is purely an antenna parameter, and can be measured accurately and conveniently with a network analyser. For designing uncorrelated antennas, or for evaluating diversity performance, measurement and estimation of the correlation coefficient is a more correct procedure, but also much more difficult and time consuming. The similarity between the normalised mutual resistance and the open circuit correlation functions of the scenario models is shown in Figure 8.4.22, above.

8.4.11.7 Network model for diversity antennas

An equivalent circuit for an array antenna with mutual coupling and a terminating impedance matrix is given in Figure 8.4.23. The open circuit voltages are denoted with the column vector

$$\boldsymbol{V}_{\mathrm{O}} = [V_1, V_2, \ldots, V_{\mathrm{M}}]^T \tag{8.4.88}$$

and similarly for the antenna currents, $\boldsymbol{I}_{\mathrm{A}}$, and the loaded voltages, $\boldsymbol{V}_{\mathrm{L}}$. The antenna impedance matrix, containing self- and mutual-impedances, is $\boldsymbol{Z}_{\mathrm{A}}$, and the load

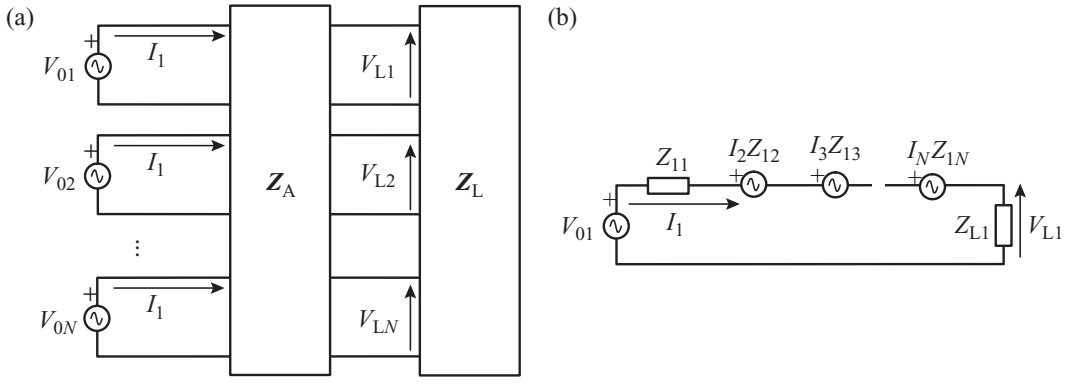


Figure 8.4.23 (a) The equivalent circuit for a terminated array antenna.(b) The equivalent circuit for a single element (denoted 1) with diagonal load impedance matrix

impedance matrix, containing the terminating resistances and any matching or mutual reactance de-coupling components, is $\boldsymbol{Z}_{\mathrm{L}}$.

For maximum power transfer, $\boldsymbol{Z}_{\mathrm{L}} = \boldsymbol{Z}_{\mathrm{A}}^{*}$, but for practical reasons, the load impedance is typically a set of 50 Ω loads directly on the antenna ports, i.e. a diagonal impedance matrix. The circuit relations are

$$\boldsymbol{V}_{\mathrm{O}} = (\boldsymbol{Z}_{\mathrm{A}} + \boldsymbol{Z}_{\mathrm{L}})\, \boldsymbol{I}_{\mathrm{A}} \tag{8.4.89}$$

$$\boldsymbol{V}_{\mathrm{L}} = \boldsymbol{Z}_{\mathrm{L}} \boldsymbol{I}_{\mathrm{A}}. \tag{8.4.90}$$

8.4.11.8 Loaded circuit and open circuit correlations

The correlation matrix of loaded, or terminated, circuit voltages is

$$\boldsymbol{L}_{\mathrm{L}} = E\{\boldsymbol{V}_{\mathrm{L}} \boldsymbol{V}_{\mathrm{L}}^{H}\} = \begin{bmatrix} \rho_{\mathrm{L}11} & \rho_{\mathrm{L}12} & \cdots \\ \rho_{\mathrm{L}12}^{*} & \rho_{\mathrm{L}22} & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} \tag{8.4.91}$$

and similarly for the open circuit voltages, $\boldsymbol{L}_{\mathrm{O}} = E\left\{\boldsymbol{V}_{\mathrm{O}} \boldsymbol{V}_{\mathrm{O}}^{H}\right\}$. From the relations above, the two correlation matrices are related by a function of the antenna and load impedances,

$$\boldsymbol{L}_{\mathrm{L}} = \boldsymbol{F} \boldsymbol{L}_{\mathrm{O}} \boldsymbol{F}^{H} \tag{8.4.92}$$

where

$$\boldsymbol{F} = \boldsymbol{Z}_{\mathrm{L}} \left(\boldsymbol{Z}_{\mathrm{A}} + \boldsymbol{Z}_{\mathrm{L}}\right)^{-1}. \tag{8.4.93}$$

This relationship allows spatial correlations, calculated from the distributions of the incident fields, to be transformed to the correlations seen at the antenna branches when the mutual coupling is significant, i.e. in compact diversity antennas. The effect of different loads on an array can be found conveniently using this equation, as well.

8.4.11.9 Received power for resistive loads

The received power is proportional to

$$\begin{aligned} P_{\mathrm{rec}} &= \left\langle \boldsymbol{V}_{\mathrm{L}}^{H} \boldsymbol{I}_{\mathrm{A}} \right\rangle \\ &= \left\langle \boldsymbol{V}_{\mathrm{L}}^{H} \boldsymbol{Z}_{\mathrm{L}}^{-1} \boldsymbol{V}_{\mathrm{L}} \right\rangle \\ &= \left\langle \boldsymbol{V}_{\mathrm{O}}^{H} \boldsymbol{Y} \boldsymbol{V}_{\mathrm{O}} \right\rangle, \end{aligned} \tag{8.4.94}$$

where

$$\begin{aligned} \boldsymbol{Y} &= \left(\boldsymbol{Z}_{\mathrm{A}} + \boldsymbol{Z}_{\mathrm{L}}\right)^{-1H} \boldsymbol{Z}_{\mathrm{L}}^{H} \left(\boldsymbol{Z}_{\mathrm{A}} + \boldsymbol{Z}_{\mathrm{L}}\right)^{-1} \\ &= \boldsymbol{F}^{H} \boldsymbol{Z}_{\mathrm{L}}^{-1} \boldsymbol{F}. \end{aligned} \tag{8.4.95}$$

If all the voltages are normalised, then this received power is, for M antenna elements,

$$\begin{aligned} P_{\text{rec}} &= \sum_{j,k}^{M} (\boldsymbol{Y})_{jk}\, \rho_{0jk} \\ &\approx \sum_{j,k}^{M} (\boldsymbol{Y})_{jk}\, r_{jk}, \end{aligned} \tag{8.4.96}$$

where the approximation applies if the assumptions discussed above hold and $r_{jk} \approx \rho_{0jk}$. For no mutual coupling and normalised average powers from the voltages, the received power reduces to

$$P_{\text{rec}}^{(\text{no mc})} = \sum_{k}^{M} (\boldsymbol{Y})_{kk}\,. \tag{8.4.97}$$

The power 'lost' owing to the mutual coupling is expressed with $P_{\text{rec}}/P_{\text{rec}}^{(\text{no mc})}$. This is the gain factor that acts to reduce the diversity gain as antennas become highly coupled. One approach for its estimate is as follows.

For real impedances and fully correlated antennas ($I_1 = I_2 \cdots = I_M$), the matched load resistance is

$$R_{\text{L}1} = R_{11} + R_{12} + \cdots + R_{1M}. \tag{8.4.98}$$

The average received power is

$$P = \left\langle \boldsymbol{V}_{\text{L}}^{H} \boldsymbol{Z}_{\text{L}}^{-1} \boldsymbol{V}_{\text{L}} \right\rangle = \frac{M}{R_{\text{L}}} \left\langle |V_{\text{L}}|^2 \right\rangle. \tag{8.4.99}$$

With the load matched, we have $V_{\text{L}} = V_{\text{O}}/2$ and the ratio of received power in the presence of mutual coupling to the received power in the absence of mutual coupling is

$$\frac{P_{\text{rec}}}{P_{\text{rec}}^{(\text{no mc})}} = \frac{\frac{M}{R_{\text{L}}(\text{with mc})} \left\langle |V_{\text{O}}|^2 \right\rangle}{\frac{M}{R_{\text{L}}(\text{with no mc})} \left\langle |V_{\text{O}}|^2 \right\rangle} = \frac{1}{1 + r_{12} + r_{13} \ldots r_{1M}}. \tag{8.4.100}$$

If all the mutual resistances are the same, which is the case for a two, and symmetric three-element antennas, and the correlation coefficient is used instead of the normalised reistances, then the above factor from this approach is $(1 + (M-1)\rho_{\text{O}})^{-1}$ for $M = 2$ or 3.

This factor can be incorporated in the cumulative distribution curves for the combined signals, from which the diversity gain is evaluated. With this power gain factor included, the diversity gain includes the effects of mutual coupling. For the three symmetric branch case, the cdf becomes

$$P_{\text{MR}}^{(3,\text{mc})}(\gamma < x) = 1 - \frac{1}{(3|\rho|)^2} \left(P_1 \text{e}^{-x} - P_2 \exp\left[-\frac{x}{\Gamma(1-|\rho|)/(1-2\rho)} \right] \right) \tag{8.4.101}$$

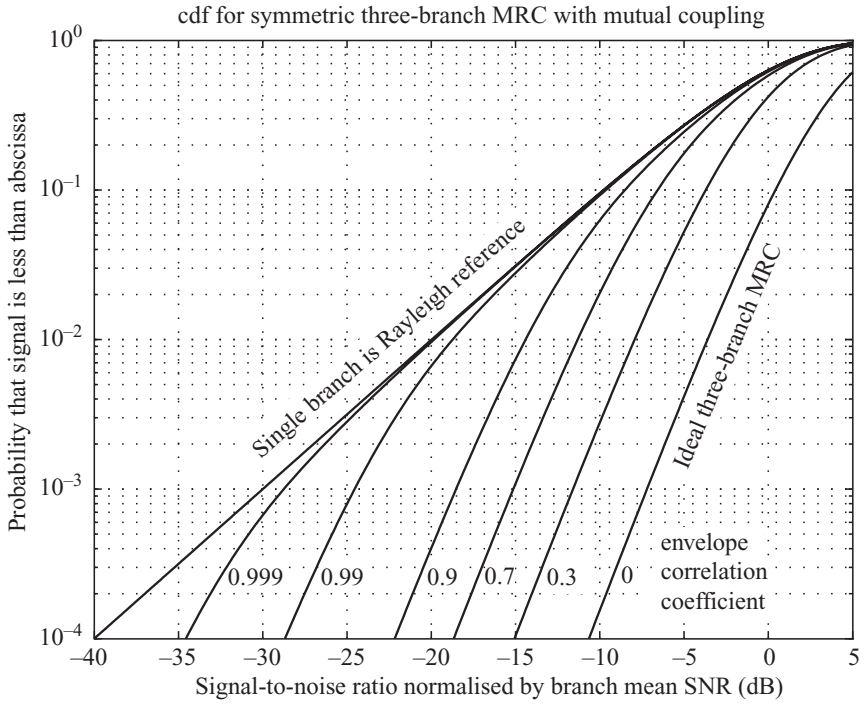


Figure 8.4.24 *The cumulative density function as a Rayleigh diagram, for the case of a symmetric three-branch MR combined signal; with envelope correlations $\rho_e \approx |\rho_0|^2$. In these curves, the effect of mutual coupling is included by setting the correlation coefficient and the mutual coupling to be the same, $r_{jk} \approx \rho_{0jk}$ and factoring in the power loss owing to the mutual coupling. The situation corresponds approximately to that of omnidirectional antenna elements in an urban/suburban or indoor environment*

where now

$$\begin{aligned} P_1 &= (1 + 2|\rho|)^2, \\ P_2 &= 3|\rho| \left(\frac{x}{\Gamma}(1 + 2\rho) + (1 - |\rho|) \right) + (1 + 2|\rho|)(1 - |\rho|). \end{aligned} \tag{8.4.102}$$

Figure 8.4.24 gives the curves for the symmetric three-branch case, and can be compared to Figure 8.4.8 above, which is the cdf for no mutual coupling.

8.4.11.10 Optimum diagonal load resistance

It is possible to optimise a diagonal load resistance, $\boldsymbol{Z}_L = \text{diag}\{R_L\}$, for maximising the received power, by solving

$$\frac{\partial P_{rec}}{\partial R_L} = \sum_{j,k}^{M} \frac{\partial Y_{jk}}{\partial R_L} \rho_{0jk} = 0. \tag{8.4.103}$$

This can be simplified with some physical conditions (the antenna details) for relating the mutual admittances to open circuit correlation coefficients. The maximum power received, with uncorrelated voltages and an optimum load (matched elements), is $M/4$ using this formulation.

8.4.11.11 Two-element array example

For a two-element array, for example, the equations are:

$$\boldsymbol{Y}^{(2)} = R_L \begin{bmatrix} R_{11} + R_L & R_{12} \\ R_{12} & R_{11} + R_{\mathrm{L}} \end{bmatrix}^{-2} \tag{8.4.104}$$

$$P_{\mathrm{rec}}^{(2)} = \frac{2r_{\mathrm{L}}}{\left((1+r_{\mathrm{L}})^2 - r_{12}^2\right)^2} \cdot \left((1+r_{\mathrm{L}})^2 + r_{12}^2 - 2\rho_{\mathrm{O12}}\, r_{12}(1+r_{\mathrm{L}})\right) \tag{8.4.105}$$

where $r_{12} = R_{12}/R_{11}$ and $r_{\mathrm{L}} = R_{\mathrm{L}}/R_{11}$. The received power is plotted in Figure 8.4.25, where the simplifying assumption is made from above that $r_{12} = \rho_{\mathrm{O12}}$,

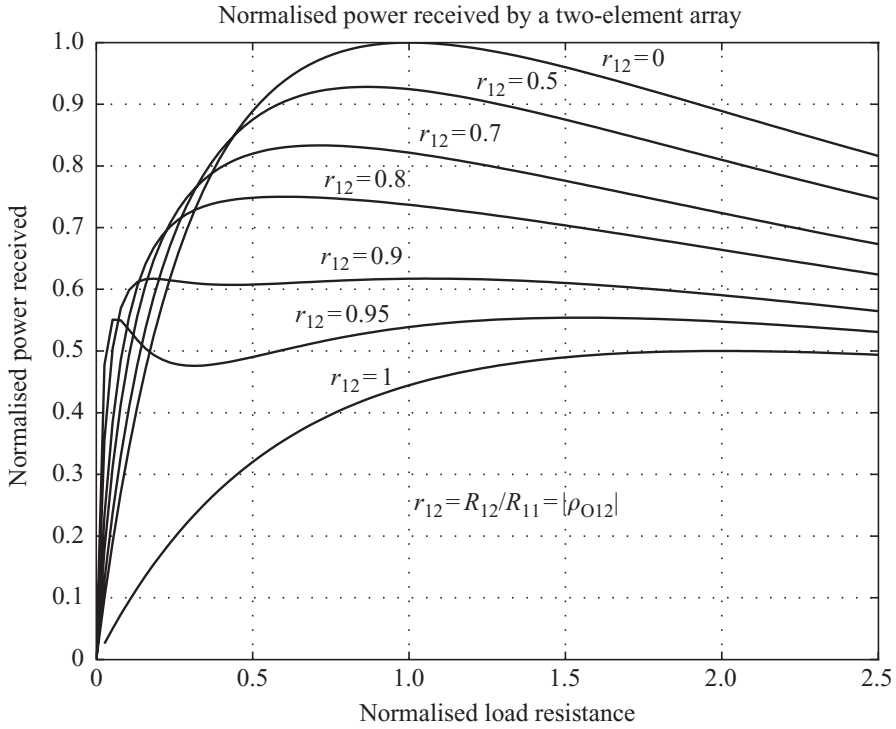


Figure 8.4.25 *For a two-element array, the received power against the normalised load resistance, which is taken to be the same as the open circuit correlation coefficient, which is shown as the graph parameter. This case is for a diagonal load matrix, meaning that there is no coupling between the branches in the load impedance. The received power is here normalised by its maximum value in the formulation of* $M/4 = 0.5$.

which allows a feel for the behaviour for the two-element diversity array. For uncorrelated branches, the best load resistance is the same as the self resistance, as expected; and for fully correlated branches, the best load is twice the self impedance, again expected since these become added in parallel as the antennas merge.

8.4.11.12 Received power using two-branch optimum load resistance

The optimum load resistance for the two-element array is found from

$$0 = \left(1 - r_{12}^2 - r_{\text{Lopt}}^2\right)\left(r_{\text{Lopt}}^2 + 2(1 - 2r_{12}^2)r_{\text{Lopt}} + (1 - r_{12}^2)\right). \qquad (8.4.106)$$

The first factor is relevant for lower values of $r_{12} = \rho_{\text{O}12}$ but heads for $(r_{12}, r_{\text{Lopt}}) = (0, 1)$ as the antennas merge into one, and so for high values the second factor provides the more correct solution, heading for the physically sensible $(r_{12}, r_{\text{Lopt}}) = (1, 2)$. The normalised optimum load resistance and the received power using the optimum load are plotted against $r_{12} = \rho_{\text{O}12}$ in Figures 8.4.26 and 8.4.27 respectively. This shows the way that the received power reduces, compensating for the otherwise-remaining 3 dB diversity gain, as the antennas merge into a single antenna.

For a symmetric three-branch circular array with all the mutual resistances and the correlation coefficients the same, the curves have a reasonably similar form to those of the two branch case. Details are given in Vaughan and Andersen (1987a).

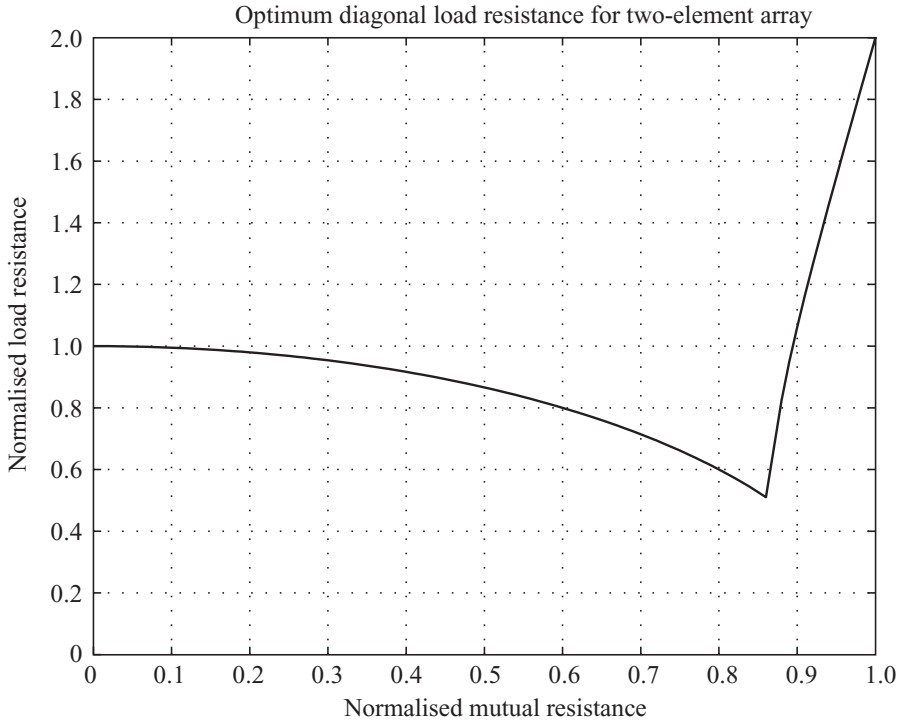


Figure 8.4.26 *For a two-element array, the optimum diagonal load resistance is plotted against the normalised mutual resistance, which is taken to be the same as the open circuit correlation coefficient*

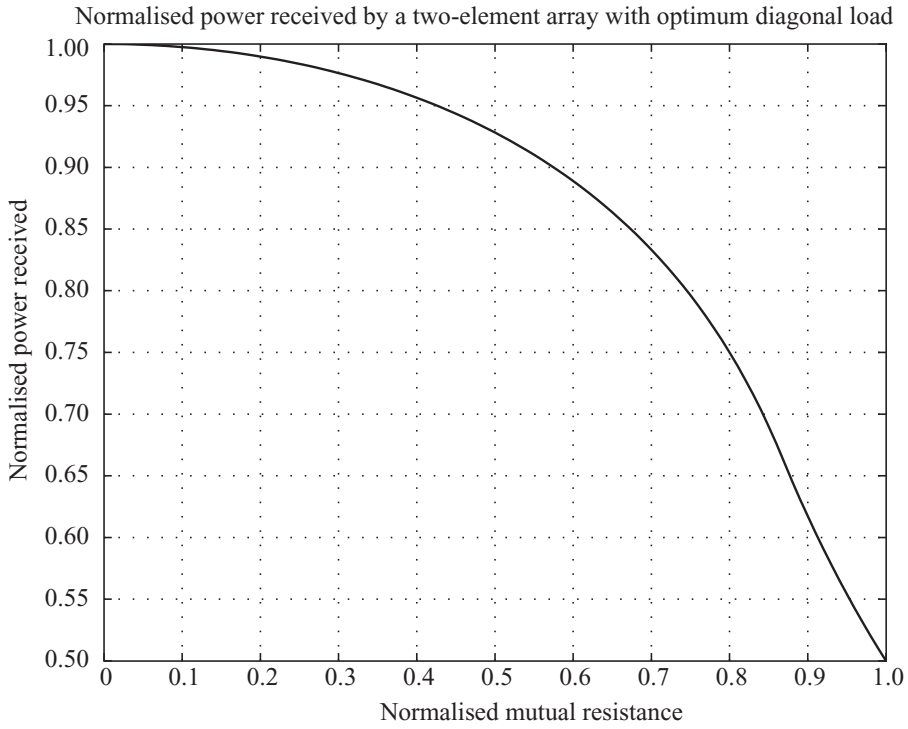


Figure 8.4.27 For a two-element array, the power received is plotted against the normalised mutual resistance, which is taken to be the same as the open circuit correlation coefficient for the load matrix being diagonal and optimised

8.4.12 Examples with measurement techniques

In the following sections, some examples are given of diversity antennas, where there is a trade-off between maintaining low correlations and keeping the size of the diversity antenna small. The sections are also used to introduce experimental techniques for evaluating the diversity action. See also the work of Pedersen *et al.* (1997, 1999), where pattern analysis is used in the evaluation of handset antennas, including the effect of the head, described in Section 8.2.

8.4.13 Example: monopoles on a groundplane

For vehicular rooftop-, panel- and even handheld chassis-mounting, the monopole is a simple, inexpensive, and effective antenna. Its bandwidth can be increased by increasing the thickness of the wire element. It is of interest to find how close such wire monopoles can be spaced, for effective diversity action. For vehicular mounting, in-line (with the direction of motion) antennas have the advantage of potentially being made the most aerodynamic of configurations. This configuration, shown in Figure 8.4.28, also allows processed measurements to confirm the theoretical results.

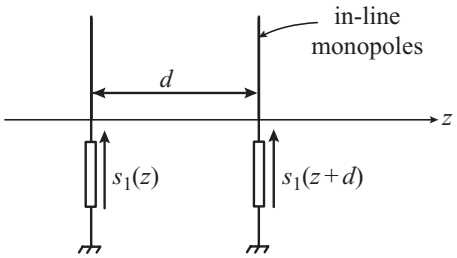


Figure 8.4.28 Two in-line monopoles allow experimentation for the spacing requirements

It is practical to use a diagonal, terminated load matrix, i.e. terminating each branch in its own load impedance. The loaded circuit correlation function, from above, depends on physical antenna spacing d and spatial cross-correlation peak displacement χ,

$$\boldsymbol{L}_{\mathrm{L}}(\chi, d) = \boldsymbol{F}(d)\ \boldsymbol{L}_{\mathrm{O}}\ (\chi, d)\ \boldsymbol{F}^{H}(d) \tag{8.4.107}$$

and here $\boldsymbol{F}(d) = \boldsymbol{Z}_{\mathrm{L}}\ (\boldsymbol{Z}_{\mathrm{A}}(d) + \boldsymbol{Z}_{\mathrm{L}})^{-1}$.

Taking the Clarke scenario, for example, the open circuit voltage matrix, or field correlation matrix, is

$$\boldsymbol{L}_{\mathrm{O}}(\chi, d) = \left\langle \boldsymbol{V}_{\mathrm{O}}(z, d)\ \boldsymbol{V}_{\mathrm{O}}^{H}(z, d+\chi) \right\rangle_z = \begin{bmatrix} J_0(kd) & J_0\left(k(d-\chi)\right) \\ J_0\left(k(d+\chi)\right) & J_0(kd) \end{bmatrix}. \tag{8.4.108}$$

This can also be estimated from closely spaced measurement samples and computer processing for the averaging. Only one of the antennas is required for such a measurement, since we are deriving all the elements from samples of an auto-correlation $L_{\mathrm{O}11} = \langle V_{\mathrm{O}1}(z) V_{\mathrm{O}1}^{*}(z+\chi) \rangle$, and in practice a single loaded element would be used for the measurement. In the measurements below, the envelope is used, rather than the complex voltage.

The loaded circuit correlation matrix comprises, here for a diagonal load,

$$L_{\mathrm{L}11}(\chi, d) = \alpha(d) L_{\mathrm{O}11}(\chi) - \beta(d) L_{\mathrm{O}12}(d, \chi) - \beta^{*}(d) L_{\mathrm{O}21}(d, \chi) \tag{8.4.109}$$

and (dropping the dependences from now on, with $\gamma(d)$ and $\epsilon(d)$ in the following),

$$L_{\mathrm{L}12} = -2\,\mathrm{Re}\{\beta\}\, L_{\mathrm{O}11} + \gamma\, L_{\mathrm{O}12} + \eta\, L_{\mathrm{O}21}. \tag{8.4.110}$$

The various terms are

$$\alpha = D\left(|Z_{\mathrm{A}12}|^2 + |Z_{\mathrm{A}11} + Z_{\mathrm{L}11}|^2 \right), \tag{8.4.111}$$

$$\beta = D\left(Z_{\mathrm{A}12}^{*}\, (Z_{\mathrm{A}11} + Z_{\mathrm{L}11}) \right), \tag{8.4.112}$$

$$\gamma = D|Z_{\mathrm{A}11} + Z_{\mathrm{L}11}|^2 \tag{8.4.113}$$

$$\eta = D|Z_{\mathrm{A}12}|^2, \tag{8.4.114}$$

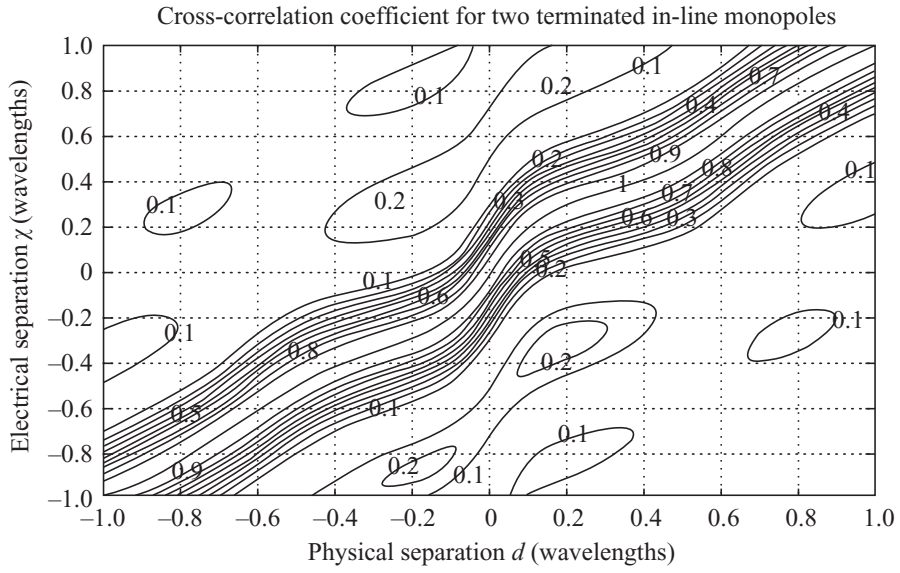


Figure 8.4.29 *The loaded circuit envelope correlation coefficient function against the physical antenna spacing and the electrical spacing showing the distortions caused by the mutual coupling*

and

$$D = \left| \frac{Z_{\mathrm{L}11}}{(Z_{\mathrm{A}11} + Z_{\mathrm{L}11})^2 - Z_{\mathrm{A}12}^2} \right|^2 . \tag{8.4.115}$$

To measure the loaded circuit correlation, the in-line configuration is used, and the measurement is a cross-correlation, $L_{\mathrm{L}12} = \left\langle V_{\mathrm{L}1}(z)\ V_{\mathrm{L}2}^*(z+\chi) \right\rangle$. In a static environment, this function peaks at a displacement which corresponds to the distance between the antennas, but in a way which depends on the antenna and load impedances. The peak value of the cross-correlation drops away as d increases (see Figure 7.1.2 in Section 7.1). Nevertheless, correlation displacement to the peak remains well-defined. The displacement, χ, is determined by measurements for different spacings, establishing numerically the position (displacement) of the peak value of the cross-correlation function, $L_{\mathrm{L}12}(\chi, d)$, and reading off the corresponding correlation displacement χ_0. This value is the electrical spacing of the antennas and it is different from the physical spacing, d, because of the mutual coupling.

A contour plot of the loaded circuit envelope correlation coefficient function, $\rho_{\mathrm{e}} \approx |\rho_{\mathrm{L}12}(\chi, d)|^2$, is given in Figure 8.4.29 (Vaughan and Scott, 1993b), the deviation from symmetry about the diagonal showing the effect of the mutual coupling. The peak of this function is given in Figure 8.4.30 (Vaughan and Scott, 1993b) with experimental results.

Small changes of the load impedances, for example matching them to $Z_{\mathrm{L}1} = Z_{\mathrm{A}11}^* + Z_{\mathrm{A}12}^*$ instead of setting them to the usual load of 50 Ω, have only a minor effect on the curve. In fact varying $R_{\mathrm{L}1}$ between 20 Ω and 60 Ω does not alter significantly the decorrelation caused by the mutual coupling at close spacings. Large variations of

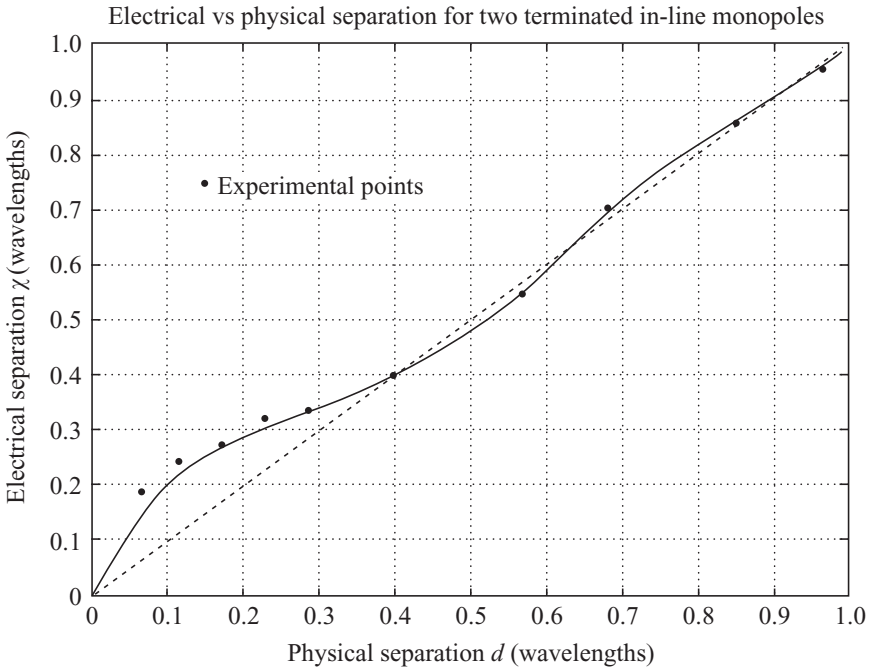


Figure 8.4.30 The ridge of the correlation function with experimental points

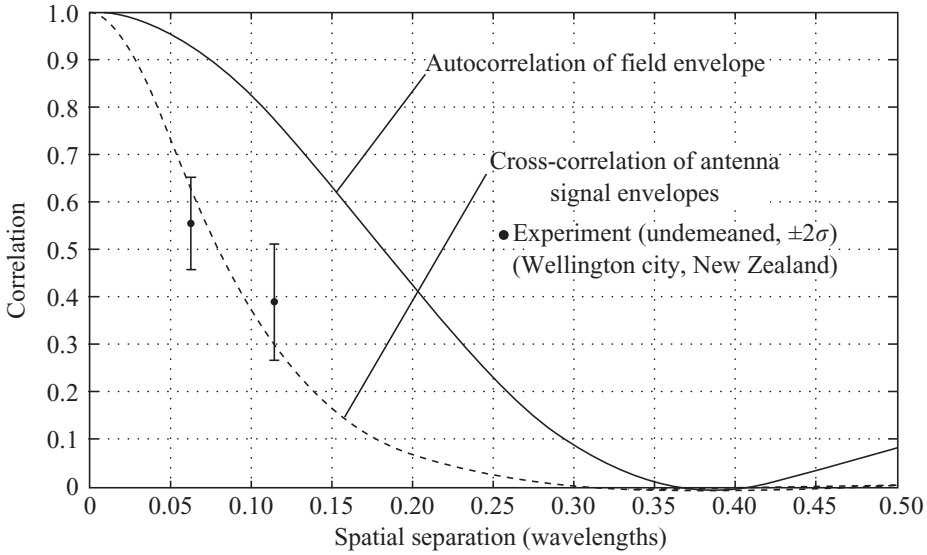


Figure 8.4.31 The loaded circuit envelope correlation coefficient function against the physical antenna spacing, with the $J_0(kd)$ function for the Clarke scenario model included. Note that at small spacings, the electrical spacing is approximately halved relative to the physical spacing, indicating a much smaller spacing for diversity antennas than suggested by the fields correlation function

the load reactance X_{L1} change the form of the curve, but modest variations, between about $-40\,\Omega$ and $+40\,\Omega$, do not affect the results much.

The cut through the contour plot at $\chi = 0$ is the spatial envelope cross-correlation coefficient function, $\rho_{e12}\,(\chi = 0, d)$. This is given in Figure 8.4.31 (Vaughan and Scott, 1993b). This can be used directly to find acceptable antenna spacings for diversity. The experimental points are from a large measurement campaign in a city environment, and the $\pm 2\sigma$ bars (95 per cent) for the correlation indicate the tight fit (cf. Section 7.6). The data for this are undemeaned, so the experimental correlations are likely to be pessimistic in the sense of having a higher value that for purely Rayleigh fading. (The changing mean, or shadow-fading component, is common, and therefore correlated, between the antennas.) The Clarke scenario envelope correlation coefficient function, $J_0^2(k_0 d)$, is included. The difference between the curves has significant implications for compact space diversity. For envelope correlation coefficients of 0.7 or 0.5, the correlation spacings are only 0.05λ and 0.08λ respectivley. These are the acceptable spacings for two monopoles on a large groundplane, to provide good diversity operating in an urban/suburban environment.

8.4.13.1 Four in-line monopoles

The question now arises, can the small spacing be maintained for more elements than two? For four equispaced elements, the envelope correlation coefficient functions are given in Figure 8.4.32 (Vaughan and Scott, 1992). The functions show that the limiting spacing is that between the inner pair. The inner pair also have lower mean effective gains. This suggests a configuration of an uneven spacing for optimum performance. However, it should be remembered that these theoretical/modelled curves are only to offer design guidelines for low correlations and similar branch powers. Fine tuning of a system for which many assumptions have been made will probably not change the measured diversity gain significantly. In particular, the finite size of a real-world groundplane may have a strong effect on these correlations.

For a handheld terminal with monopole elements, an extremely small 'groundplane' comprises a facet of the chassis, and this behaves as the other part of an asymmetric dipole. The chassis is common to all the elements so much higher cross-correlations can be expected than when evaluated using a large groundplane. Experimental measures for evaluating the diversity gain are called for.

8.4.13.2 Experimental estimation of diversity gain by equivalent number of ideal branches

Handset diversity including the effect of the head has been discussed by Pedersen and Andersen (1999); see Section 8.2. The closely spaced monopole on a handheld chassis design has been tested for an indoor environment without head effects in (Nørklit *et al.*, 2001). Here, the method used was also by mobile multiport signal logging and analysis, but this time using a circular servo to move the antennas, rather than a vehicle. This indoor measurement set-up is shown in Figure 8.4.33.

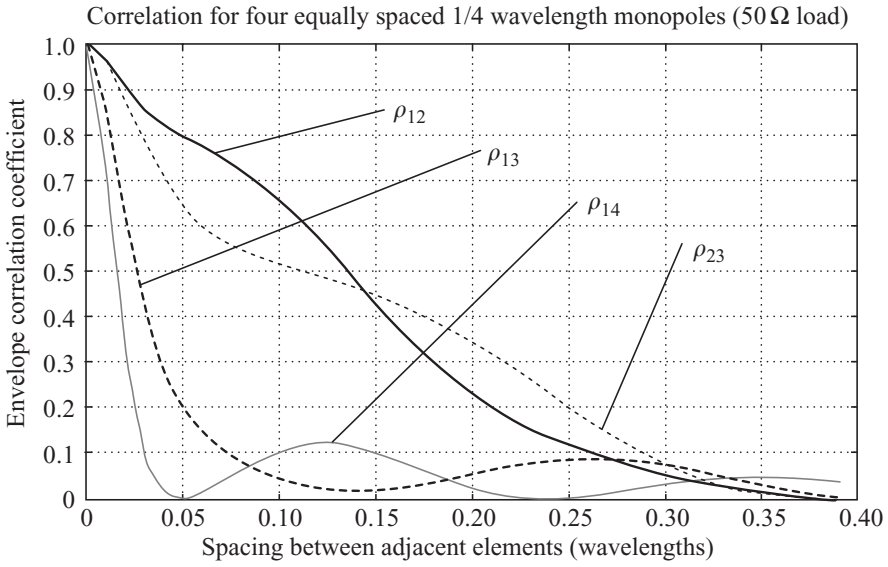


Figure 8.4.32 *The envelope correlation coefficient for an array of four monopoles on an infinite groundplane (taken from the relations for dipoles in free space). The correlations between the inner elements are seen to decrease more slowly than those of the outer elements*

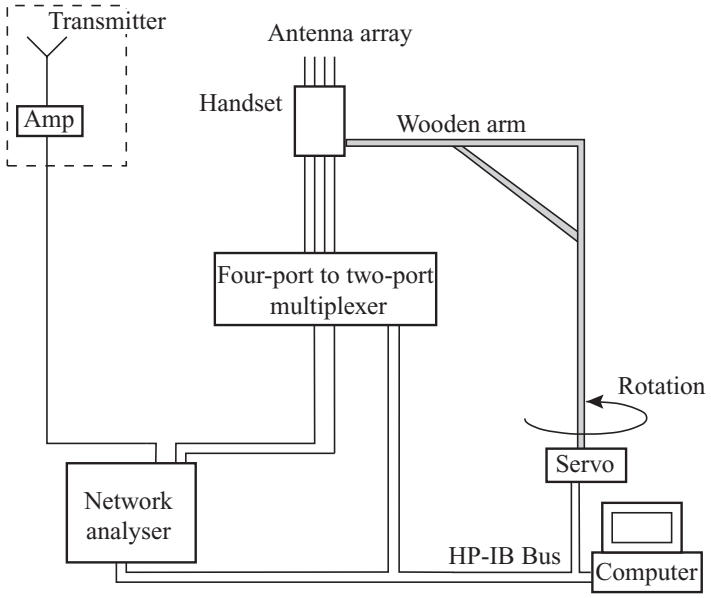


Figure 8.4.33 *An experimental set-up for four closely spaced monopoles on a hand-held terminal, moved in an indoor scenario by rotation of a boom. The signals are logged and processed for estimating useful design parameters such as the mean gains of the elements, the correlations, and the diversity gain*

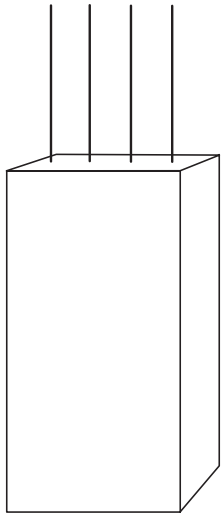


Figure 8.4.34 Four closely spaced monopoles on a handheld chassis. Note the very small groundplane, with a chassis which acts as an asymmetric element (making a dipole-like arrangement with the monopole) which is common to all the monopoles

The candidate diversity design is moved through space (rotated on a boom in this case) and all the branch signals are densely sampled. The time series are processed to estimate the mean effective gains and the correlations. However, even with the mean gains and correlation coefficients established, there is no simple equation for finding the diversity gain, since the correlations are non-zero and different, and the branch powers are different.

One way to estimate the diversity gain in these circumstances is to combine the branch signals in the computer and then calculate the diversity gain from the cumulative distribution of the combined signal. The diversity gain is read off as shown in Figure 8.4.2, above. The design example is given in Figure 8.4.34, where the four-element spacing is 0.06λ, making a linear aperture width of 0.18λ. A similar array of monopoles can be placed along the bottom of a short handset case, and these are all uncorrelated with all of those on top (Vaughan and Scott, 1992).

Here the measurement frequency is 1.9 GHz, the monopoles are 10 mm apart and the aperture across the feeds is 32 mm. The elements are mounted on a chassis $50\,\text{mm} \times 80\,\text{mm} \times 100\,\text{mm}$, with antennas mounted on a facet of $50\,\text{mm} \times 80\,\text{mm}$. The diversity gain, defined for the 0.5 per cent level, is measured to be the equivalent of only 2.5 maximum ratio, uncorrelated branches. A three-element version, occupying the same linear aperture (now the element spacing is 0.1λ, or 16 mm) has a diversity gain of 2.3 such branches. This indicates that for a linear array, an array length of 0.18λ is well filled (in a performance sense) with three, rather than two or four elements. The four-element array can be arranged to give almost the full four-branch diversity gain, through bending the monopoles apart, a technique pursued below originally for a rooftop vehicular application.

8.4.14 Example: sloping monopoles on a groundplane

The mutual coupling circuit model analysis above showed the close spacings that are possible for omni-type scenarios such as indoor and urban, for what is essentially purely space diversity. (In fact if the patterns are measured from each of the elements, with the other elements terminated, the amplitudes are somewhat different, owing to the mutual coupling. If each pattern is measured with the other elements open circuited, then the patterns are essentially the same because the elements are essentially minimum scattering antennas.) However, it is clear that no direct use has been made of the polarisation, and not much with the amplitude degrees of freedom for assisting with the pattern orthogonality. Although the elements are identical and have stand-alone omnidirectionality in 2D, a pattern measurement based on each element has amplitude variation owing to the presence of the other terminated elements. By bending the monopoles in different directions and angles, the polarisation, amplitude and phase are all used for the pattern orthogonality (Vaughan *et al.*, 1988). An advantage is that the feed positions can be made more closely spaced than with purely space diversity, for a given open circuit correlation. A special case of bending the monopoles apart is the circular array configuration, and for three elements, the mutual couplings are all the same. This section approaches diversity evaluation using the mutual resistances and pattern orthogonality rather than by experimental time series analysis of the branch signals.

8.4.14.1 Element patterns

The analysis starts with finding the element pattern for the sloping monopole on a groundplane, which is the half-space equivalent of a bent dipole. The pattern of the monopole along the vertical z-axis in free space (no groundplane), $\boldsymbol{h}_{\mathrm{M}z}(\theta, \phi) = h_{\mathrm{M}z}(\theta)\hat{\theta}$, is rotated to a new coordinate system, namely from (x, y, z) to (x', y', z') to have an elevation angle α from the y-axis. Figure 8.4.35 shows the geometry. The monopole pattern is found by the transform $\boldsymbol{h}_{1\mathrm{S}}(\theta, \phi) = \boldsymbol{h}_{\mathrm{M}z}(\theta', \phi')$. Finally, the groundplane is added and the mirror image of the sloping monopole is added and the pattern of the sloping monopole on a groundplane is found.

The coordinate transformations are

$$x' = x, \quad y' = y\sin\alpha - z\cos\alpha, \quad z' = y\cos\alpha + z\sin\alpha; \tag{8.4.116}$$

and in the spherical systems,

$$z' = r, \tag{8.4.117}$$

and from $r = r'\cos\theta'$, the following relations are required:

$$\begin{aligned} \cos\theta' &= r\sin\theta\cos\phi\cos\alpha + r\cos\theta\sin\alpha, \\ \hat{\theta}'\sin\theta' &= (\sin\theta\sin\alpha - \cos\theta\sin\phi\cos\alpha)\,\hat{\theta} - (\cos\phi\cos\alpha)\hat{\phi}. \end{aligned} \tag{8.4.118}$$

The length of the monopole is L_{M}. (Recall that for a dipole analysis, the monopole length is denoted $L/2$.) It is again assumed that the current distribution is sinusoidal. The pattern of the monopole alone, i.e. with no groundplane, is found by the method

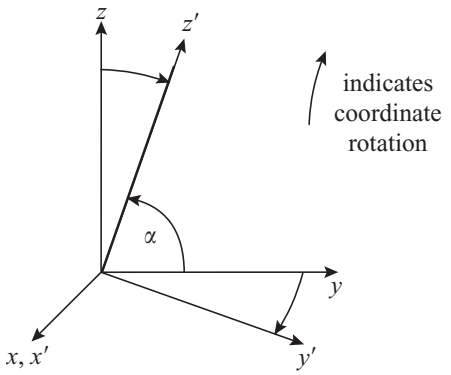


Figure 8.4.35 *The coordinate rotation from (x, y, z) to (x', y', z') for monopoles which are sloped and rotated. Here the rotation is about the x-axis, such that the monopole forms an elevation angle α with the y-axis, or groundplane*

in Section 8.1, that is, by calculating the magnetic vector potential from the assumed current distribution, and integrating. The result is

$$\boldsymbol{h}_{\mathrm{M}z}(\theta) = \mathrm{j}Z_0 I_0 \frac{\mathrm{e}^{-jk_0 r}}{4\pi r}\left(\frac{\mathrm{e}^{-\mathrm{j}k_0 L_{\mathrm{M}}\cos\theta} - \mathrm{j}\cos\theta\sin(k_0 L_{\mathrm{M}}) - \cos(k_0 L_{\mathrm{M}})}{\sin\theta}\right)\hat{\theta}. \tag{8.4.119}$$

For a monopole on the xy-plane and sloping along the yz-plane in the positive y-direction with elevation angle α, the element pattern transforms to

$$\boldsymbol{h}_{1\mathrm{s}}(\theta,\phi) = \mathrm{j}Z_0 I_0 \frac{\mathrm{e}^{-\mathrm{j}k_0 r}}{4\pi r}\left(\frac{\mathrm{e}^{\mathrm{j}k_0 L_{\mathrm{M}} U_{\mathrm{S}}} - \mathrm{j}U_{\mathrm{S}}\sin(k_0 L_{\mathrm{M}}) - \cos(k_0 L_{\mathrm{M}})}{1 - U_{\mathrm{S}}^2}\right)\boldsymbol{W}_{\mathrm{S}} \tag{8.4.120}$$

where

$$\boldsymbol{W}_{\mathrm{S}} = -V_{\mathrm{S}}\hat{\theta} - \cos\phi\cos\alpha\hat{\phi} \tag{8.4.121}$$

and

$$\begin{aligned} U_{\mathrm{S}} &= \sin\theta\sin\phi\cos\alpha + \cos\theta\sin\alpha, \\ V_{\mathrm{S}} &= \cos\theta\sin\phi\cos\alpha - \sin\theta\sin\alpha. \end{aligned} \tag{8.4.122}$$

The pattern contribution from the monopole's mirror image, $\boldsymbol{h}_{1\mathrm{S}}^{(\mathrm{image})}$, is found by the substitutions $L_{\mathrm{M}} = -L_{\mathrm{M}}$ and $\alpha = \pi - \alpha$, giving

$$\boldsymbol{h}_{1\mathrm{S}}^{(\mathrm{image})}(\theta,\phi) = \mathrm{j}Z_0 I_0 \frac{\mathrm{e}^{-\mathrm{j}k_0 r}}{4\pi r}\left(\frac{\mathrm{e}^{\mathrm{j}k_0 L_{\mathrm{M}} U_{\mathrm{M}}} - \mathrm{j}U_{\mathrm{M}}\sin(k_0 L_{\mathrm{M}}) - \cos(k_0 L_{\mathrm{M}})}{1 - U_{\mathrm{M}}^2}\right)\boldsymbol{W}_{\mathrm{M}} \tag{8.4.123}$$

where

$$\boldsymbol{W}_{\mathrm{M}} = -V_{\mathrm{M}}\hat{\theta} + \cos\phi\cos\alpha\hat{\phi} \tag{8.4.124}$$

and

$$\begin{aligned} U_{\mathrm{M}} &= \cos\theta \sin\alpha - \sin\theta \sin\phi \cos\alpha, \\ V_{\mathrm{M}} &= \cos\theta \sin\phi \cos\alpha + \sin\theta \sin\alpha. \end{aligned} \tag{8.4.125}$$

The total pattern is then

$$\boldsymbol{h}_{1\mathrm{S}}^{(\mathrm{total})} = \boldsymbol{h}_{1\mathrm{S}} + \boldsymbol{h}_{1\mathrm{S}}^{(\mathrm{image})}. \tag{8.4.126}$$

For numerical calculations, the values corresponding to $U_{\mathrm{S}}, U_{\mathrm{M}} = \pm 1$ need to be treated with care. The problem terms have limits:

$$\begin{aligned} &\left\{ \underset{U_S = \begin{bmatrix} + \\ - \end{bmatrix} 1}{\mathrm{Lim}} \right\} \left(\frac{\mathrm{e}^{\mathrm{j}k_0 L_{\mathrm{M}} U_{\mathrm{S}}} - \mathrm{j}U_{\mathrm{S}} \sin(k_0 L_{\mathrm{M}}) - \cos k_0 L_{\mathrm{M}}}{1 - U_{\mathrm{S}}^2} \right) \\ &\qquad = \begin{bmatrix} - \\ + \end{bmatrix} \frac{1}{2} \left(\mathrm{j}k_0 L_{\mathrm{M}} \cos(k_0 L_{\mathrm{M}}) \begin{bmatrix} - \\ + \end{bmatrix} \left(k_0 L_{\mathrm{M}} \begin{bmatrix} + \\ - \end{bmatrix} \mathrm{j} \right) \sin(k_0 L_{\mathrm{M}}) \right), \end{aligned} \tag{8.4.127}$$

and similarly for $U_M = \pm 1$ but with signs [] reversed.

With the element pattern for the sloping monopole, the way is now clear to rotate and translate the patterns to get the patterns of the other elements in an array. For example, three outward sloping monopoles, arranged in a circular array with spacing $|d| = |d_{12}| = |d_{13}|$, have patterns

$$\begin{aligned} \boldsymbol{h}_{2\mathrm{S}}(\theta, \phi;\ \alpha, L_{\mathrm{M}}) &= \boldsymbol{h}_{1\mathrm{S}}^{(\mathrm{total})}(\theta, \phi - 120^\circ;\ \alpha, L_{\mathrm{M}})\ \mathrm{e}^{-\mathrm{j}\boldsymbol{k}_0 \cdot \boldsymbol{d}_{12}} \\ \boldsymbol{h}_{3\mathrm{S}}(\theta, \phi;\ \alpha, L_{\mathrm{M}}) &= \boldsymbol{h}_{1\mathrm{S}}^{(\mathrm{total})}(\theta, \phi - 240^\circ;\ \alpha, L_{\mathrm{M}})\ \mathrm{e}^{-\mathrm{j}\boldsymbol{k}_0 \cdot \boldsymbol{d}_{13}}. \end{aligned} \tag{8.4.128}$$

These patterns can be substituted into the formulas for diversity design and the parameters of length, spacing distance, and sloping angle, adjusted for a suitable trade-off for the scenario at hand.

8.4.14.2 Distributed directivity

The distributed directivity of the sloping monopole with a sinusoidal current distribution into the vehicular scenario is

$$D_{1\mathrm{S}} = \frac{\int_{\theta=\pi/3}^{\pi/2} \int_{\phi=0}^{2\pi} \left(|h_\theta(\theta, \phi)|^2 + |h_\phi(\theta, \phi)|^2 \right) \sin\theta \ \mathrm{d}\theta \ \mathrm{d}\phi}{\int_{\theta=0}^{\pi/2} \int_{\phi=0}^{2\pi} \left(|h_\theta(\theta, \phi)|^2 + |h_\phi(\theta, \phi)|^2 \right) \sin\theta \ \mathrm{d}\theta \ \mathrm{d}\phi} \tag{8.4.129}$$

where the vertically polarised pattern for example is $h_\theta(\theta, \phi) = h_{1\mathrm{S}\theta}^{(\mathrm{total})}(\theta, \phi)$. This distributed directivity, referenced to that of a vertical quarter wavelength monopole, is given in Figure 8.4.36 (Vaughan *et al.* 1988). The variation of the directivity with sloping angle is significant for the longer monopoles. For near-vertical monopoles (slope angle, or elevation angle, larger than about 70°), the pattern is more confined to the low angles of the vehicular scenario than the pattern for a quarter-wavelength monopole and so the distributed directivity is higher for these cases than for near-horizontal dipoles.

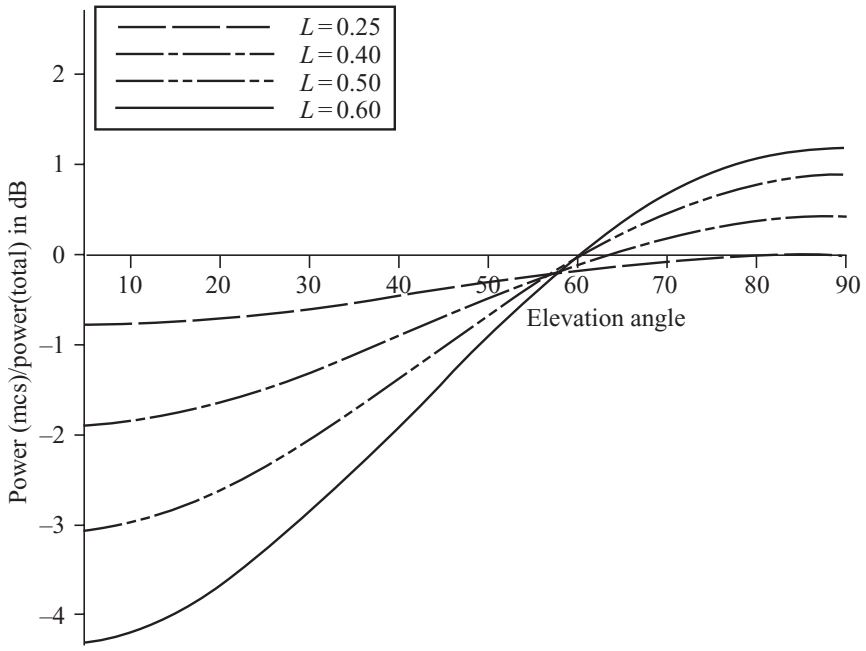


Figure 8.4.36 The distributed directivity in dB for the sloping monopole on an infinite groundplane for the vehicular scenario. The reference value (0 dB) is for a vertical quarter wave monopole, which corresponds to about 73 per cent of the power in the wanted directions. At near-vertical sloping angles, the longer antennas have an advantage over short ones because their pattern is more confined to the lower angles of the vehicular scenario

8.4.14.3 Radiation resistance

Using the techniques of Section 8.1, the radiation resistance of a sloping monopole is calculated from the far fields of the monopoles on an infinite groundplane,

$$\frac{1}{2}\,|I(0)|^2\,R_{11} = \frac{1}{2}\iint \left(\boldsymbol{E}\times\boldsymbol{H}^*\right) r^2 \sin\theta\;\mathrm{d}\theta\;\mathrm{d}\phi. \tag{8.4.130}$$

Setting $\boldsymbol{E}(\theta,\phi) = h_{1\mathrm{S}}^{(\mathrm{total})} = \mathrm{j}I(0)Z_0G(r)\boldsymbol{F}(\theta,\phi)$,

$$R_{11}^{(1\mathrm{S,total})} = \frac{Z_0}{16\pi^2}\int_{\theta=0}^{\pi/2}\int_{\phi=0}^{2\pi}\left|\boldsymbol{F}_{1\mathrm{S}}^{(\mathrm{total})}\right|^2 \sin\theta\;\mathrm{d}\theta\;\mathrm{d}\phi. \tag{8.4.131}$$

This expression is the self impedance referred to the current maximum, and so for finding the feedpoint impedance, the quantity must be divided by $\sin^2(k_0L)$. The expression is plotted in Figure 8.4.37 (Vaughan *et al.*, 1988). The limiting values in the figure are as expected, heading for zero as the monopole bends to the groundplane, and to the correct dipole-equivalent values for a vertical monopole.

For a specific antenna with a specific wire thickness, the method of moments, where no assumption is made regarding the current distribution, can also be used

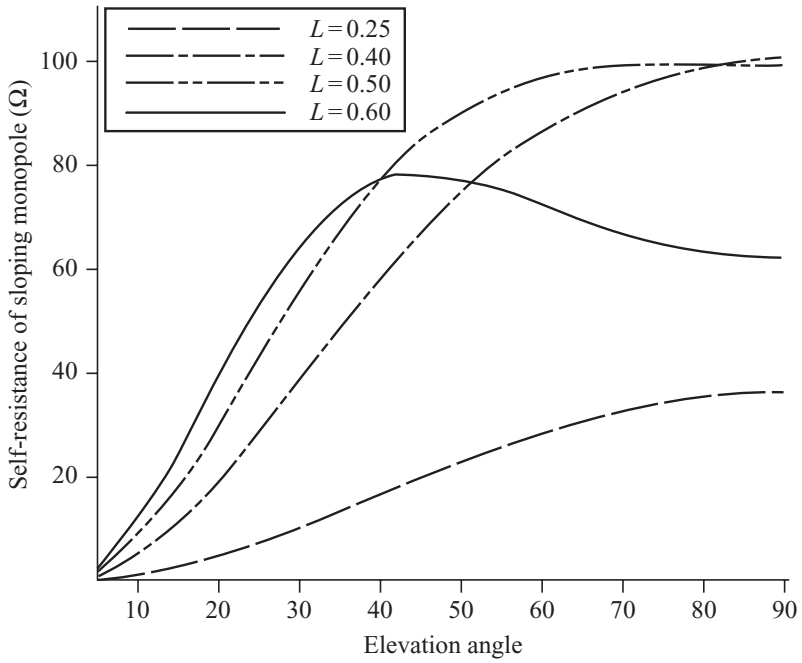


Figure 8.4.37 *The radiation resistance of sloping monopoles on an infinite groundplane, as calculated using the integral of the power pattern, which is estimated by assuming sinusoidal current distribution. The values here are referred to the current node, and need to be divided by* $\sin^2(k_0 L)$ *to get the feedpoint impedance*

for finding the feedpoint impedance. The wires need to be very thin for most moment method solutions to be reliable. Specific wire thicknesses will have different impedances. Numerical far-field patterns can be found reliably from the method of moments. An example of the self-resistance of a sloping monopole, found from the moment method, is given in Figure 8.4.38, for a wire radius of 0.00001λ. Here the wire is impracticably thin, but thin enough for the resistive part of the moment method to be reliable. Although the input impedance can be calculated reliably from the theoretical techniques for short monopoles (up to about a quarter wavelength), this is not the case for longer ones. For the vertical 0.25λ long monopole, the moment method gives $Z_{\text{IN}} = 39 + \text{j}21.6\,\Omega$, and for the 0.4λ monopole, it gives $Z_{\text{IN}} = 408 + \text{j}912\,\Omega$. This latter figure has been measured as $415 + \text{j}300\,\Omega$, for a wire of radius 0.00005λ and on a finite groundplane of electrical diameter about 13 wavelengths. For the 0.4λ vertical monopole of Figure 8.4.37 (Vaughan *et al.*, 1988), the feedpoint impedance is about $280\,\Omega$ (note that this calculation does not include wire thickness), which is quite different from the value for a very thin monopole.

The pattern orthogonality for a pair of elements from the three-element circular array, in the form of the normalised mutual impedance is given in Figure 8.4.39 (Vaughan *et al.*, 1988) for a relatively large spacing of 0.1λ. Here, to illustrate the techniques, it is again assumed that the elements behave as minimum scattering antennas and have sinusoidal current distributions so that the patterns derived above

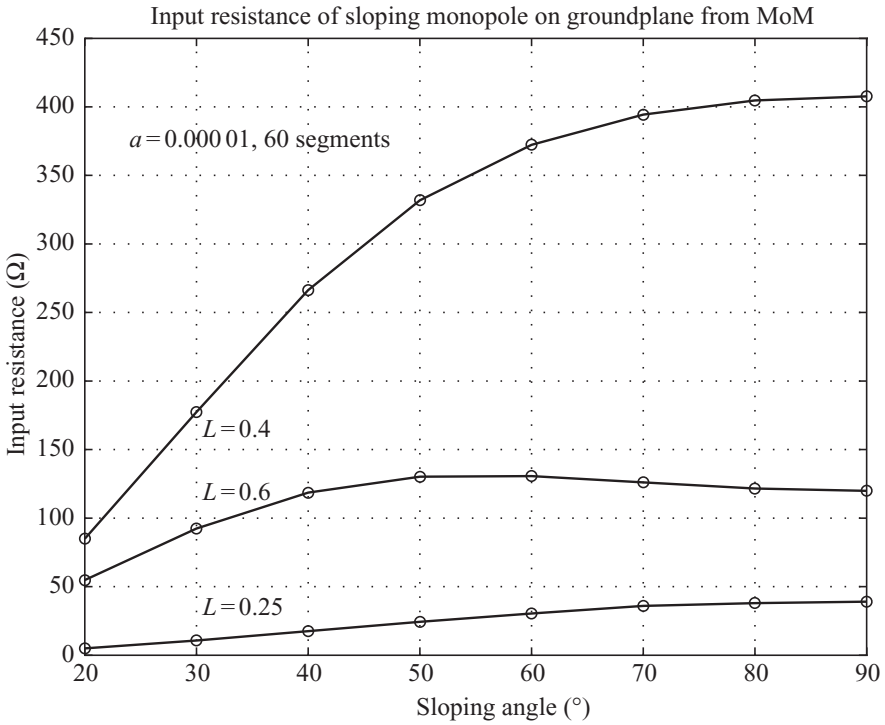


Figure 8.4.38 The self resistance of sloping monopoles as calculated by a moment method (NEC II) as seen from the feed position. The wire radius $a = 0.00001\lambda$, and there are 60 segments. For accurate impedance of the vertical monopole, see Figure 8.1.38, Section 8.1

can be used. For a finite groundplane, and for longer monopoles, the moment method would be a more accurate approach to get the patterns and the mutual coupling. The mutual resistance is

$$R_{12}^{(1S,2S,\text{total})} = \frac{Z_0}{16\pi^2} \int_{\theta=0}^{\pi/2} \int_{\phi=0}^{2\pi} \boldsymbol{F}_{1S}^{(\text{total})} \cdot \boldsymbol{F}_{2S}^{(\text{total})^*} \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi. \qquad (8.4.132)$$

For the short monopole, such as the quarter wavelength monopole, the normalised mutual resistance is rather insensitive to its elevation angle (as long as the angle is over about 60°), suggesting that the diversity action is not overly sensitive to this parameter, for this spacing and separation angle. Increasing the length of the monopoles provides a good mechanism for decorrelating the antennas, and this is because the effective spacing between them is being increased.

8.4.14.4 Correlation coefficient

With the mutual resistances known (either from the above calculations or from a measurement), and the open circuit voltage correlation coefficient function calculated from scenario models (or estimated from measurements), the loaded circuit voltage correlation coefficient function can be calculated.

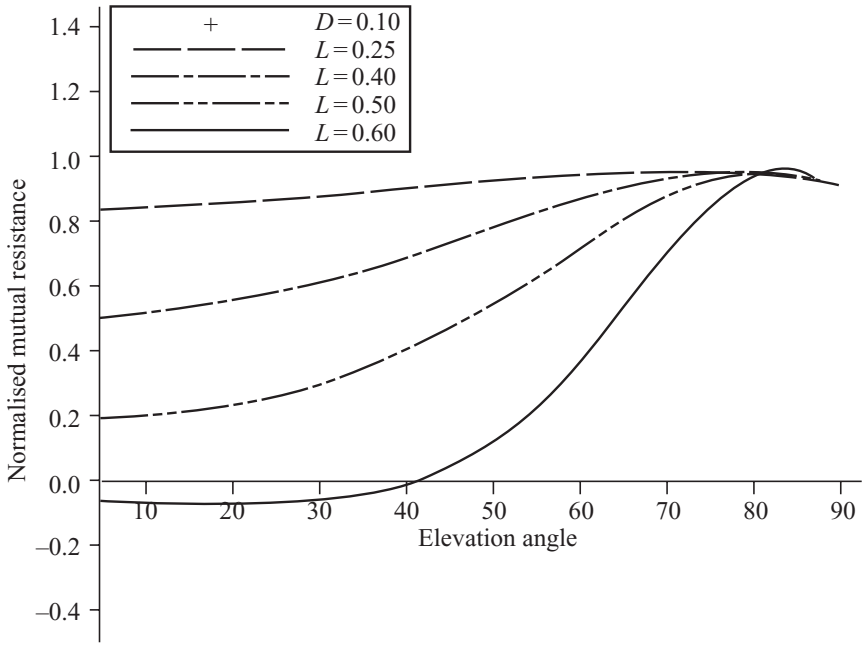


Figure 8.4.39 The normalised mutual resistance between two elements of a three-element circular array of outward sloping monopoles on an infinite groundplane, as calculated from the integral of the patterns

This is shown in Figure 8.4.40 (Vaughan *et al.*, 1988) for a vehicular scenario (Section 6.7), and is seen to momentarily decrease as a pair of the elements approaches being parallel. This behaviour is only hinted at in the normalised mutual impedance curves. The effect is at the sensitive values of correlation, where it drops from maxima of well over 0.84 (this corresponds to the larger values where the diversity gain has started to drop off very quickly) to about 0.7, suggesting an envelope correlation coefficient of about 0.5 which is sufficiently low for good diversity gain.

For a two-element array, the decorrelation of the received signals caused by bending the monopole elements apart is shown in Figure 8.4.41 (Vaughan *et al.*, 1988). A two element diversity antenna can be built with very close feedpoint spacing. However, the antenna is more expansive owing to the bent nature of the monopoles.

8.4.14.5 Diversity gain

The diversity action can be confirmed by calculating the diversity gain, since there is a formula for three identical branches with identical mutual correlations. Figure 8.4.42 (Vaughan *et al.*, 1988) shows the 0.5 per cent maximum ratio diversity gain, calculated with reference to one of the elements of the array, *operating in the presence of the other elements*, so that the mutual coupling effects are included. The calculation is also for the vehicular scenario. The maximum diversity gain from three MRC branches for this probability is almost 25 dB, from Figure 8.4.6. Here the diversity gains are mostly over 18 dB, the maximum value for two branch MRC. When the three elements are

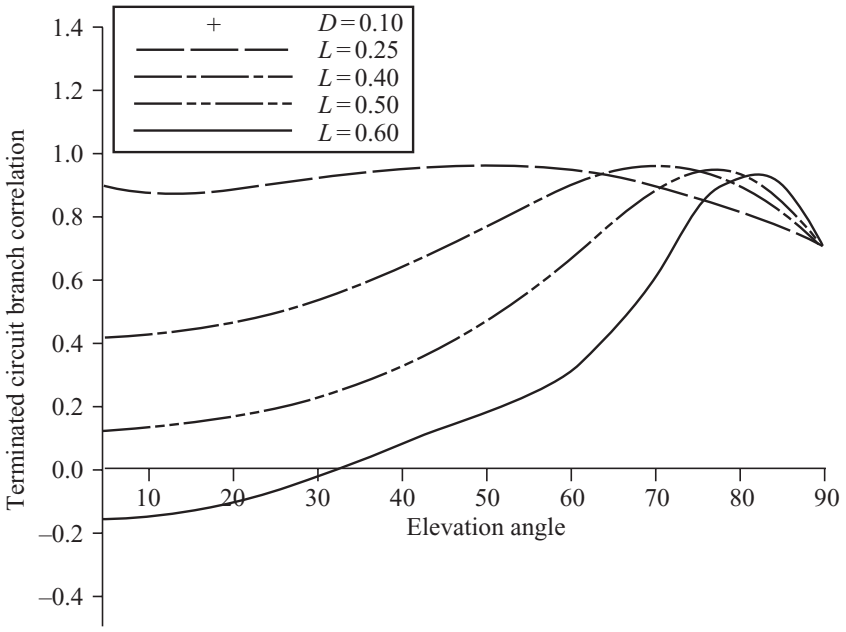


Figure 8.4.40 *The terminated circuit correlation coefficient for two elements of a circular array of three outward sloping monopoles on a large groundplane. The feedpoint spacing is 0.1λ. Good diversity action can be expected for values less than about 0.8*

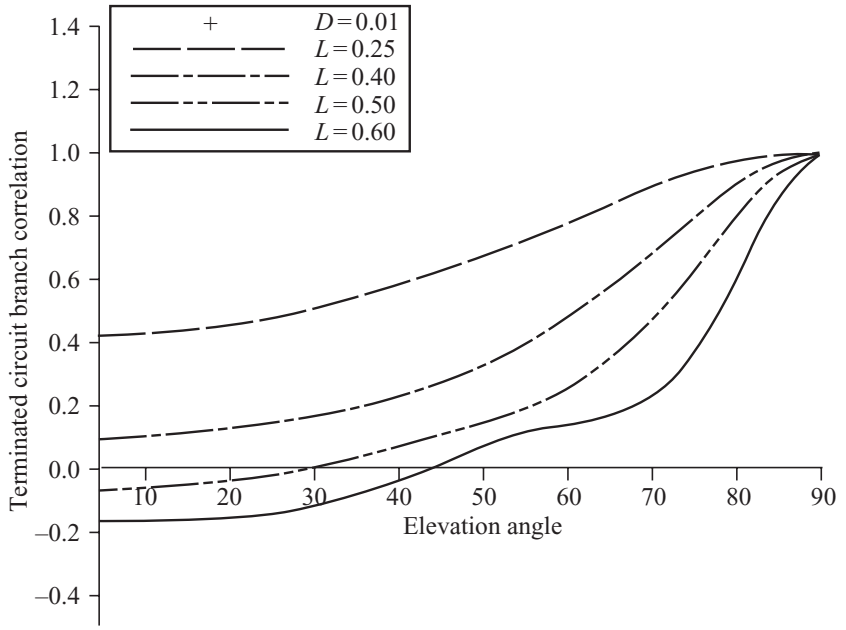


Figure 8.4.41 *The terminated circuit correlation coefficient for the two elements symmetrically bent apart (two-element circular array) on a large groundplane. The feedpoint spacing is 0.01λ*

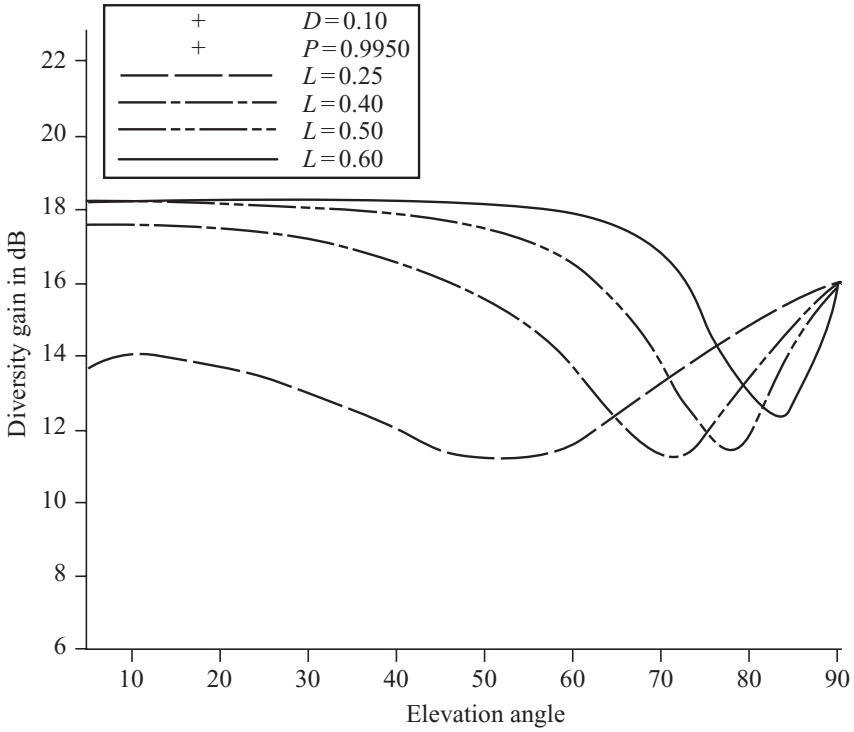


Figure 8.4.42 *The 0.5 per cent diversity gain of the circular three element array, defined relative to one of the elements operating in the presence of the array, against the sloping angle of the monopoles. The diversity gain for the three-element space diversity case of vertical monopoles, is the same as that of two widely spaced monopoles*

parallel so that there is a three element circular array of monopoles mutually spaced by 0.1λ, the 0.5 per cent diversity gain is the equivalent of that of two uncorrelated, equal-power branches.

To compare the diversity gain referenced to a different stand alone reference antenna, say a quarter-wavelength monopole, then the mean branch power must be calculated and the diversity gain recalculated using this reference.

8.4.15 Example: patch antennas on a groundplane

Patch antennas offer practicable and interesting possibilities for diversity in mobile communications. This is because patches can operate in different modes, including degenerate pairs, which can serve as the diversity elements. The different modes can be selected via different feed positions in the patch.

For circular patches, the basic modes have field dependence of $\cos(n(\phi - \phi_0))$. The far-field patterns are given by (James *et al.*, 1981)

$$h_\theta(\theta, \phi) = \mathrm{e}^{\mathrm{j}n\pi/2} \frac{V_0^{(n)}}{2} (k_0 a)\ \cos(n(\phi - \phi_0))\,(J_{n+1} - J_{n-1}) \qquad (8.4.133)$$

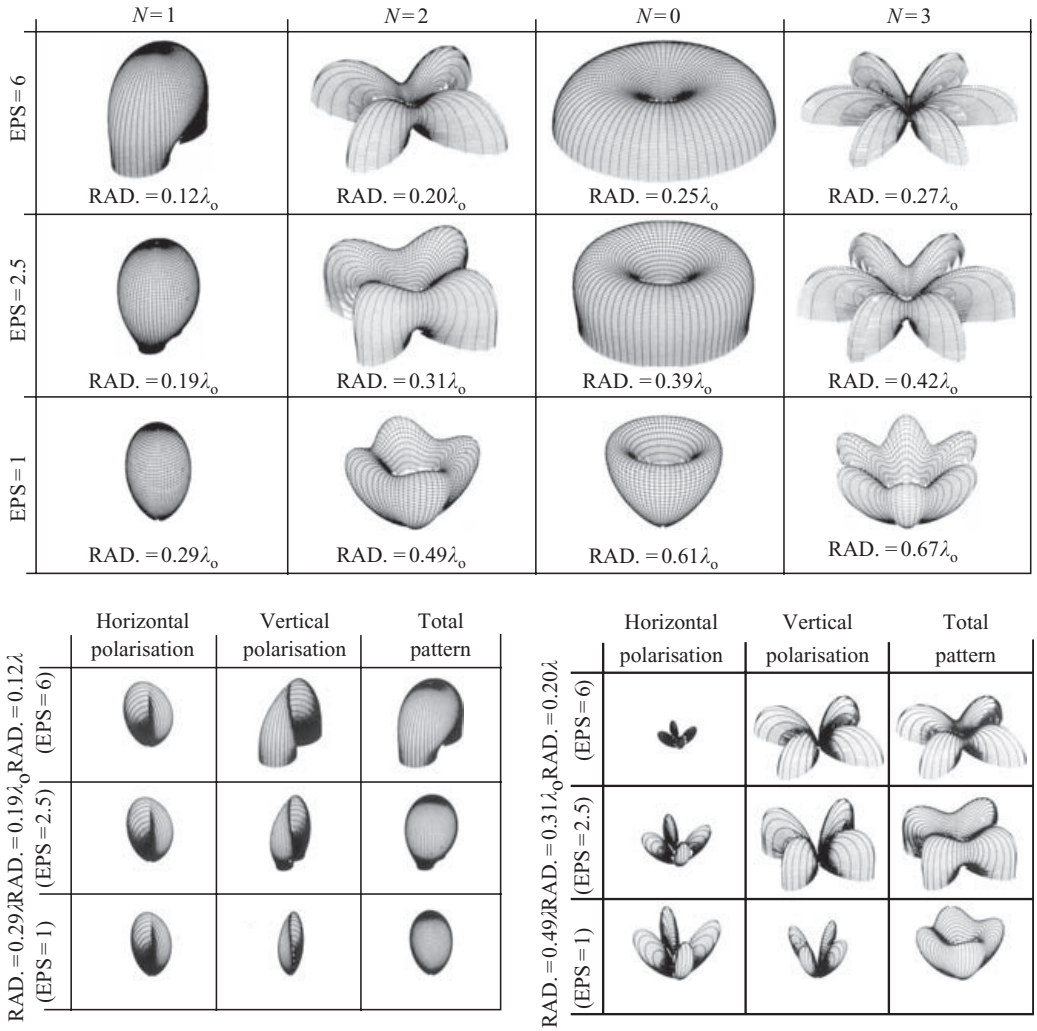


Figure 8.4.43 The power pattern shapes, and polarisation breakdown, on a linear scale, from the basic modes of circular patch antennas. The radius of the patch for the given relative permittivity is included for the patch

and

$$h_\phi(\theta, \phi) = -\mathrm{e}^{\mathrm{j}n\pi/2} \frac{V_0^{(n)}}{2} (k_0 a) \cos\theta \sin\left(n(\phi - \phi_0)\right) (J_{n+1} + J_{n-1}), \qquad (8.4.134)$$

where $J_n = J_n(ka \sin\theta)$ and V_0 is the peak voltage at the edge of the patch which has radius a. On a groundplane, the power patterns with a polarisation breakdown are given in Figure 8.4.43 (Vaughan, 1988). The radius of the resonant circular patch is included for a few practical values of substrate permittivity. The variety of patterns gives some choice for good distributed directivity in a given situation. In a similar way, different modal patterns can be arranged for a square patch, where the modes have field dependence inside the patch of $\sin(kx)\sin(ky)$.

If the scenario is omnidirectional, such as the Clarke or the full 3D spherical scenario cases, then it is clear that the different modes will have orthogonal patterns, seen from an azimuthal integration only,

$$0 = \int_0^{2\pi} \cos\left(n(\phi - \phi_0)\right) \cos\left(m(\phi - \phi_0)\right)\, \mathrm{d}\phi \quad m \neq n, \quad (m, n) \neq (0, 0). \tag{8.4.135}$$

So the orthogonality is achieved through pattern amplitude shape only, with no explicit recourse to the polarisation or phase degrees of freedom for achieving orthogonality.

Degenerate modes, represented by the real and imaginary parts of the $\exp\left(\mathrm{j}n(\theta - \theta_0)\right)$ modes, also give orthogonality. Again, this is seen through the azimuthal variation only,

$$0 = \int_0^{2\pi} \cos\left(n(\phi - \phi_0)\right) \sin\left(n(\phi - \phi_0)\right) \mathrm{d}\phi \quad n \neq 0, \tag{8.4.136}$$

and this is also angle diversity since the patterns are rotated with respect to each other, as seen from $\sin\left(n(\phi - \phi_0)\right) = \cos\left((n(\phi - \phi_0) - (\pi/(2n))\right)$

It is straightforward to realise the degenerate, orthogonal modes, from the same patch, by rotating the off-centre feeds by some multiple of $\pi/(2n)$. In principle, a centre feed to the same patch can excite only the zero mode which provides three orthogonal modes, however, it is practical to build a separate patch on top for this, as shown in Figure 8.4.44 (Vaughan and Andersen, 1987). This concept also allows a transmit port and two diversity receive ports. This latter configuration gives duplexer action through the isolation between the transmit and receive ports.

The patch antenna offers many possibilities for design innovation, both for single elements and for diversity. Circular patch examples have been given here but PIFA and other types lend themselves to imaginative and compact designs, including a mixture of patch and wire antennas.

8.4.16 Example: switched diversity by switched parasitic elements

The fact that only the pattern need be changed to give diversity branches fosters variations in the way a diversity antenna can be realised. For example, the antenna pattern can be changed by near scatterers if the scatterers are included as part of the antenna (Vaughan, 1996, 1999; Scott *et al.*, 1999; Thiel and Smith, 2001). So by placing parasitic elements, which can have minimum scattering antenna properties, near the receiving element so there is appreciable mutual coupling, element arrangements can be sought whose patterns give diversity action. The parasitic elements can be realised by wires and switches in the form of electronic means such as PIN diodes or using mechanical means.

The basic arrangement is shown in the form of a monopole over a groundplane, in Figure 8.4.45 (Vaughan, 1999). When the switch is closed, the mutual coupling action cuts in and the pattern is altered. A problem is that the input impedance changes as the element is switched, but this can be solved with the arrangement shown in (b), where the monopole always 'sees' the same configuration.

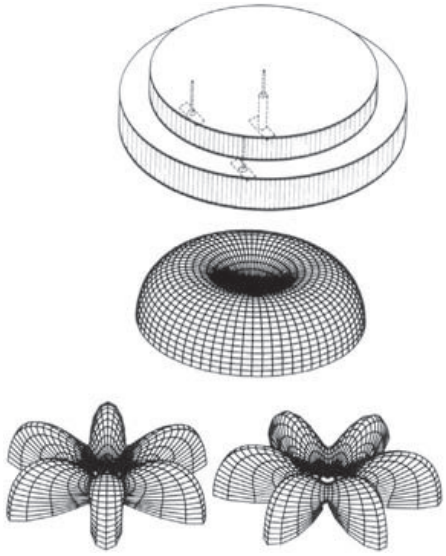


Figure 8.4.44 *Feed positions separated by a multiple of 30° (here 90° is used) and the patterns of the degenerate $n = 3$ modes of the circular patch antenna. The stacked $n = 0$ mode patch gives possibilities of a third branch or a transmit port, which is strongly isolated from the receive ports, giving duplexer action*

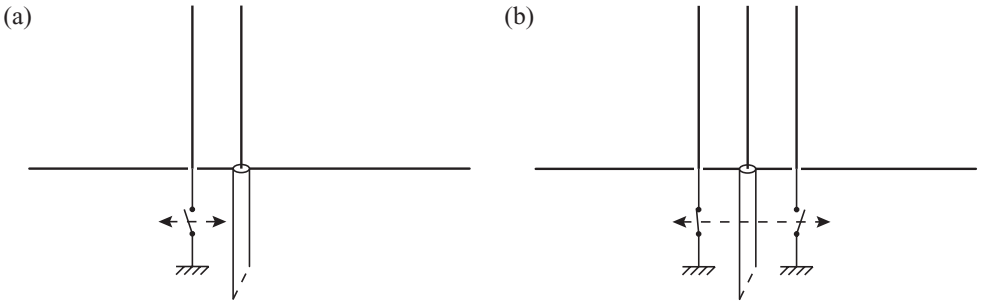


Figure 8.4.45 *Basic concept for switched parasitic diversity. When the parasitic element is open circuited, it is essentially invisible to the monopole, and when the switch is closed the mutual coupling between the elements acts to alter the pattern. The configuration on the right acts to maintain a constant input impedance for the monopole*

An example of a switched diversity antenna whose pattern can be changed or rotated gradually is shown in Figure 8.4.46. From a moment method analysis, the pattern has an HPBW of about 90°, and the impedance is about $21 - j6\ \Omega$. The pattern can be rotated $(360/16)°$ by rotating the switch setting of the parasitic elements.

The envelope correlation function as a function of rotation angle for this antenna pattern is shown in Figure 8.4.47. For adjacent patterns (spacing angle of 22.5°), the open circuit correlation is about 0.95 (envelope correlation about 0.91), and patterns spaced by 45° have an envelope correlation of 0.7 for the complex patterns.

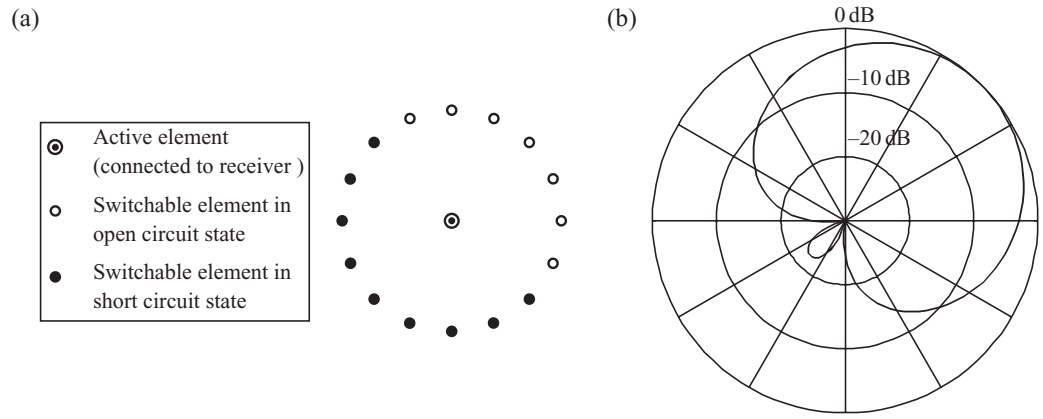


Figure 8.4.46 A 16-element configuration with 0.25λ *radial spacing, which gives the directive pattern (30 dB full scale) in (b)*

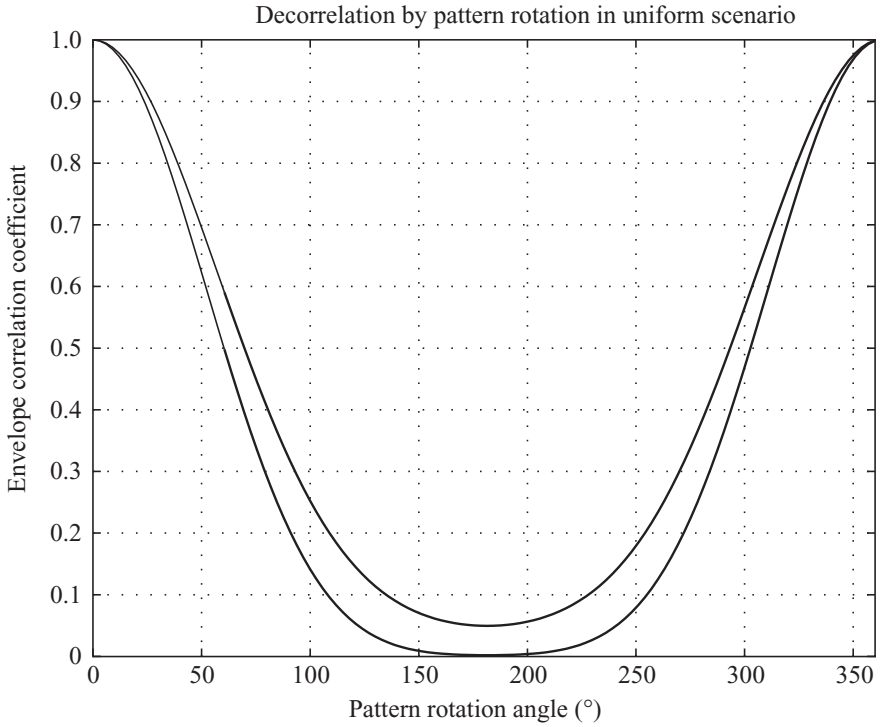


Figure 8.4.47 The envelope correlation coefficient of signals from two patterns of the 16 parasitic element antenna operating in a uniform, uncorrelated scenario, as a function of the rotation angle between the two patterns. The upper curve is for the amplitude pattern only and the lower curve is for the complex pattern. The similarity of the curves shows that the amplitude, rather than the phase is the main decorrelation mechanism for this pattern (cf. Figure 8.4.18)

Other patterns for other diversity branches can be realised with different switch settings. Having many switchable elements gives rise to the possibility of gradient-type algorithms that use relatively gentle transitions between the various patterns. This can decrease potentially error-causing abrupt transitions in the received signal caused by the abrupt pattern changes in the usual switched diversity systems. Also, different terminations can be used instead of simply an open or short circuit, and an example of this technique is given in Scott *et al*. (1999).

The advantage of the switched parasitic concept is that it is a relatively simple system which can give the adaptive antenna performance of many-branch selection- or switched-diversity. Here, the receiver is connected to one element only, so combining, such as equal gain or maximum ratio or optimum, is not possible. The approach can be applied to other antenna types such as slots and switched shorting pins in patch antennas to change the modal structure. The antenna has a single port, so there are no switches in the direct RF path, although there are RF currents flowing in the parasitic elements of course. A disadvantage of using PIN diodes, or other nonlinear elements, in the switches, is the possibility of generating spurious, including out-of-spectrum, signals.

8.4.17 Example: polarisation diversity

Polarisation diversity depends on the polarisations in the scenarios being mutually uncorrelated. Measurements of the polarisation tend to support the hypothesis that the polarisations are indeed uncorrelated, as described in Section 6.7.

8.4.17.1 Mobile terminals

It is hard to make antennas which rely on pure polarisation diversity in a general scenario. If the scenario is omni-like, then dipoles or other linearly polarised antennas at right-angles to each other represent angle diversity, and in this situation it is the amplitude pattern differences that act to decorrelate the received signals.

In an omni scenario, a co-located small wire dipole and small wire loop have the same patterns, except for the polarisation. (A set of these in the three Cartesian directions forms the energy density antenna.) However, if a co-located horizontal loop and vertical dipole are placed closely above a groundplane, such as in a rooftop-mounted antenna for vehicles, then the array factors for the elements and their respective images are different. This means that the amplitude patterns (and the phase) are different, as well as the polarisations (Figure 8.4.48).

The array pattern for the electrical dipole above the groundplane is denoted A_{E}, and A_{M} for the magnetic dipole. The array patterns are

$$A_{\mathrm{E}} = 1 + \mathrm{e}^{\mathrm{j}k_0 2\cos\theta}, \quad A_{\mathrm{M}} = 1 - \mathrm{e}^{\mathrm{j}k_0 2\cos\theta} \tag{8.4.137}$$

and with the element pattern $\sin\theta$, the correlation coefficient for just the amplitude patterns in the presence of the groundplane is

$$\rho_{\mathrm{O}} = \frac{\int A_{\mathrm{E}} \cdot A_{\mathrm{M}}^{*} \sin\theta \, \mathrm{d}\theta}{\sqrt{\int |A_{\mathrm{E}}|^2 \sin^3\theta \, \mathrm{d}\theta \cdot \int |A_{\mathrm{M}}|^2 \sin^3\theta \, \mathrm{d}\theta}} \tag{8.4.138}$$

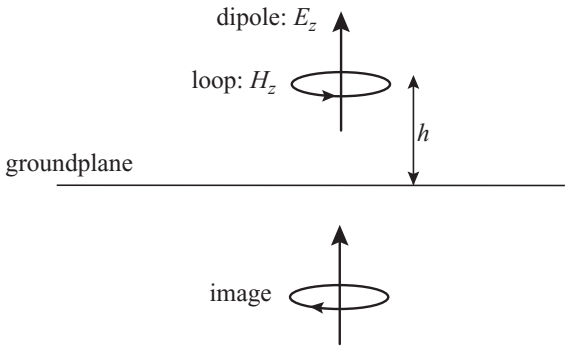


Figure 8.4.48 *A co-located dipole and loop have the same element pattern but different polarisations. Above a large groundplane, the patterns become different because of the array behaviour*

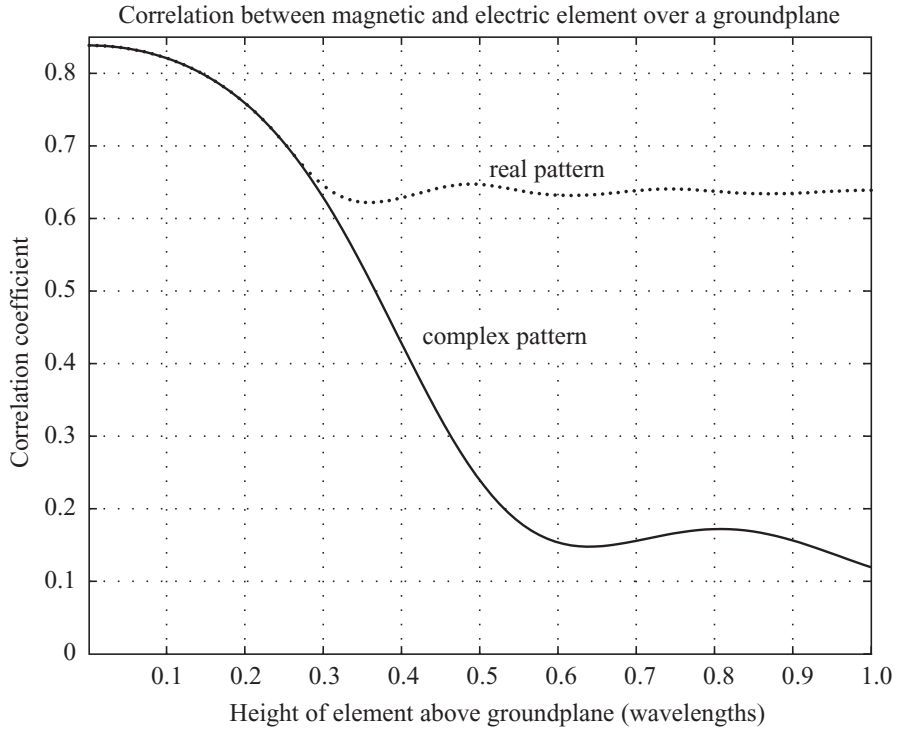


Figure 8.4.49 *The correlation coefficient for the patterns of an electric element and a magnetic dipole above an infinite groundplane in a 3D omni scenario. Although the polarisations are different, the plot shows that the amplitude differences between the patterns is also a decorrelating mechanism*

and this is given in Figure 8.4.49, as well as the correlation for the amplitudes of the patterns, i.e. using $|A_{\mathrm{E}}|$ and $|A_{\mathrm{M}}|$ instead of A_{E} and A_{M}. It is clear that for the range of practical heights for microwave antennas, the array factor is a strong decorrelating

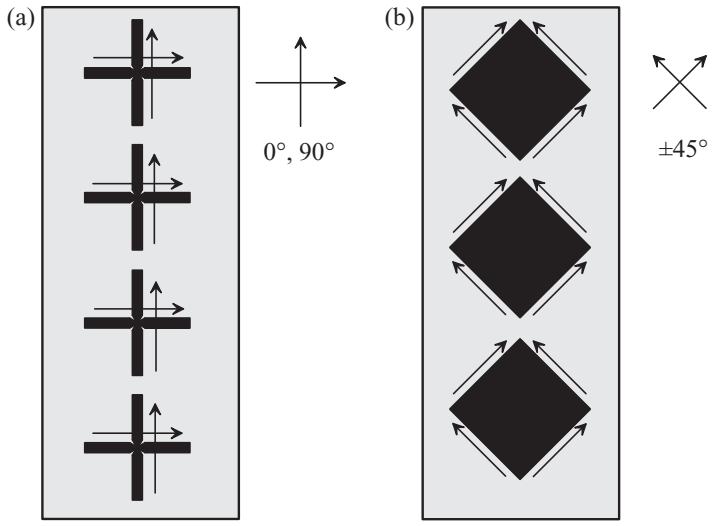


Figure 8.4.50 An outdoor panel arrangement for a base station has polarisation diversity, implemented using orthogonal dipoles (a) or degenerate mode square patches (b). Here the dipoles are aligned to give vertical and horizontal diversity and the patches are aligned to give ±45° *polarisations*

mechanism. The amplitude patterns themselves are sufficient to reduce the correlation to below 0.7 ($\rho_e \approx 0.5$) and when the phase is included, the patterns are essentially uncorrelated for heights above the groundplane of over a half-wavelength.

So it is difficult to identify exactly which degree of freedom dominates the decorrelation mechanism. It practice, it does not matter to a user which particular degree of freedom is the more effective. The fact that both the amplitude patterns are different and the polarisations are different, is an effective way to aim for the uncorrelated branches sought in diversity design. Once the height above the groundplane becomes several wavelengths, the presence of the groundplane becomes less important for incident waves from higher elevation angles.

8.4.17.2 Base stations

At a base station, in particular the outdoor situation, the angular spread from any one mobile terminal is often small. This means that dual-polarised base-station antennas have become popular diversity choices, typically implemented using crossed dipoles or using degenerate mode patches with orthogonal, linear polarisations. Figure 8.4.50 depicts such polarisation diverse base stations. The choice of polarisations makes little difference, since the total energy is collected from a small angular spread with any set of orthogonal polarisations. However, ±45° linear has become the main choice because it is easier to design for matching coverage in an outdoor sector, and generally easier to design for linear rather than circular polarisation. Variable beam downtilt has become a feature of such base stations as well, to allow optimisation of coverage within the cell and interference outside the cell.

8.5 References

ANDERSEN, J. BACH.: 'Admittance of infinite and finite cylindrical metallic antennas', *Radio Science*, June 1968, **3** (6), pp. 607–621.

ANDERSEN, J. BACH.: *Metallic and Dielectric Antennas*, Polytekhnisk Forlag, Copenhagen, 1971.

ANDERSEN, J. BACH.: 'Low- and medium-gain antennas' in *The Handbook of Antenna Design* Ed. A. Rudge et al., Peter Peregrinus, 1982.

ANDREWS, M.A., MITRA, P.A., and deCARVALHO, R.: 'Tripling the capacity of wireless communications using electromagnetic polarisation', *Nature*, January 2001, **409**, pp. 316–318.

BALANIS, C.: *Antenna Theory Analysis and Design*, Harper and Row, 1982.

BELL, D., DESSOUKY, P., ESTABROOK, P., and SUE, M.K.: 'MSAT-X antennas: noise temperature and mobile receiver G/T', *MSAT-X Quarterly*, 1988, No. 16, pp. 13–15.

BONEFAČIĆ, D., BARTOLIĆ, J., and KOCEN, D.: 'Stacked shorted patch antenna with tilted parasitic radiator', *Electronics Letters*, 30th August 2001, **37** (18), pp. 1109–1110.

BOOKER, H.G.: 'Slot aerials and their relation to complementary wire aerials (Babinet's principle)', *JIEE*, 1946, **93**, Part IIIA, pp. 620–626.

CHU, L.J.:'Physical limitations of small antennas', *Journal of Applied Physics*, December 1948, **19**, pp. 1163–1175.

COCKRELL, C.R.: 'The input admittance of the rectangular cavity-backed slot antenna', *IEEE Transactions*, May 1976, **AP-24**, pp. 288–294.

COLLIN, R.E.: *Antennas and Radiowave Propagation*, McGraw-Hill, 1985.

COLLIN, R.E. and ROTHSCHILD, S.: 'Evaluation of antenna Q', *IEEE Transactions*, December 1964, **AP-12**, pp. 23–27.

COLLIN, R.E. and ZUCKER, F.J.: *Antenna Theory*, McGraw-Hill, 1969.

DU TOIT, C.F., CHADWICK, P.E., BUTLAND, R.J., and VAUGHAN, R.G.: 'Adjustable multifilar helical antenna with reduced scanning' *Proceedings of the IEEE International Symposium on Antennas and Propagation*, 1999, **1** (1), pp. 18–21.

ELLIOTT, R.S.: *Antenna Theory and Design*, Prentice-Hall, 1981.

ELLIOTT, R.S.: *An Introduction to Guided Waves and Microwave Circuits*, Prentice-Hall, 1993.

FANO, R.M.: 'Theoretical limitations on the broad-band matching of arbitrary impedances', *Journal of the Franklin Institute*, January 1950, **249**, pp. 57–83; February 1950, pp. 139–154,

FUJIMOTO, K., HENDERSON, A., HIRASAWA, K., and JAMES, J.R.: *Small Antennas*, Research Studies Press/Wiley, 1987.

GILBERT, E.N.: 'Energy reception for mobile radio', *Bell Systems Technical Journal*, 1965, **44**, pp. 1779–1803.

GOUBAU, G.: 'Multi-element monopole antennas', *Proceedings of the Workshop on Electrically Small Antennas*, Fort Monmonth, New Jersey, May 1976, pp. 63–67.

HANSEN, R.C.: *Phased Array Antennas*, Wiley, 1998.

HARRINGTON, R.F.: *Time-Harmonic Electromagnetic Fields*, McGraw-Hill, 1961.
HUANG, J. and DENSMORE, A.C.: 'Microstrip yagi antenna for mobile satellite vehicle application', *IEEE Transactions*, July 1991, **AP-39** (7), pp. 1024–1030.
IEEE Standard Test Procedures for Measuring Antennas, IEEE Std 149-1979, The Institute of Electrical and Electronic Engineers, Inc., New York, in conjunction with Wiley-Interscience, 1979.
JAKES, W.C. (Ed.): *Microwave Mobile Communications*, AT&T press, 1974; also IEEE Press, 1993.
JAMES, J.R. and HALL, P.S. (Eds): *Handbook of Microstrip Antennas* **1** and **2**, Peregrinus, 1989.
JAMES, J.R., HALL, P.S., and WOOD, C.: *Microstrip Antenna Theory and Design*, IEE Peter Peregrinus, 1981.
JENSEN, M.A. and RAHMAT-SAMII, Y.: 'Performance analysis of antennas for handheld transceivers using FDTD', *Proceedings of the IEEE*, January 1995, **83**, pp. 7–17.
JONES, D.S.: *Methods in Electromagnetic Propagation*, Oxford University Press, 1979.
JORDAN, E.C. (Ed.): *Reference Data for Engineers: Radio, Electronics, Computer and Communication*, 7th Edition, Sams, 1985.
KAC, M. and SIEGERT, A.J.F.: 'On the theory of noise receivers with square law detectors', *Journal of Applied Physics*, 1947, **18**, p. 396.
KAHN, W.K. and KURSS, H.: 'Minimum scattering antennas', *IEEE Transactions*, September 1965, **AP-13**, pp. 671–675.
KILGUS, C.C.: 'Resonant quadrifilar helix', *IEEE Transactions*, July 1968, **AP-16**, pp. 499–500.
KLOPFENSTEIN, R.W.: 'A transmission line taper of improved design', *Proceedings of the IRE*, 1956, **44**, pp. 31–35.
KOZONO, S., AOYAMA, S., and TANAKA, T.: 'Prediction of propagation delay spread in mobile radio communications', *Antennas and Propagation Society Symposium Digest*, London, Ontario, June 1991, **3**, paper 88.7, pp. 1548–1551.
KRAFT, U.R.: 'Gain and G/T of multielement receive antennas with active beam forming networks', *IEEE Transactions*, December 2000, **AP-45** (12), pp. 1818–1029.
KRAUS, J.D.: *Antennas*, McGraw-Hill, 1950; second edition, 1988.
LEE, W.C.Y.: *Mobile Communications Engineering*, McGraw-Hill, 1982.
LONG, S.A.: 'Experimental study of the impedance of cavity-backed slot antennas', *IEEE Transactions*, January 1975, **AP-23**, pp. 1–7.
MATTHAEI, G.L., YOUNG, L., and JONES, E.M.T.: *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, McGraw-Hill, 1964.
McLEAN, J.S.: 'A re-examination of the fundamental limits on the radiation limits on the radiation Q of electrically small antennas', *IEEE Transactions*, May 1996, **AP-44** (5), pp. 672–676.
MONTGOMERY, C.G., DICKE, R.H., and PURCELL, E.M.: *Principles of Microwave Circuits*, McGraw-Hill, 1948.

NØRKLIT, O., EGGERS, P.C.F., and ANDERSEN, J. BACH.: 'Jitter diversity in multipath environments', *Proceedings of the IEEE Vehicular Technology Society Conference*, Chicago, 1995, **2** (5), pp. 853–857.

NØRKLIT, O., TEAL, P.D., and VAUGHAN, R.G.: 'Measurements and evaluation of multi-antenna handsets in indoor mobile communication', *IEEE Transactions*, March 2001, **AP-49** (3).

PEDERSEN, G.F. and ANDERSEN, J.B.: 'Integrated antennas for handheld telephones with low absorption', *Proceedings 44th IEEE Vehicular Technology Conference*, Stockholm, June 1994, pp. 1537–1541.

PEDERSEN, G.F. and ANDERSEN, J.B.: 'Handset antennas for mobile communications – integration, diversity, and performance', *Review of Radio Science 1996–1999*, Oxford University Press, August 1999, pp. 119–138.

PEDERSEN, G.F., NIELSEN, J.Ø., OLESEN, K., and KOVACS, I.Z.: 'Antenna diversity on a UMTS handheld phone', *Proceedings of the 10th International Symposium on Personal, Indoor and Mobile Communications*, September 1999, **1**, pp. 152–156.

PEDERSEN, G.F. and SKJAERRIS, S.: 'Influence on antenna diversity for a handheld phone by the presence of a person', *Proceedings of the 47th IEEE Vehicular Technology Conference*, 1997, **3**, pp. 1768–1772.

POZAR, D.M.: *Microwave Engineering*, Wiley, 1998.

RAPPAPORT, T.S. and HAWBAKER, D.A.: 'Wide-band microwave propagation parameters using circular and linear polarized antennas for indoor wireless channels', *IEEE Transactions*, February 1992, **Com-40** (2), pp. 240–245.

RUMSEY, V.H.: 'Horn antennas with uniform power pattern around their axes', *IEEE Transactions*, 1966, **AP-14**, p. 156.

RUSTAKO, A.J., YEH, Y.S., and MURRAY, R.R.: 'Performance of feedback and switch space diversity 900 MHz FM mobile radio system with Rayleigh fading', *IEEE Transactions*, November 1973, **Comm-21**, pp. 1257–1268.

SCOTT, N.L., LEONARD-TAYLOR, M.O., and VAUGHAN, R.G.: 'Diversity gain from a single-port adaptive antenna using switched parasitic elements illustrated with a wire and monopole prototype', *IEEE Transactions*, June 1999, **AP-47** (6), pp. 1066–1070.

SLATER, J.C.: *Microwave Transmission*, McGraw-Hill, 1942.

SMITH, G.S.: 'Radiation efficiency of electrically small multiturn loop antennas', *IEEE Transactions on Antennas and Propagation*, September 1972, **AP-20** (5), pp. 656–657.

STEIN, S.: Part III of SCHWARTZ, M., BENNETT, W.R., and STEIN, S.: *Communication Systems and Techniques*, McGraw-Hill, 1966.

STUTZMAN, W.L. and THIELE, G.A.: *Antenna Theory and Design*, Wiley, 1981.

TAGA, T.: 'Analysis for mean effective gain of mobile antennas in land mobile radio environments', *IEEE Trans. Vehicular Technology*, May 1990, **39**, pp. 117–131.

TAYLOR, C.D., ARONSON, E.A., and HARRISON Jr., C.W.: 'Theory of coupled monopoles', *IEEE Transactions*, 1970, **AP-13** (3), pp. 360–366.

TEAL, P.D., LEONARD-TAYLOR, M.O., and VAUGHAN, R.G.: 'Quad patch PCS antenna', Industrial Research Limited Report, September 1997.

THIEL, D.V. and SMITH, S.: *Switched Parasitic Antennas for Cellular Communications*, Artech House, 2001.

TOFTGÅRD, J., HORNSLETH, S.N., and ANDERSEN, J.B.: 'Effects on portable antennas of the presence of a person', *IEEE Trans. Antennas and Propagation*, June 1993, **41** (6), pp. 739–746.

TURIN, G.L.: 'The characteristics and function of a Hermitian quadratic form in a complex normal variable', *Biometrika*, 1960, **47**, pp. 199–201.

VAUGHAN, R.G.: 'Two port higher mode circular microstrip antennas', *IEEE Transactions*, March 1988, **AP-36** (3), pp. 1365–1374.

VAUGHAN, R.G.: 'Switched parasitic elements for antenna diversity', New Zealand Patent Application 286131, March 1996.

VAUGHAN, R.G.: 'Estimation and reduction of power absorbed in the head from handheld terminals', *Proceedings of the 7th International Symposium on Personal, Indoor and Mobile Radio Communications* (PIMRC 96), Taipei, October 1996, **3**, pp. 477–481.

VAUGHAN, R.G.: 'Switched parasitic elements for antenna diversity', *IEEE Transactions*, February 1999, **AP-47** (2), pp. 399–405.

VAUGHAN, R.G. and ANDERSEN, J. BACH.: 'Polarisation properties of the axial mode helix antenna', *IEEE Transactions*, January 1985, **AP-33** (1), pp. 10–21.

VAUGHAN, R.G. and ANDERSEN, J. BACH.: 'Antenna diversity for mobile communications', *Proceedings of the Nordic Seminar on Digital Land Mobile Radiocommunications*, Espoo, February 1985, pp. 87–96.

VAUGHAN, R.G. and ANDERSEN, J. BACH.: 'Antenna diversity in mobile communications', *IEEE Transactions*, November 1987, **VT-36** (4), pp. 149–172.

VAUGHAN, R.G. and ANDERSEN, J. BACH.: 'A multiport patch antenna for mobile communications', *Proceedings of the 14th European Microwave Conference*, Liege, 1984, pp. 607–692; also reprinted in *Advanced Antenna Technology*, **2**, Ed. P. Clarricoats, Microwave Exhibitions and Publishers, UK, 1987.

VAUGHAN, R.G., LANGHORN, M. and ANDERSEN, J. BACH.: 'Circular array of outward sloping monopoles' *IEEE Transactions*, October 1988, **AP-36** (10), pp. 309–321.

VAUGHAN, R.G. and SCOTT, N.L.: 'Closely spaced terminated monopoles', *Radio Science*, Nov-Dec 1993, **28** (6), pp. 1259–1266.

VAUGHAN, R.G. and SCOTT, N.L.: 'Terminated in-line monopoles for vehicular diversity', *Proceedings of the 1992 Internationial Union of Radio Science (URSI) Triennial Symposium on Electromagnetic Theory*, Sydney, August 1992, pp. 111–113.

VAUGHAN, R.G. and SCOTT, N.L.: 'Composite antenna for handheld or portable communications', New Zealand Patent Application 248283, July 1993a; (US Patent 5771022, June 1998).

VAUGHAN, R.G. and SCOTT, N.L.: 'Antennas for FPLMTS', *Proceedings of the Fourth International Symposium on Personal, Indoor, and Mobile Radio Communications* (PIMRC 93), Yokohama, September 1993, pp. A4.2.1–A4.2.5.

VAUGHAN, R.G. and SCOTT, N.L.: 'Evaluation of antenna configuration for reduced power absorption in the head', *IEEE Transactions*, September 1999, **VT-48** (5), pp. 1371–1380.

VAUGHAN, R.G., SCOTT, N.L., and JENNESS, C.A.: 'Steerable beam helix structure', New Zealand provisional patent 242298, August 1994; (US Patent 6512707, May 1997).

VAUGHAN, R.G., SCOTT, N.L., and JENNESS, C.A.: 'Bungy adjustable helix structure', New Zealand provisional patent 264314, August 1994.

WASYLKIWSKYJ, W. and KAHN, W.K.: 'Theory of mutual coupling among minimum scattering antennas', *IEEE Transactions*, March 1970, **AP-18** (2), pp. 204–216.

WATERHOUSE, R.B.: 'Design and performance of small printed antennas', *IEEE Transactions*, November 1998, **AP-46** (11), pp. 1629–1633.

WATERHOUSE, R.B.: 'Design of probe-fed stacked patches', *IEEE Trans. Ant. Prop.*, December 1999, **47** (12), pp. 1780–1784.

WEINER, M.W., CRUZE, S.P., LI, CHO-CHOU, and WILSON, W.J.: *Monopole Elements on Circular Ground Planes*, Artech House, 1987.

WHEELER, H.A.: 'Fundamental limitations of small antennas', *Proceedings IRE*, December 1947, **35**, pp. 1479–1484.

WOLFF, E.A.: *Antenna Analysis*, Wiley, 1966.

Chapter 9

Array antennas in a multipath environment

9.1 Introduction

In the previous chapters receiver diversity related to antenna arrays has been discussed to some extent, but the additional flexibility derived from having more than one antenna element leads to many more uses, some of which are rather recent and which only have attracted theoretical attention. The interdependence between the antennas and the propagation environment is even more clear in this new connection, since the signal correlation between elements is very important for the potential applications, and this correlation depends on the angular spread as seen from each array antenna, as has been shown in Section 6.4. The situation may be described as given in Figure 9.1.1.

In general there will be M transmitting antennas in a transmit array and N receiving antennas in a receive array. Although there will be reciprocity between any two antennas, it will be necessary to distinguish between the transmit and receive situation, since the knowledge of the channel parameters may be different in the two situations. Also in the most general case the angular spread of the same environment will be different seen from the two sides.

Another important distinction will be between one-port and multi-port arrays. In the first case the different elements may have different complex weights, but combined into one port, whereas in the second case each element may be connected to a network with several ports. This is illustrated in Figure 9.1.2.

- ***A one-port array***

  The situation is symbolically described by Figure 9.1.2(a), where the symbol sequence $s_1, s_2, \ldots,$ is transmitted sequentially, and the antenna weights $w_1, \ldots, w_n$ are used to optimise some property, such as a specific shape of the radiation pattern, or an optimum combination for uncorrelated elements, but the information transmitted or received is the same for all elements, justifying the notation of a one-port array. It is the traditional use of an array, as in space diversity applications.

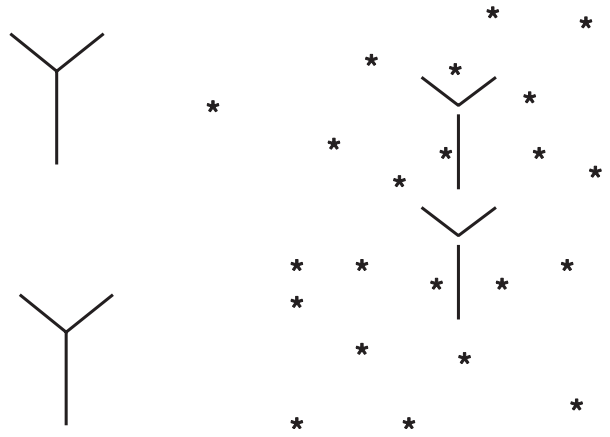


Figure 9.1.1 *Two antenna arrays in a scattering environment. The angular spread seen from the two arrays may be different. The antennas to the left see a small angular spread and those on the right see a large spread even though the modelled scatterers (marked *) are common. The situation is characteristic of a macrocellular environment with the base station to the left and the mobile to the right*

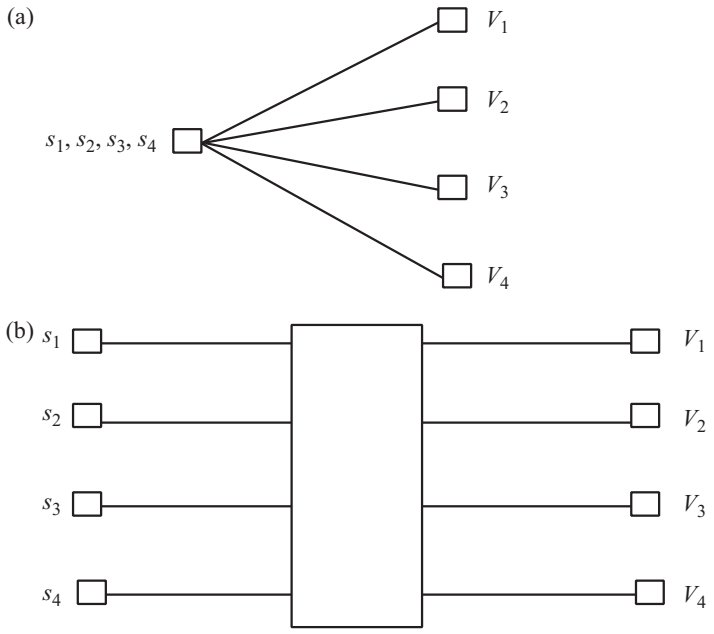


Figure 9.1.2 *(a) The array is a one-port antenna with the same information transmitted over all elements; (b) the array is a multi-port antenna, where the information signals are spread over the array elements*

- ***A multi-port array***

  The situation is depicted in Figure 9.1.2(b). Several independent information symbols are treated at the same time, combined or delayed, so the information symbols sent to the individual ports are different. The combination could be a

Butler matrix, where each input port would correspond to a different direction in space, or a base band coding, where each information signal coding is spread over the antennas leading to essentially a number of quasi-independent channels.

A final distinction is between the use in a cellular network, where the interfering signals come from different users, and the situation of multi-element arrays at each end of the link. Here the user elements effectively act as self-interferers. First the cellular network case will be treated.

9.2 Adaptive antennas in cellular networks

For outdoor environments adaptive receive antennas will usually be applied at a base station for a macrocell. We assume there are K interferers with one antenna each and an array at the base with M elements. The adaptive antenna will in principle be able to enhance the performance of the network by 'focusing' on the wanted user, thus diminishing the path loss, and reducing the influence of the unwanted users, i.e. the interferers. In order to achieve this enhancement, which ultimately leads to higher user capacity, the antenna system must be able to distinguish between user and interferer, and know something about the channel transfer function to the antenna elements from all users. This is not always possible, even in principle, since the channel may be non-reciprocal in the sense that the downlink and the uplink transmission may take place at different carrier frequencies (i.e. the frequencies are separated by more than the coherence bandwidth) as in GSM. The uplink is the transmission from mobile to base and the downlink is the corresponding transmission from base to mobile. Most practical situations currently relate to one antenna element at the mobile, and several antenna elements at the base, although it will be shown later that considerable improvement occurs when there are multiple antenna elements at each end. The uplink case, where the base is receiving, is the most useful from an array point of view, since the channels corresponding to each user at each base station element may be estimated on reception. This assumes that there is a 'flag' identifying the signal of the wanted user, which could be CDMA code in a CDMA system.

9.2.1 Interference rejection

As shown earlier the spatial correlation depends on the angular spread of the environment seen from the array. The angular spread can vary from less than 1° to 10–20° for macrocells, so the correlation coefficient will in practice vary between zero ('uncorrelated') and one ('correlated'). In the following, two extremes will be treated, the typical case being somewhere in between. First it is assumed that there are completely uncorrelated signals. To be more precise the signals are coherent, they are replicas of the same original signal, but they are coming simultaneously from many different directions, so the spatial correlation between elements along the array is zero. In this case the notion of a radiation pattern loses its meaning, a null in a pattern will be filled by signals coming from other directions, and a description simply assuming Rayleigh fading signals at each antenna from each user will be applied. Only the narrowband

case will be treated. The antenna is assumed to have M elements, and the geometry and polarisation are arbitrary as long as the condition of zero correlation is satisfied.

9.2.1.1 Spatially uncorrelated signals

The signal from user number 1 is thus given by $s_1 \boldsymbol{H}_1$, where s_1 is the information-bearing part and $\boldsymbol{H}_1$ is the part belonging to the propagation channel

$$\boldsymbol{H}_1 = (V_1, V_2, \ldots, V_M)^{\mathrm{T}} \tag{9.2.1}$$

where V_1 is the complex transmission factor at element i, with a Gaussian distribution of the real and the imaginary part, leading to a Rayleigh distribution in magnitude. Note that for any given position the $\boldsymbol{H}$ vector is assumed to be constant; it is only when moving (or when the environment changes) that the statistical properties enter. Similar signals $s_k \boldsymbol{H}_k$ exist from the other users. A weight factor $\boldsymbol{w}$ is applied to the array (see also Section 9.4)

$$\boldsymbol{w} = (w_1, w_2, \ldots, w_M)^{\mathrm{T}} \tag{9.2.2}$$

the elements being complex, the norm unity, and superscript 'T' meaning transpose. The power received from user 1 is then given by

$$P_1 = \boldsymbol{w}^H \boldsymbol{H}_1 \boldsymbol{H}_1^H \boldsymbol{w} = \boldsymbol{w}^H \boldsymbol{R}_1 \boldsymbol{w} \tag{9.2.3}$$

where ^{H} means conjugate transpose. The signal part s_1 is supposed to have unit magnitude in a CDMA system and be orthogonal (uncorrelated) to signals from the other users, such that the propagation vector $\boldsymbol{H_1}$ may be reliably estimated. $\boldsymbol{R_1}$ is a Hermitian matrix with one non-zero, real eigenvalue, which is the maximum value (over all weights) of P_1, and the corresponding eigenvector is the optimum weight factor for this eigenvalue problem. $\boldsymbol{R_1}$ is often called the correlation matrix since it may be determined by correlating the signals with the assumed independence properties.

Similar expressions may be derived from the other users, and assuming a noise power described by a diagonal matrix $\boldsymbol{N}_0$ with equal elements the following expression is found for the ratio of the signal power over the sum of all interferer powers and noise,

$$SINR = \frac{\boldsymbol{w}^H \boldsymbol{R_1} \boldsymbol{w}}{\boldsymbol{w}^H (\boldsymbol{N_0} + \boldsymbol{R_2} + \boldsymbol{R_3} + \cdots) \boldsymbol{w}} = \frac{\boldsymbol{w}^H \boldsymbol{R_1} \boldsymbol{w}}{\boldsymbol{w}^H \boldsymbol{R} \boldsymbol{w}}, \tag{9.2.4}$$

where number 1 is assumed to be the wanted user and the others unwanted interferers. Note that the noise is the antenna noise modified by the weights, although strictly speaking this depends on implemention details. Maximising this *SINR* ratio is called optimum combining, and it has been discussed by Winters *et al.* (1994). For zero interference it reduces to maximal ratio combining. In contrast to equation (9.2.3) this is a generalised eigenvalue problem, and the solution is rapidly computed using known algorithms. It will be discussed in relation to some numerical calculations showing the distributions of SINR over many realisations of the random channel. It is also well known that the optimum weight vector is given by

$$\boldsymbol{w_0} = \boldsymbol{R}^{-1} \boldsymbol{H}_1^* \tag{9.2.5}$$

and the maximum value is given by

$$SINR_{\max} = \boldsymbol{H}_1^H \boldsymbol{R}^{-1} \boldsymbol{H}_1 \tag{9.2.6}$$

Note that a standard technique in adaptive filtering, the least mean squares (LMS) method (Haykin, 1991), uses the total correlation matrix $\boldsymbol{R}$, i.e.

$$\boldsymbol{R} = \boldsymbol{N}_0 + \boldsymbol{R}_1 + \boldsymbol{R}_2 + \cdots \tag{9.2.7}$$

which includes the wanted signal. This is formulated as minimising the mean square difference between the signal and the signal transmitted by user 1, corresponding to maximising the ratio of signal power over signal + noise + interference. The solution is the same (see Section 9.4).

In Figure 9.2.1 is shown a case of one interferer as a function of the average interference-to-signal ratio with antenna array size (number of uncorrelated elements) as a parameter. The averaging is over many scenarios of spatially uncorrelated transmission factors. Note that it is assumed that the weights are found for each snapshot (or frame) and that the channel is assumed to be constant during this time.

For small values of ISR (interference-to-signal ratio) the interference is negligible and the normal 3 dB combining gain for two elements is noted, leading to an *SNR*

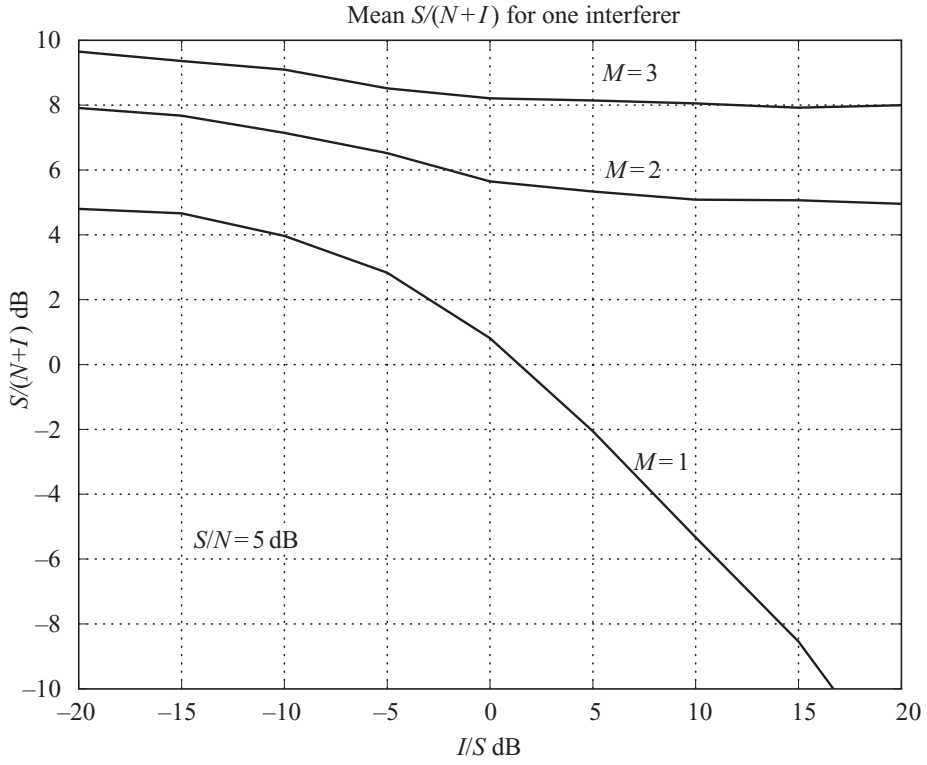


Figure 9.2.1 Average SINR with spatially uncorrelated signals and one interferer. M is the number of array elements, and the signal to noise ratio (SNR) at one element is 5 dB

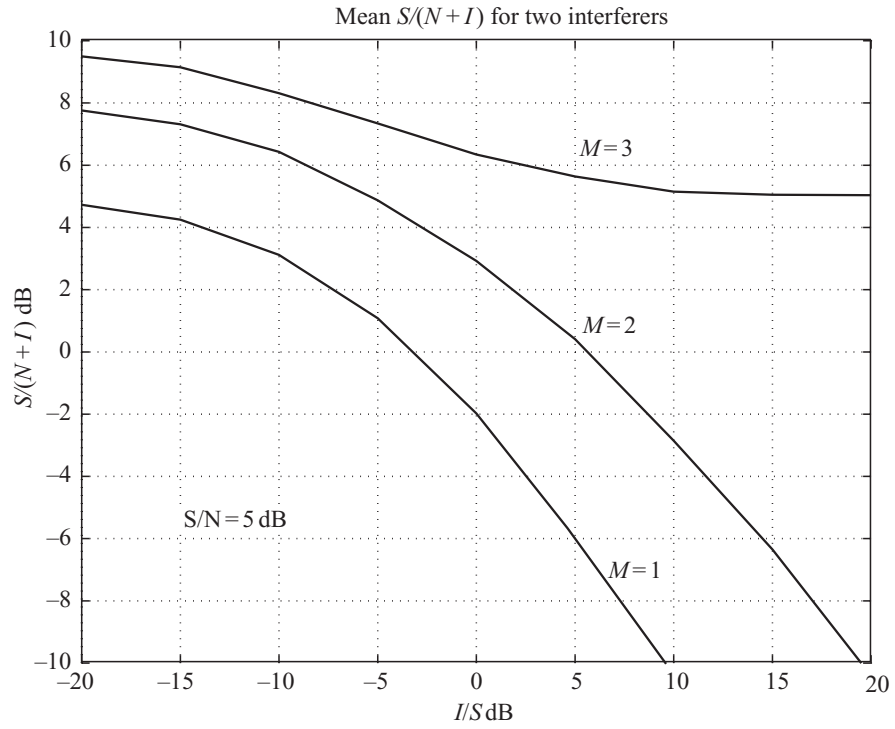


Figure 9.2.2 Average SINR with spatially uncorrelated signals and two interferers present. M is the number of array elements, and the signal to noise ratio at one element is 5 dB

of 8 dB with an *SNR* of 5 dB for a single element. For $M = 2$ the combiner is able to cancel the interferer completely for large values of interference, but this means that there are no degrees of freedom left for the diversity action, and thus the value of *SINR* returns to 5 dB. Note that the cancellation can be interpreted as putting the interferer in a deep fade, not as a null in a radiation pattern. Adding another element ($M = 3$) provides this extra freedom giving second order diversity and one interference cancellation. As pointed out by Winters *et al.* (1994), M antennas may enable cancellation of K interferers and $M-K$ order diversity. This is further illustrated in Figure 9.2.1 for two interferers assumed to be independent with equal average power. Although some gain is available for $M = 2$, only $M = 3$ will cancel out both interferers.

The cumulative probability distribution for the same case is shown in Figure 9.2.3. The *ISR* is 10 dB, and the diversity effect (seen as the slope of the probability curves) is still there, but it does not increase from $M = 2$ to $M = 3$ because all the degrees of freedom are used to cancel the interferers.

9.2.1.2 Spatially correlated signals

Here the other extreme is chosen with spatially completely correlated signals from each source, i.e. the angular spread is zero (or just much smaller than the effective beamwidth of the array), and the individual users appear as point sources as seen

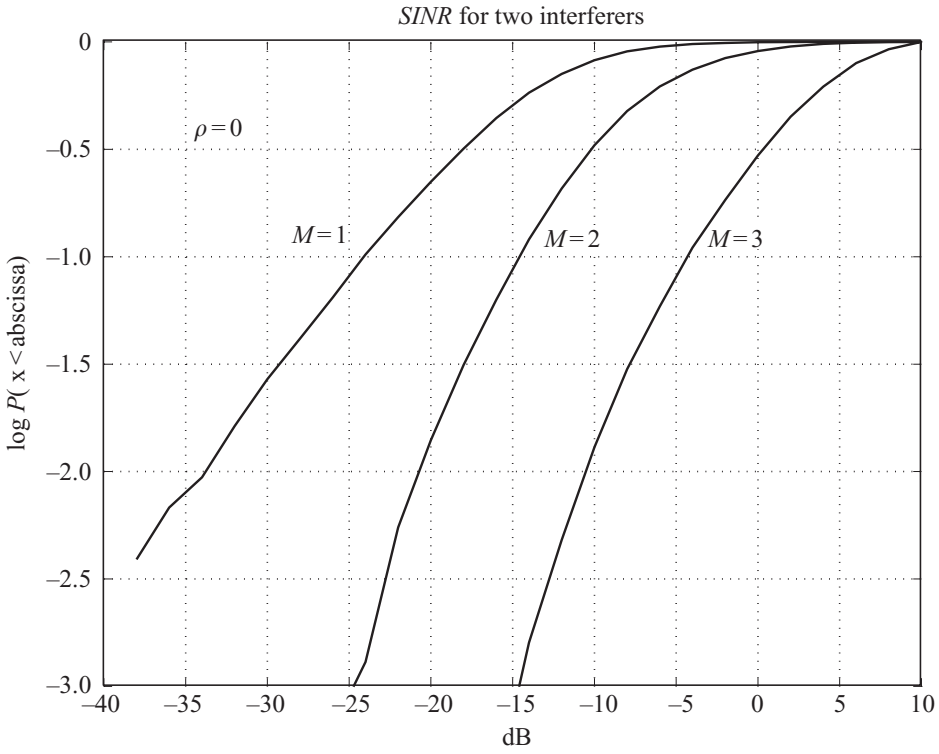


Figure 9.2.3 Cumulative distribution of SINR for two interferers with spatially uncorrelated signals. Mean $ISR = 10\,dB$

from the array. It is then relevant to use the angular domain, so instead of the random vectors in equation (9.2.1) the incident wave from each point source creates a linear phase shift across the array. When the angular spread at the mobile is large, which is most often the case, the magnitudes of the signals at all of the elements will have a common Rayleigh fading.

The signal from user number 1 at the output of an array with constant spacing is now

$$\boldsymbol{H}_1 = V_1(\exp(-\mathrm{j}u, \exp(-\mathrm{j}2u), \ldots, \exp(-\mathrm{j}Mu))^{\mathrm{T}} \tag{9.2.8}$$

where here V_1 is a random complex factor and

$$u = k_0 d \sin(\theta_1), \tag{9.2.9}$$

with k_0 the free-space wavenumber, d the distance between elements and θ_1 the direction to user number 1 measured from the normal to the array. The interferers will have a similar signal vector but with a different direction (hopefully!). The methods described above for the uncorrelated case may be used again, the only difference being the distribution of angles. In the numerical examples a linear array with half a wavelength between elements has been used and θ_1 is set equal to zero. The angles for the interferers are uniformly distributed between $-60°$ and $+60°$ around broadside.

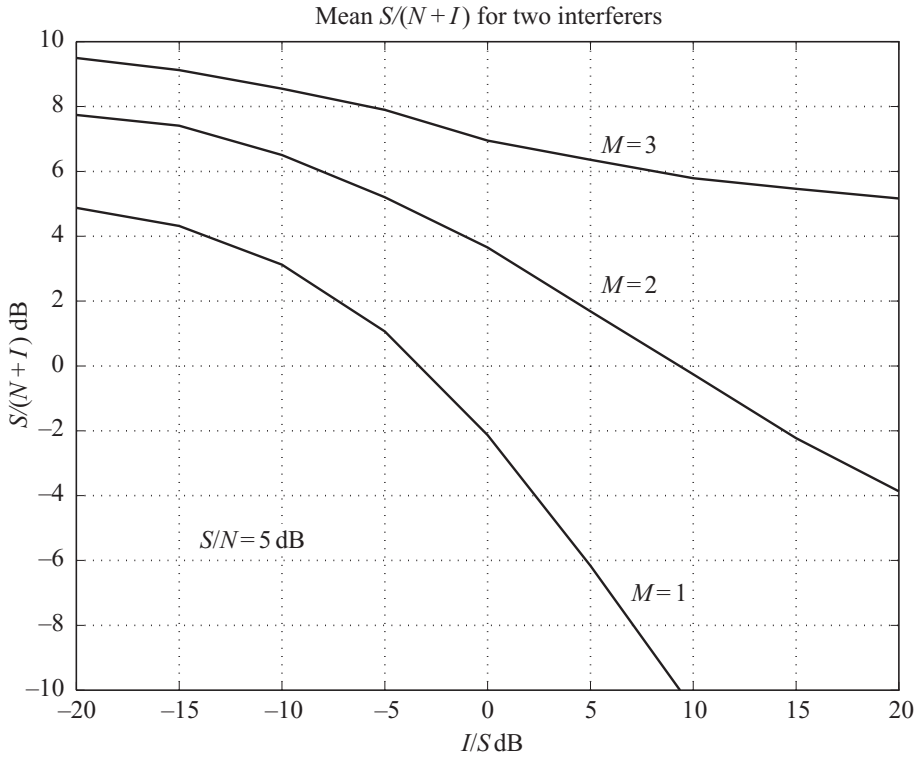


Figure 9.2.4 Average SINR with spatially correlated signals and two interferers. M is the number of array elements, and the signal to noise power at one element is 5 dB. Narrow band Rayleigh fading is assumed for all users. The interferers are distributed uniformly over a 120° sector

For one interferer the result is somewhat surprising; the mean value for the *SINR* is the same as for the completely uncorrelated case as shown in Figure 9.2.1! The discussion above could be repeated, except that antenna gain would be inferred instead of diversity gain; the result is the same. There will be a difference however if the cumulative probability distributions are considered, say at the 1 per cent level. At the left side of Figure 9.2.1 there is diversity gain for the uncorrelated case, while in the correlated case the signals always will exhibit Rayleigh fading. For two interferers the result shown in Figure 9.2.4 can be compared with Figure 9.2.2. Again the results are rather similar, although the correlated case works better when the number of interferers equals the number of antennas, $M = 2$.

The distribution for this latter case shows (Figure 9.2.5) that only antenna gain is active, since the curves are just shifted without changing the shape, in contrast to the case of Figure 9.2.3.

9.2.1.3 Direction of arrival, DOA

The calculation of the optimum weights was performed above by assuming that the correlation matrix $\boldsymbol{R}$, and the array factor $\boldsymbol{H_1}$ are known. This is a problem in itself in estimation theory and a topic for various adaptive filter methods, outside the scope of this book. However, certain differences between the correlated and uncorrelated

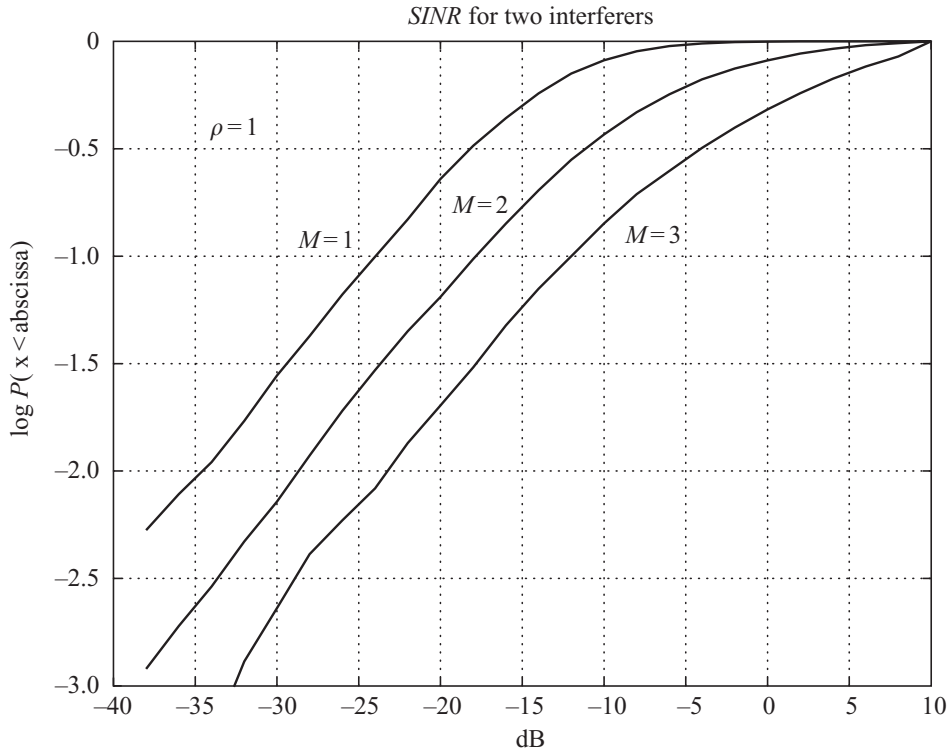


Figure 9.2.5 Cumulative distribution function of SINR for two interferers with spatially correlated signals. Mean $(I/S) = 10\,dB$

cases may be noted. In order to estimate $\boldsymbol{H_1}$ from equation (9.2.1) within a complex constant, all the antenna signals from a particular user need to be estimated. This is a demanding task, but necessary for using equation (9.2.4) to find the maximum possible useful power for the uncorrelated signals. In the correlated case, equation (9.2.8), only one parameter needs to be estimated, namely the direction parameter u of the angle θ. This is much simpler, since it is only one parameter per user. In most cases with a non-zero angular spread this corresponds to some decorrelation between the elements, but the average value of θ should be sufficiently good for enhancement of the wanted user and rejection of the largest interferers.

In practice it has been found useful to use simple techniques like angular spectrograms, or scanning all the angles of interest with steps of a few degrees. Such information derived from the uplink may then be used for downlink transmission, when the channel usually is unknown, but the average direction is approximately unchanged.

9.2.1.4 User discrimination: uplink case

In the uplink situation it is thus possible at the base station to discriminate different users by performing optimum combining of the individual users. This is done in an uncorrelated environment by separating the users by a distance of the order of a

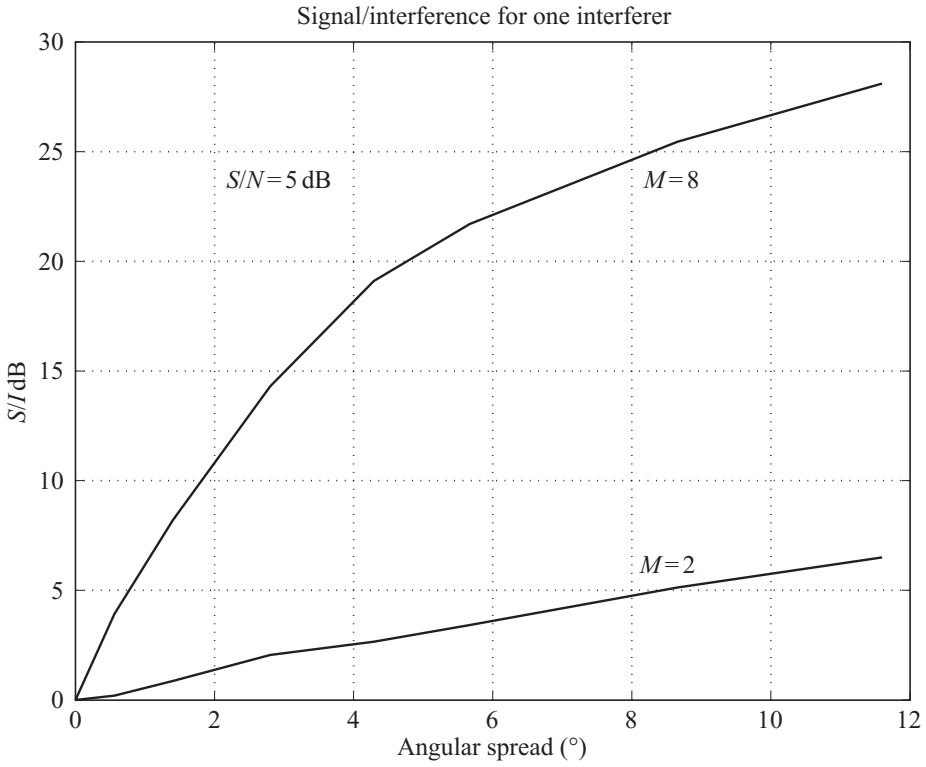


Figure 9.2.6 The signal to interference ratio between two closely spaced users of equal average power as a function of the angular spread seen from the base array. Optimum combining is used and the average SNR is 5 dB. M is the number of elements at the base station

wavelength, or by separating by a distance giving sufficient separation in angle in the correlated environment. In most practical cases we have a situation in between the two extreme cases with some average angle surrounded by an angular spread. An interesting question for finding the maximum density of users is, how closely spaced can two users be in the general case, half a metre (shoulder to shoulder) or 50 m? An example is given in Figure 9.2.6, which shows the interference suppression in a simulated case, where two users are separated by a distance d and surrounded by a collection of scatterers yielding a large angular spread seen from the mobile and a variable spread (x-axis in Figure 9.2.6) as seen from the base. The answer to the question above is that the distance d is not important. At microwave frequencies, as long as d is larger than a few centimetres, the scatterers will be illumi nated in an uncorrelated manner seen from the mobile, and it is the angular spread seen from the base relative to the size of the array which is important. The distance between elements at the base is half a wavelength, and clearly two elements have only limited suppression capability due to the high spatial correlations at the base. For $M = 8$ elements the suppression is quite high even for a few degrees of angular spread. When the angular spread becomes very large the curves saturate corresponding to the completely uncorrelated case (Figure 9.2.1).

9.2.1.5 Downlink case with unknown channel

For the downlink when the channel is unknown, the situation is somewhat different. In an FDD system, if the frequency separation in uplink and downlink is larger than the coherence bandwidth, the short term fading will be unknown. Thus complex signals (or rather complex weights) are out of the question, but it can be assumed with some confidence that the angular power distribution is approximately the same for up- and downlink. The best that can be achieved is then to optimise the SINR on a power basis, i.e. creating a far-field pattern from the array that does that. This may easily be put into the same framework as before where instead of equation (9.2.8) we get

$$\boldsymbol{H}_{\boldsymbol{w},\boldsymbol{k}} = \sqrt{P_{\boldsymbol{w},\boldsymbol{k}}}(\exp(-\mathrm{j}u_k), \exp(-\mathrm{j}2ui_k), \ldots, \exp(-\mathrm{j}Mu_k)), \tag{9.2.10}$$

where $P_{\boldsymbol{w},\boldsymbol{k}}$ is the wanted power in direction u_k, and similarly for $H_{\mathrm{un},k}$ for the unwanted power. The SINR is then given by

$$SINR = \frac{\boldsymbol{w}^H(\boldsymbol{R}_{\boldsymbol{w},\mathbf{1}} + \boldsymbol{R}_{\boldsymbol{w},\mathbf{2}} + \cdots)\boldsymbol{w}}{\boldsymbol{w}^H(\boldsymbol{N}_{\mathbf{0}} + \boldsymbol{R}_{\boldsymbol{un},\mathbf{1}} + \boldsymbol{R}_{\boldsymbol{un},\mathbf{2}} \cdots)\boldsymbol{w}} \tag{9.2.11}$$

where the summation is over as many angles as necessary. Note that it is possible to take the angular spread into account by shaping the powers as a function of angle. The addition of powers implies uncorrelated scattering.

9.2.2 Gain enhancement

Especially in the downlink with an unknown channel and just an average direction known, it is important to have the extra gain available from the array. It should be mentioned that increased gain is a mixed blessing in an interference limited system, since the increased gain will lead to larger interference in some directions. However, in general it is an advantage in easing penetration into buildings and difficult areas. Gain here means the power relative to one element with the antenna weights set for maximum free space gain in an LOS situation. This means that all the weights of the antenna elements have the same magnitude, but with the phase progression adjusted to the direction in question. This is also denoted the *beam mode* in the following. Assuming that the wanted user is in the broadside direction, the normalised weight vector, comprising M elements, reduces then to

$$w = (1, 1, 1, \ldots, 1)^{\mathrm{T}}/\sqrt{M}. \tag{9.2.12}$$

The gain now depends on the correlation matrix, or the angular spread. The general expression in equation (9.2.4) may still be used if $\boldsymbol{R}$ is replaced by an identity matrix. Again, the two cases of completely uncorrelated and completely correlated signals will be used, and for simplicity a two element array will be used in the examples.

9.2.2.1 Uncorrelated case

For the uncorrelated case the $\boldsymbol{R_1}$ matrix is given by

$$\boldsymbol{R_1} = \begin{pmatrix} |V_1|^2 & V_1 V_2^* \\ V_1^* V_2 & |V_2|^2 \end{pmatrix} \tag{9.2.13}$$

and the average value over many different realisations (i.e. phases, magnitudes and locations of the scatterers with a given scenario) is

$$E\{\boldsymbol{R_1}\} = \begin{pmatrix} E\{|V_1|^2\} & 0 \\ 0 & E\{|V_2|^2\} \end{pmatrix}. \tag{9.2.14}$$

Using equations (9.2.10) and (9.2.12) in the modified (9.2.4) yields a gain relative to one element

$$\text{Gain} = \big(E\{|V_1|^2\} + E\{|V_2|^2\} + \cdots\big)/M = 1. \tag{9.2.15}$$

Thus the mean gain in the uncorrelated elements case for a fixed beam is unity, independent of the number of elements. Another way of interpreting this is that as the beam gets more and more narrow, it illuminates less and less of the medium, and the total power stays constant. It is the same mechanism that makes a measurement of the antenna temperature of a uniform spatial distribution of antenna noise independent of the antenna gain.

9.2.2.2 Correlated case

In the completely correlated case

$$\boldsymbol{R_1} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{9.2.16}$$

and the gain equals

$$\text{Gain} = (M + M + \cdots + M)/M = M \tag{9.2.17}$$

which is the normal array gain relative to one element. It is interesting to observe that the average maximum gain in the general case with optimum weights is given by equation (9.2.6)

$$\text{Gain}_{\max} = \big(E\{|V_1|^2\} + E\{|V_2|^2\} + \cdots\big) = M \tag{9.2.18}$$

which is independent of the correlation or angular spread. But for the beam mode the gain decreases from M to one when the correlation varies from one to zero. The in-between cases are treated numerically in Figure 9.2.7, which shows the gain for $M = 4$ and $M = 8$ as a function of angular spread.

For a 10° spread, which is not unrealistic for an outdoor macro cell, the eight-element array loses about 3 dB of its free space gain.

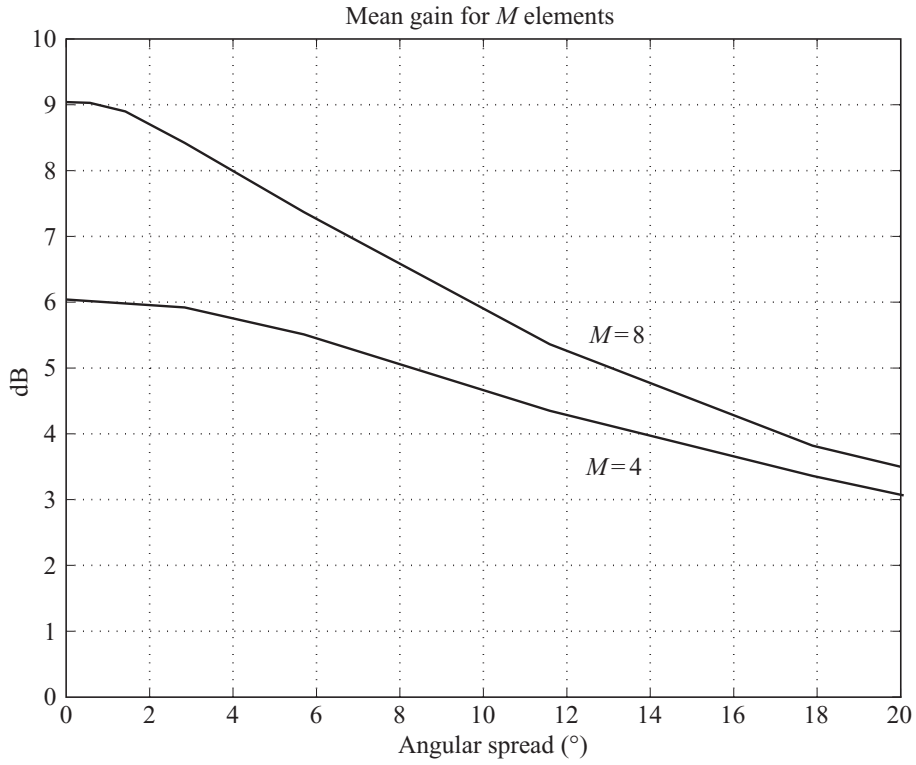


Figure 9.2.7 The average broadside antenna gain (relative to one element) for an array with M elements in a beam mode using constant weights as a function of the angular spread of the scatterers as seen from the antenna

9.2.3 Discussion

The optimum combining technique used on the uplink at a base station is an excellent method for reducing the level of interference and thus increasing the capacity of the network. In this narrowband description the only requirements needed are as follows.

- The wanted user can be identified through a training sequence of a special user code, thus making it possible to estimate the array factor or propagation factor $\boldsymbol{H_1}$. The accuracy depends on the length of the training sequences or codes, their orthogonality, and the *SNR*.
- The correlation matrix $\boldsymbol{R}$ can be estimated correctly as a sum of the correlation matrices of the individual users. This presupposes an averaging process where all cross terms between the signals from different users are eliminated due to orthogonality.

It is noteworthy that when these conditions are satisfied, then equation (9.2.6) will always give the maximum value of the signal energy relative to interference and noise, *independent of the correlation between elements and of the number of interferers.*

For the completely correlated case with zero angular spread seen from the base, two users may not come from the same direction. If they do, the vectors of transmission

factors are the same. However, even a small angular spread will help in separating the vectors, even if the mean directions are unresolvably close.

For the downlink, only the average direction may be used for maximum gain, when the channel is unknown, and the beam mode gain gets smaller as the angular spread increases. There are, however, cases with TDD (time division duplex) and slowly moving terminals, when the channel is reciprocal, and the optimum combining will work equally well in downlink.

9.3 Multi-element arrays: MIMO systems

In the previous section it was assumed that there was only one antenna element at one end of the link, or rather the other antennas belonged to different users. Here, the case of several elements at each end is treated, and by utilising these extra channels additional gain or information capacity may be achieved. The situation may be described by Figure 9.3.1 in the simple case of two two-element arrays.

The transmission from the transmitter (assumed to be at the left) to the receiver is now described by a matrix

$$\boldsymbol{H} = \begin{pmatrix} H_{11} & H_{21} \\ H_{12} & H_{22} \end{pmatrix}. \tag{9.3.1}$$

At this point it is advantageous to use the singular value decomposition (Scharf, 1991; Teletar, 1996) of the matrix, which lends itself to an elegant description of the situation.

9.3.1 Singular value decomposition, SVD

The SVD is a way of diagonalising the matrix and finding the eigenvalues. The matrix $\boldsymbol{H}$ will in general be rectangular with N rows and M columns corresponding to the N transmit elements and the M receive elements. Associated with it will be an inner product Gram matrix, which will be quadratic with M columns and rows ($M > N$),

$$\boldsymbol{G} = \boldsymbol{H}^H \boldsymbol{H} \tag{9.3.2}$$

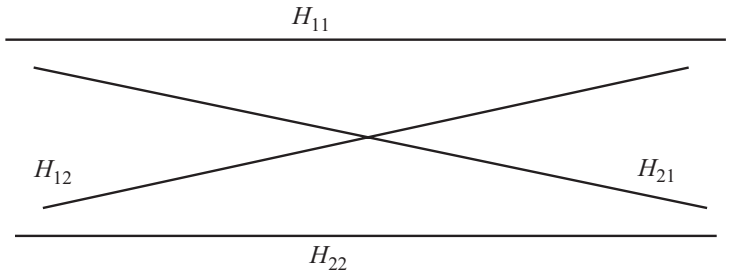


Figure 9.3.1 Two two-element arrays with four paths

and in general this Hermitian matrix will have $n = \min(M, N)$ distinct, real positive eigenvalues, and the remaining will be zero

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_n \geq 0. \tag{9.3.3}$$

An SVD expansion is a description of $\boldsymbol{H}$ itself as

$$\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^H \tag{9.3.4}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of real, non-negative singular values, the square roots of the eigenvalues of $\boldsymbol{G}$, above. The columns of the orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are the corresponding singular vectors. Since $\boldsymbol{G}$ may be written as follows

$$\boldsymbol{G} = \boldsymbol{H}^H\boldsymbol{H} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\boldsymbol{T}}\boldsymbol{V}^H, \tag{9.3.5}$$

$\boldsymbol{V}$ contains the eigenvectors of $\boldsymbol{G}$.

The SVD is particularly useful for interpretation in the antenna context. Writing equation (9.3.4) differently, using just one of the eigenvalues and its eigenvector,

$$\boldsymbol{H}\boldsymbol{V}_{\boldsymbol{1}} = \sqrt{\lambda_1}\boldsymbol{U}_{\boldsymbol{1}} \tag{9.3.6}$$

and it is clear that $\boldsymbol{V}_{\boldsymbol{1}}$ is the transmit weight factor for excitation of the singular value $\sqrt{\lambda_1}$ with the gain of λ_1. A receive weight factor of $\boldsymbol{U}_1^H$, a conjugate match, gives the receive voltage and the square of the received power

$$\begin{aligned} V_\mathrm{r} &= \boldsymbol{U}_{\boldsymbol{1}}^H\boldsymbol{U}_{\boldsymbol{1}}\sqrt{\lambda_1} = \sqrt{\lambda_1} \\ P_\mathrm{r} &= |V_\mathrm{r}|^2 = \lambda_1. \end{aligned} \tag{9.3.7}$$

Thus the eigenvalues and their distributions are important measures of the possible uses that can be extracted from the medium. An important property is that the trace of $\boldsymbol{G}$ is the sum of the eigenvalues,

$$\mathrm{Tr}(\boldsymbol{G}) = \sum_i \lambda_i. \tag{9.3.8}$$

The $(M, N) = (2, 2)$ case may illustrate the case. $\boldsymbol{G}$ is given by

$$\boldsymbol{G} = \begin{bmatrix} |H_{11}|^2 + |H_{12}|^2 & H_{11}H_{21}^* + H_{12}H_{22}^* \\ H_{11}^*H_{21} + H_{12}^*H_{22} & |H_{22}|^2 + |H_{21}|^2 \end{bmatrix} = \begin{bmatrix} a & c \\ c^* & b \end{bmatrix} \tag{9.3.9}$$

and the two eigenvalues are

$$\lambda_{\max} = \frac{1}{2}\left(a + b + \sqrt{(a-b)^2 + 4|c|^2}\right) \tag{9.3.10}$$

and

$$\lambda_{\min} = \frac{1}{2}\left(a + b - \sqrt{(a-b)^2 + 4|c|^2}\right). \tag{9.3.11}$$

Note that

$$\mathrm{Tr}(\boldsymbol{G}) = \sum_i \lambda_i = a + b = |H_{11}|^2 + |H_{12}|^2 + |H_{21}|^2 + |H_{22}|^2. \tag{9.3.12}$$

Although the sum in equation (9.3.12) on average equals four (this is relative to one element at each end), the distribution among the eigenvalues is uneven,

$$E\{\lambda_{\max}\} = 3.5, \quad E\{\lambda_{\min}\} = 0.5 \tag{9.3.13}$$

The pdf of ordered eigenvalues may be found in James (1964) from which the distributions for $\lambda_{\min}$ and $\lambda_{\max}$ may be derived. The results are, recalling that $\lambda \geq 0$,

$$p_{\lambda_{\min}} = 2\mathrm{e}^{-2\lambda} \tag{9.3.14}$$

$$p_{\lambda_{\max}} = \mathrm{e}^{-\lambda}(\lambda^2 - 2\lambda + 2) - 2\mathrm{e}^{-2\lambda}, \tag{9.3.15}$$

which for the minimum eigenvalue indicates a Rayleigh distribution with mean power 0.5. The distributions are shown in Figure 9.3.2, where the maximum eigenvalue follows the fourth order maximum ratio diversity distribution (the distribution

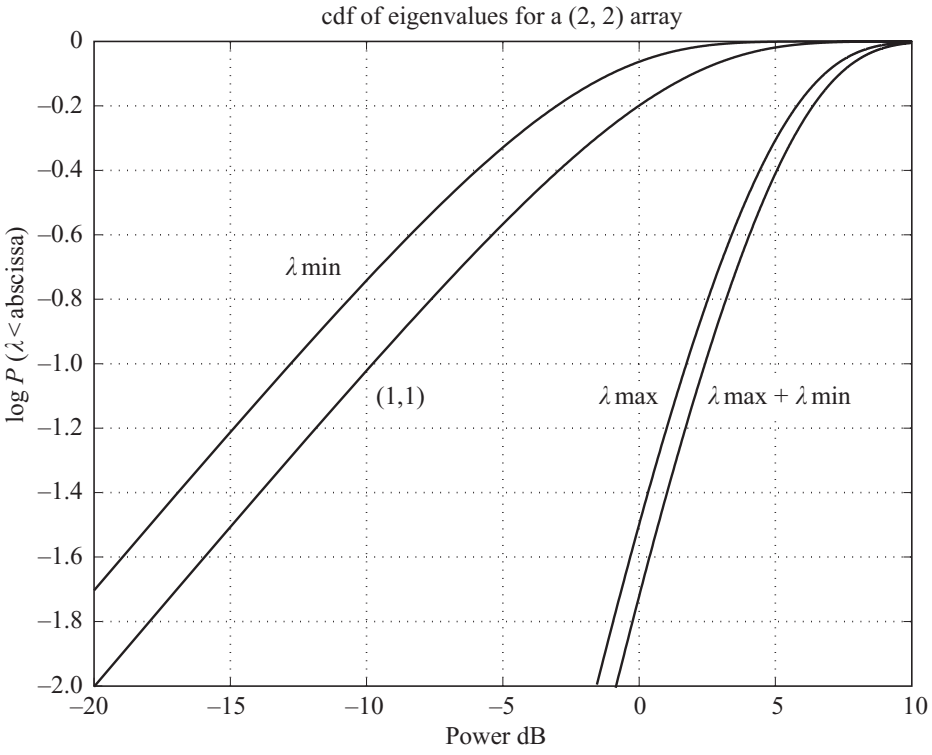


Figure 9.3.2 *The cdf of eigenvalues for an $(M, N) = (2, 2)$ array with four uncorrelated paths. The maximum eigenvalue follows closely the fourth-order diversity valid for the sum of the two eigenvalues. The gain is normalised to the average power of one path. (1,1) means one antenna at each end*

of the trace and the sum of the two eigenvalues) quite closely. The cumulative density function for $\lambda_{\max}$ is

$$\begin{aligned} \text{Prob}(\lambda_{\max} < x) &= 1 - \mathrm{e}^{-\lambda}(\lambda^2 + 2) + \mathrm{e}^{-2\lambda} \\ &= \lambda^4/12 \quad \lambda \ll 1, \end{aligned} \tag{9.3.16}$$

indicating that the (2, 2) case displays full fourth-order diversity but with twice the probability relative to the (4, 1) case, or a shift of $2^{1/4} = 0.75\,\text{dB}$ in power for the same probability.

In order to make full benefit of these eigenvalues it is of course a necessity to know the channel at the transmitter, otherwise the eigenvectors cannot be found. An alternative would be to select a beam mode from the transmitter, but this would reduce the situation to the one eigenvalue case described in the previous sections.

9.3.2 Maximum gain and transmit–receive diversity for a known channel

It is natural again to divide the discussion between the two cases of a known channel on the transmit side and an unknown channel. First we will look at the known channel case. The (2, 2) case above may be generalised to more elements, and as an example consider the (4, 4) case in Figure 9.3.3.

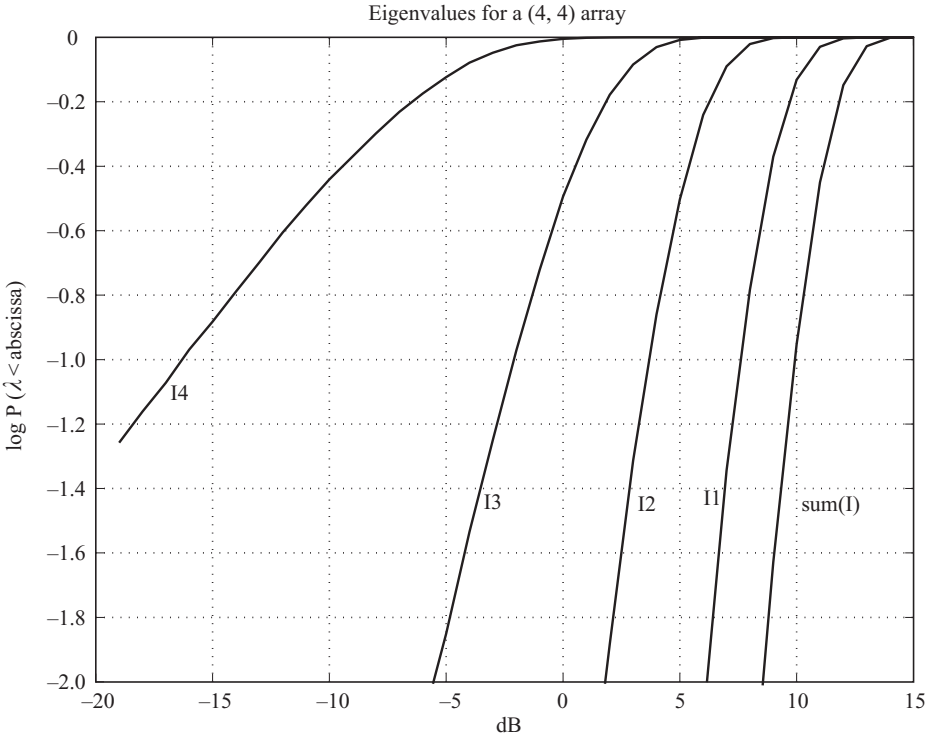


Figure 9.3.3 Cumulative distribution of eigenvalues (power) for two arrays each of four elements including the total power equal to the sum of eigenvalues

The array has 16 different uncorrelated paths, so if we can utilise all the eigenvalues in a transmit-diversity situation we will achieve gain of 16, exactly as in free space without the scattering medium, and a diversity order of 16. This is, however, not possible since the maximum gain corresponds to the largest eigenvalue, which has a mean of about 10 (see Figure 9.3.5). The reason for this reduction is that the available weight vectors on both sides only allow a total of eight complex weights, not enough for the combining of 16 complex paths. This reduction of the gain is a general feature of MIMO systems, since $M + N < MN$.

The smallest eigenvalue λ_4 has an insignificant value, like the smaller eigenvalue in the previous case, but the first three all add significantly to the total power, and they are all above the (1, 1) path.

For a specific eigenvalue the SVD description in equation (9.3.6) gives directly the proper weight factors at the transmit and receive side. This is illustrated for the first two eigenvalues in Figure 9.3.4.

Recent results concerning the distribution of the eigenvalues of a random Hermitian matrix can give some insight into the maximum gain and how it varies with M and N. In the asymptotic limit when N is large it may be shown (German, 1980; Silverstein, 1985; Haargerup and Thorbjønsson, 1998) that the largest eigenvalue is bounded above by

$$\lambda_{\max} < \left(\sqrt{c} + 1\right)^2 N, \quad c \geq 1, \tag{9.3.17a}$$

where the integer c equals M/N, and the smallest eigenvalue is bounded below by

$$\lambda_{\min} > \left(\sqrt{c} - 1\right)^2 N, \quad c > 1. \tag{9.3.17b}$$

In the previous examples $c = M/N = 1$, and the upper asymptotic bound is thus $4N$.

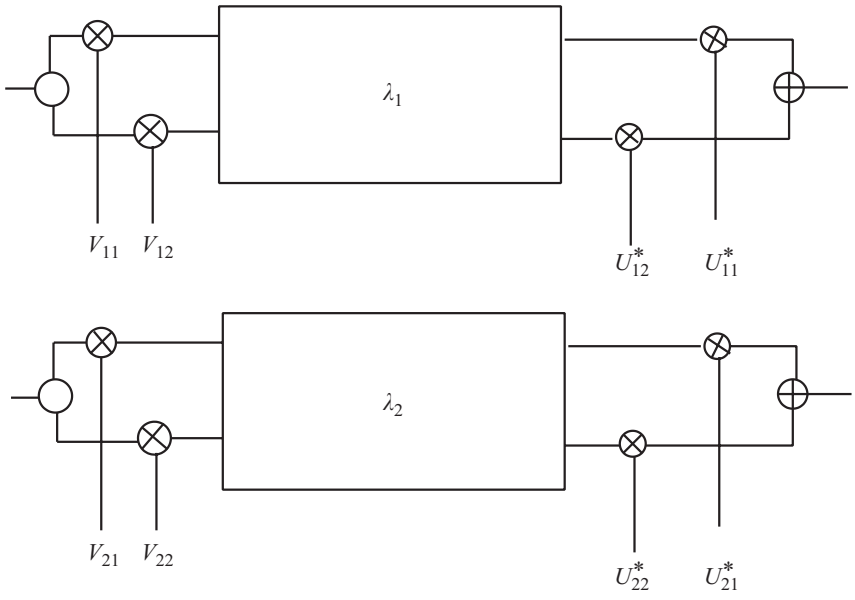


Figure 9.3.4 Excitation of separate eigenvalues for realising transmit–receive diversity in agreement with equation (9.3.4)

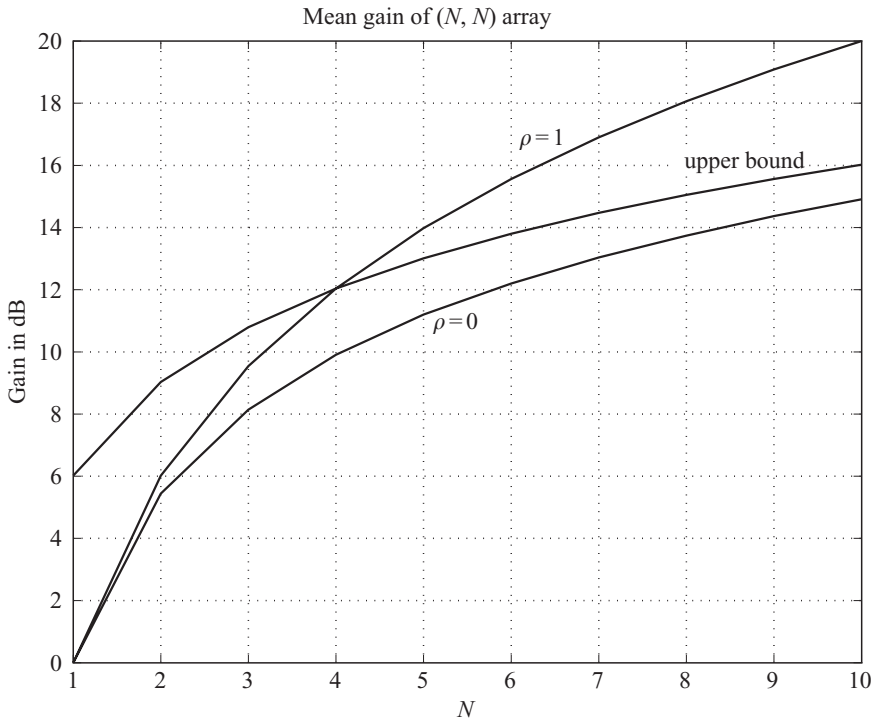


Figure 9.3.5 The gain relative to one element of (N, N) arrays in a correlated situation ($\rho = 1$), which is N^2 and in an uncorrelated case ($\rho = 0$), which is N. The upper bound is $4N$, and is the asymptotic upper bound for the maximum eigenvalue for large N

The mean gains for the maximum eigenvalue for uncorrelated signals ($\rho = 0$) are shown in Figure 9.3.5 together with the upper bound and the gain for the correlated, $\rho = 1$, case, N^2. For $N = 10$, for example, the true mean gain is just 1 dB below the upper bound. Thus the price to pay for the random scattering is a diminishment of the gain from N^2 to $4N$ for N large. For a partly correlated case we can expect the gain to lie between the $\rho = 0$ and the $\rho = 1$ cases.

In some situations it might be advantageous to have more antennas on one side than on the other, especially for asymmetric situations with heavy downloading of data from a base station. Again the asymptotic, upper bound for the largest eigenvalue is useful (equation (9.3.17a)). Introducing M explicitly we find

$$G_{\text{upper bound}} = \left(\sqrt{M} + \sqrt{N}\right)^2, \tag{9.3.18}$$

which asymptotically will approach M for large values of M and fixed N, and is $4M$ for $M = N$.

9.3.3 Maximum gain and transmit–receive diversity for an unknown channel

An unknown channel means, as before, that the channel is unknown at the transmitter, but tracked by the receiver, e.g. in our (2, 2) example all four paths are tracked. The

situation is symbolically described in Figure 9.3.6, where the power is divided equally between the two input antenna ports in lack of knowledge of the channel. The signals at the two transmit antennas are, however, coded differently, symbolically noted by a and b. We need not be concerned at this stage how this coding is performed, only that we are able to extract all four paths on the receive side.

The total power received is then given by

$$P_{\max} = \tfrac{1}{2}|H_{11}|^2 + \tfrac{1}{2}|H_{12}|^2 + \tfrac{1}{2}|H_{21}|^2 + \tfrac{1}{2}|H_{22}|^2 \qquad (9.3.19)$$

so there is a 3 dB loss relative to full transmit–receive diversity, but it is still fourth-order diversity. In general with M antennas on both sides, the average gain is M and the diversity order is M^2.

It seems advantageous to use the time dimension in so-called space-time coding to achieve the results above. Winters (1998) uses re-transmission of symbols from the different antennas in the general M antenna case, while Alamouti (1998) transmits two symbols over the two antennas in one time slot and the same symbols shifted and conjugated in the following time slot.

9.3.4 Summary of gain and diversity for multi-arrays

The various situations with low and high correlations (or high or low angular spreads) at the two ends of the link may be summarised in Tables 9.3.1, 9.3.2 and 9.3.3 (Andersen, 2000a) for M antennas at one end and N antennas at the other end. For ease of explanation one side is called the transmitter (M antennas) and the other the receiver (N antennas), but of course there is reciprocity.

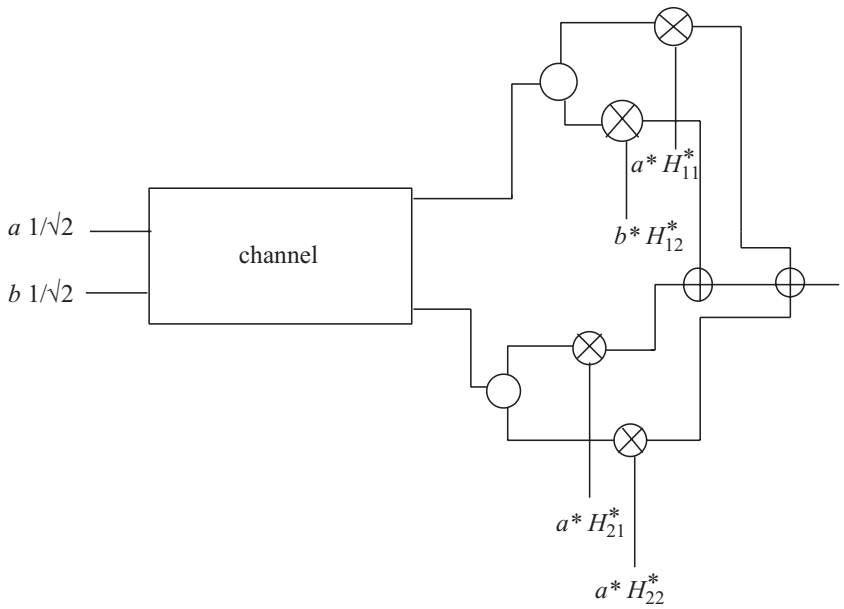


Figure 9.3.6 Transmit signals, channel, and the receiver for the (2, 2) case of transmit–receive diversity when the channel is known at receiver only

Table 9.1 Mean link gain for beam mode for both arrays

	Receiver, low ang. spread	Receiver, high ang. spread
Transmitter, low ang. spread	MN	M
Transmitter, high ang. spread	N	1

Table 9.2 Mean link gain for maximum gain combining for both arrays

	Receiver, low ang. spread	Receiver, high ang. spread
Transmitter, low ang. spread	MN	MN
Transmitter, high ang. spread	MN	$\left(\sqrt{M}+\sqrt{N}\right)^2$

Table 9.3 The diversity order of the link for maximum gain combining for both arrays

	Receiver, low ang. spread	Receiver, high ang. spread
Transmitter, low ang. spread	1	N
Transmitter, high ang. spread	M	MN

As an example consider the upper right corners of the tables with low angular spread from the transmit side and high angular spread on the receive side. In the beam mode the transmitter sees a point source and has a gain of M, the receiver a gain of one, so the link gain is M. In the maximum combining mode at the receiver the mean gain is N, so the total gain is MN. The diversity order (the number of independent paths) is N, since there is no diversity possible at the transmitter side. Further examples of finite correlations are given in Section 9.3.5

9.3.4.1 Gain and capacity impact on data rates, range and frequency

The data rate achievable over a link is a function of many factors like modulation, error distribution, coding, etc., but in all cases a certain energy per bit is required, the well-known E_b/N_0. When the data rate increases the needed power increases with the data rate R. This is why the link antenna gain, the mean value and the diversity gain,

are so important for wideband services, since the achievable data rate is proportional to the power gain.

The impact on the range depends on the decay of power with distance, so for a power law like $P = P_0 d^{-n}$ the range will vary with gain like $d = d_0 G^{1/n}$. As an example we can choose $n = 3.5$. If the required power is increased by a factor of ten for a given data rate, the range is reduced by a factor of two, unless the gain is increased correspondingly by a factor of ten. Four elements at each end will give roughly a mean gain of about 10 dB.

As far as the carrier frequency is concerned the situation is more complicated, since it depends on how the path loss varies with frequency. The famous Hata law for urban propagation gives a frequency dependence of $f^{-2.6}$ for the received power, which corresponds approximately to the free space law and a shadow diffraction. This is valid for constant gain antennas, but if we instead consider constant area antennas filled with adaptive antenna elements, the situation changes dramatically. For all the situations discussed above where the gain was MN the received power now increases with the square of the frequency. The worst case is the case of wide angular spreads, where the gain is reduced, and in this situation the received power is independent of frequency. Thus there seem to be possibilities of going up to higher frequencies such as 10–20 GHz for the mobile links, if the adaptive arrays can be made economically feasible.

9.3.5 Spectral efficiency of parallel channels

A look at Figure 9.3.4 with two independent channels makes it clear that there are other uses of the eigenvalues than diversity combining them for maximum power. Another use is to keep them as parallel channels with independent information. Shannon's capacity measure gives an upper bound on the realisable information rates through parallel channels (Gallagher, 1968), and how the power should be distributed over the channels to achieve maximum capacity through 'water filling'. The basic expression for the spectral efficiency measured in bits s^{-1} Hz^{-1} for one Gaussian channel is given by

$$C = \log_2(1 + P) \text{ bits s}^{-1} \text{ Hz}^{-1} \tag{9.3.20}$$

where P is the signal-to-noise ratio, SNR, for one channel or the signal power for unit noise power. In the context of transmit–receive diversity, or multiple-input-multiple-output (MIMO) systems, C is called simply *capacity*, although in most other areas of communications this would be called the capacity efficiency since the units are bits s^{-1} Hz^{-1}.

For N such channels the capacity is

$$C = N \log_2 \left(1 + \frac{P}{N}\right), \tag{9.3.21}$$

where the total power P is divided equally between the N channels. It is noticed that there are substantial increases in capacity as a function of N, if P/N is large, and this

is exactly what can be achieved in a random medium, where P includes the transmit and receive diversity gains discussed above.

The eigenvalue expansion in the previous sections is very useful here again, since it may be used for the optimum distribution of power between the parallel channels.

Assuming all noise powers of the eigenchannels seen at the receiver to be the same, P_i is the input power to the ith eigenchannel. At the receiver, the *SNR* of the ith eigenchannel is $P_i\lambda_i$ (note that in this section the terminology $P_i = SNR_i/\lambda_i$ is used) so for unity noise power of the eigenchannels at the receiver, the equivalent noise power at the transmitter end is $1/\lambda_i$. The 'water filling' concept is the solution for the maximum capacity, where each channel is filled up to a common level D,

$$\frac{1}{\lambda_1} + P_1 = \frac{1}{\lambda_2} + P_2 = \frac{1}{\lambda_3} + P_3 = \cdots = D, \tag{9.3.22}$$

as in Figure 9.3.7.

Thus the channel with the highest gain receives the largest share of the power. The constraint on the powers is that

$$\sum P_i = P. \tag{9.3.23}$$

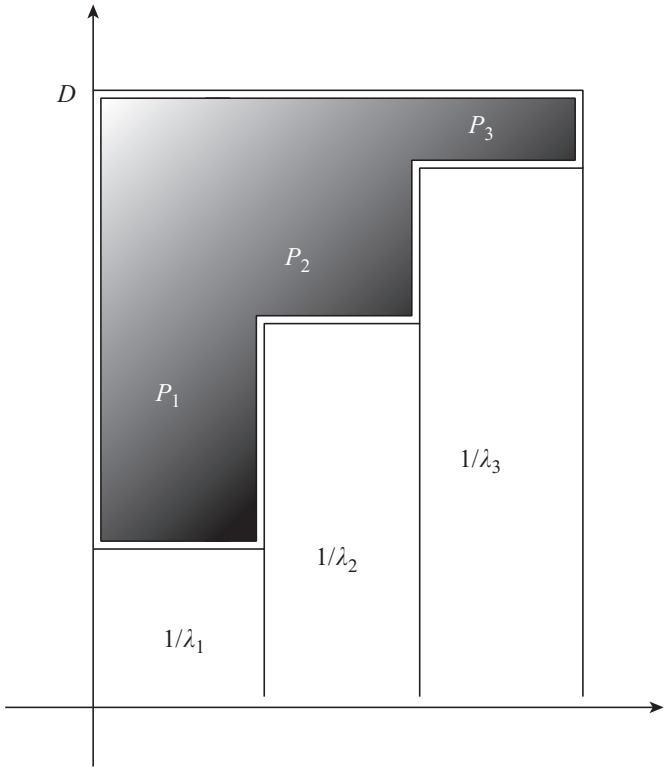


Figure 9.3.7 'Water filling' the independent channels with different gains λ_i. P_i is the average SNR to the ith channel

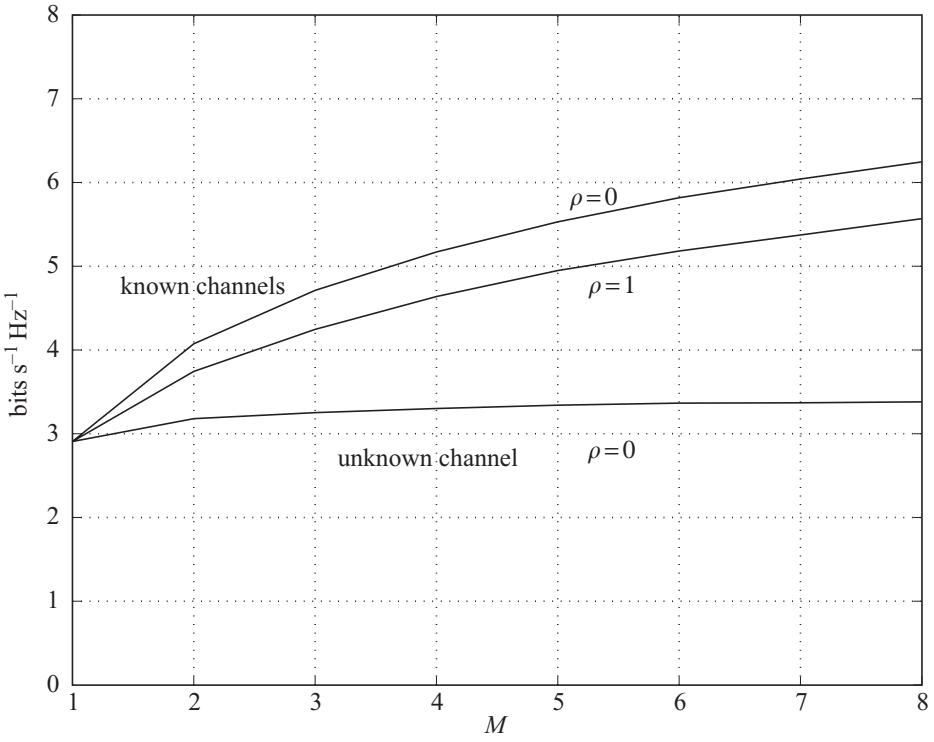


Figure 9.3.8 *The mean capacity for a transmit array as a function of the number of transmit array antennas, and one receive antenna. The channel may be known or unknown at the transmitter. The antenna signals are spatially uncorrelated for $\rho = 0$, while they are correlated for $\rho = 1$. The SNR at the receiver is 10 dB*

In case the level D drops below the highest $1/\lambda$ then that power is set to zero; for example P_3 would be set to zero in Figure 9.3.7 for $D\lambda_3 < 1$.

The capacity (bits s^{-1} Hz^{-1}) equals

$$C = \sum_{N'} \log_2(1 + \lambda_i P_i) = \sum_{N'} \log_2(\lambda_i D), \tag{9.3.24}$$

where the summation is over all channels with non-zero powers. The water filling is of course dependent on the knowledge of the channels on the transmit side, i.e. the known channel in the terminology of the previous sections.

In the case where the channel is unknown at the transmitter the only reasonable division of power is a uniform distribution over the channels, i.e.

$$P_i = \frac{P}{M}, \tag{9.3.25}$$

M being the number of transmit antennas.

9.3.5.1 Capacity for M transmit antennas and one receive antenna

Since the number of eigenvalues is $\min(N, M)$ there is only one eigenvalue in this case, so

$$C_{\text{known}} = \log_2\left(1 + P\lambda_1^{(M)}\right) \tag{9.3.26}$$

for a known channel, and

$$C_{\text{unknown}} = \log_2\left(1 + \frac{P}{M}\lambda_1^{(M)}\right) \tag{9.3.27}$$

for an unknown channel, and where the distribution of the single eigenvalue of course depends on M, the number of transmit elements. As noted earlier, both the average value of the sum of eigenvalues and the diversity order equals MN, so in this case the average value of λ equals M, and the mean value of the argument of the log function is $1 + P$. The mean value of C_{unknown} is thus close to the case of a single antenna at each end, but the probability distribution is quite different owing to the diversity effect.

The normal receive case with $M = 1$ and N receive antennas, is also covered by equation (9.3.26), since there is no difference between receive and transmit diversity for a known channel. The mean values are shown in Figure 9.3.8 for *SNR* equal to 10 dB. Also shown is the correlated case which differs from the uncorrelated case only because of the log function; the mean of the argument is the same as discussed earlier.

9.3.5.2 Capacity of the (M, N) array

The situation is much more interesting when there are several antennas at each end of the link, and here the discussion is limited to the symmetric array with equal numbers of antennas at each end. It follows from Section 9.3.1 that there are several independent orthogonal channels to be used simultaneously.

Figures 9.3.9 and 9.3.10 support this approximate analysis for the mean values of capacity.

The mean capacity $\langle C \rangle$ is found by simulation, where the individual matrix elements are assumed to be uncorrelated complex Gaussians with zero mean and unit variance. Note that the mean capacity is different from the capacity of the mean power.

Figure 9.3.9 shows the symmetric case where $M = N$ for an SNR of 10 dB, and it is noted that capacity grows linearly with N for the uncorrelated case, while for the correlated case there is only one channel, which then grows only logarithmically with N. Figure 9.3.10 indicates that the incremental increase in capacity with an increase in power is much larger for the multi-antenna case.

The previously shown upper and lower bounds on the eigenvalues may be used to give some approximate bounds on the capacity. It follows (Andersen, 2000b) that

$$\left(\sqrt{M} - \sqrt{N}\right)^2 < \lambda_i < \left(\sqrt{M} + \sqrt{N}\right)^2. \tag{9.3.28}$$

Thus if $M > N$ the N eigenvalues tend to cluster around M, and they will all be allocated the same power, namely P/N, so we arrive at the following approximate formula for C

$$C = N \log_2(1 + MP/N). \tag{9.3.29}$$

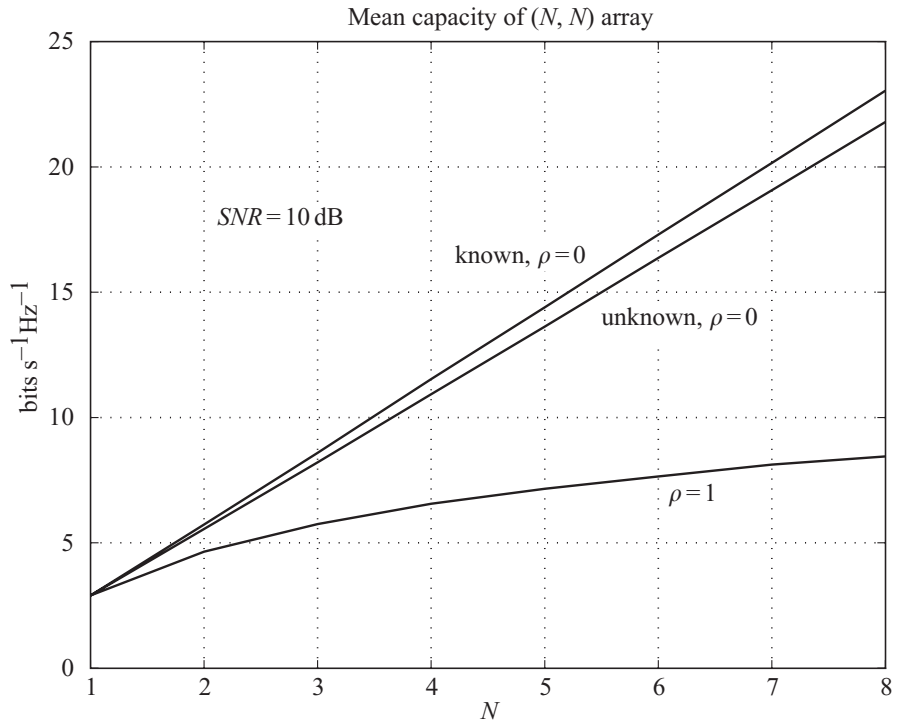


Figure 9.3.9 Mean capacity for two arrays each of N elements. The capacity grows linearly with the number of elements and is approximately the same for the known and the unknown channel. The total transmitted power is constant

The interpretation is simple: there are N independent channels, which share the power equally, and they all have a gain of M.

In the case where the channels are known at the transmitter it is straightforward to realise these capacities, while space-time coding is needed when this is not the case. Some interesting results are available in Foschini (1996) and Tarokh *et al.* (1998).

9.3.6 Effect of correlations and pinholes

In the preceding sections the channel matrix was assumed to consist of uncorrelated, zero mean complex Gaussians. This is an ideal situation not always fulfilled in practice, since small angular spreads will lead to non-zero correlations among the antennas. In the following various environments will be defined, which highlight the various phenomena. The model (Figure 9.3.11) consists of antenna arrays of M and N elements illuminating two sheaths of random scatterers of width W. The distances between the arrays and the neighbouring sheaths are d_M and d_N, respectively, and the distance between the sheaths is d_S. In order to simplify the situation symmetry is assumed, $L_M = L_N = L$.

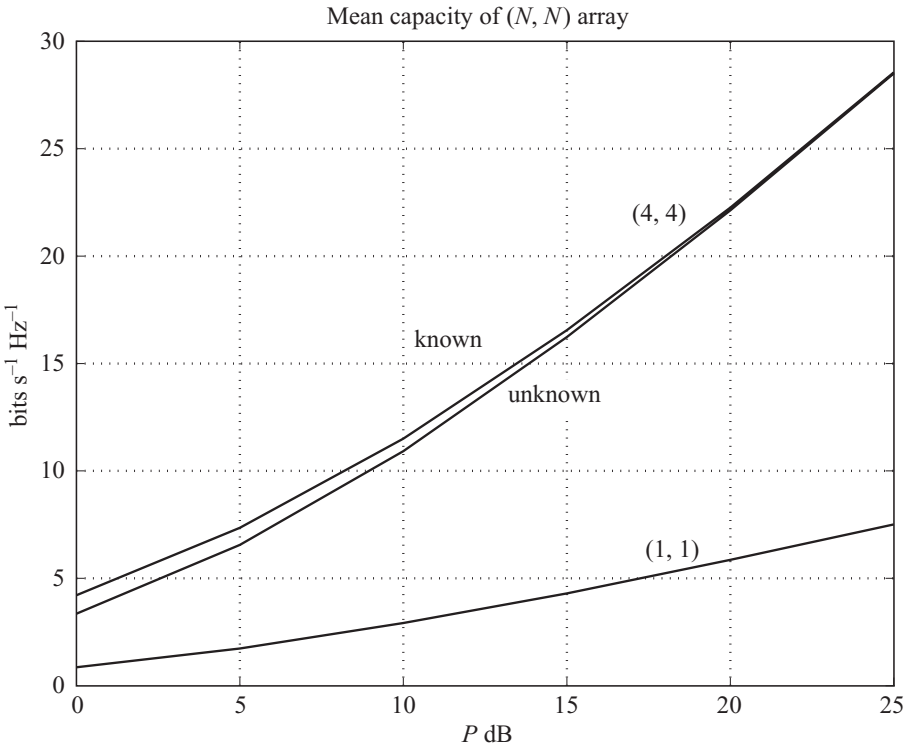


Figure 9.3.10 Mean capacity of (4, 4) array as a function of signal-to-noise ratio

In order to classify the various situations two dimensionless parameters are introduced

$$F_{\text{cor}} = \frac{LW}{d\lambda} \tag{9.3.30}$$

$$F_{\text{pin}} = \frac{W^2}{d_{\text{S}}\lambda} \tag{9.3.31}$$

where λ is the wavelength, $d = d_M = d_N$ is the distance from array to nearer screen, and d_{S} the distance between screens. We have here assumed symmetry around the middle, but it is clear that parameters may be defined separately for the two sides. The parameter F_{cor} is roughly the difference in electrical length between the array ends and the screen edge, and similarly for F_{pin} between the two screens. When F_{cor} is small ($\ll 1$) the screen looks like a point source seen from the array, and the antenna weights are uniform with no possibility for diversity. On the other hand, for $F_{\text{cor}} \gg 1$ the fields are spatially fading over the array, and diversity gain may be achieved. When F_{pin} is small the two screens are in each other's far fields, while for F_{pin} large, they are in the radiating near fields of each other. We can roughly define four different situations.

9.3.6.1 Effect on gain and diversity

(a) $F_{\text{cor}} \ll 1$ and $F_{\text{pin}} \ll 1$
This corresponds to a situation, where everything looks like a point source. The total gain equals MN, and there is no diversity gain. There is only one eigenvalue,

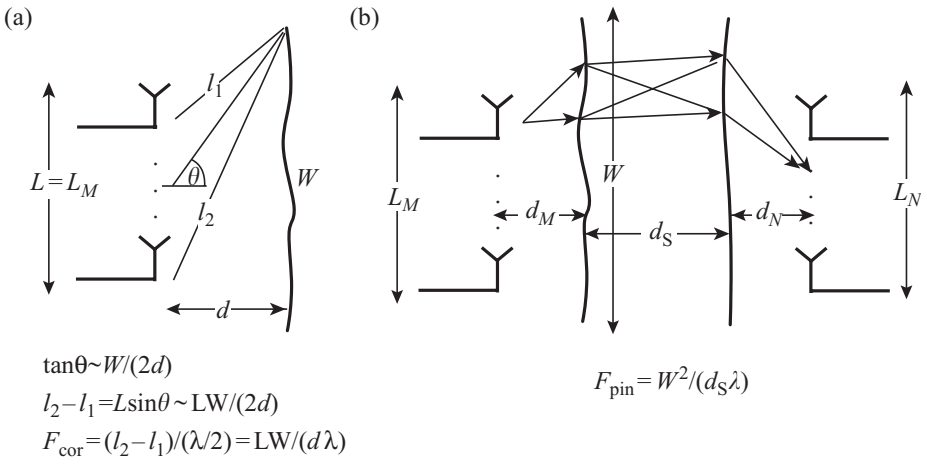


Figure 9.3.11 *(a) The parameter F_{cor} is the electrical path difference between the array edges and the screen edge. (b) Two arrays with M and N elements of total length L_M and L_N spaced apart by a distance $D = d_M + d_S + d_N$ illuminate a number of scatterers on two sheaths of width W. Distance from array to scatterer is d_M and d_N. F_{pin} is similar to F_{cor}, but for the screens. A few ray paths are shown*

since there is effectively only one effective path between the two screens. It acts like a pinhole (Gesbert *et al.*, 2000; Chizhik *et al.*, 2000), which explains the name F_{pin}. Since the link consists of a product of the two Rayleigh fading scatterers, the distribution is now a double-Rayleigh. Mathematically the probability density for the power is a modified Bessel function of zero order (Haagerup and Thorbjørnsson, 1998).

(b) $F_{cor} \gg 1$ and $F_{pin} \ll 1$
Locally around the arrays there is wide angular scattering, but the ensemble of scatterers still looks like a point from the other side, and there is only one eigenvalue. Each side may be optimised independently with full diversity gain, which also means a mean link array gain of MN, and a diversity order of MN.

(c) $F_{cor} \ll 1$ and $F_{pin} \gg 1$
The two sheets act like one scattering point in the far field of the arrays. Independent optimisation on each side, no diversity, mean gain of MN, and the distribution is Rayleigh.

(d) $F_{cor} \gg 1$ and $F_{pin} \gg 1$
Complete decorrelation at all points; the optimisation at the receiver depends on the optimisation at the transmitter, so it is a case of joint optimisation. The gain is reduced relative to the other situations. The diversity order is high.

Several situations are simulated by assuming that the screen consists of a number (40) of random scatterers, modelled as complex Gaussians, and each scatterer on one screen is radiating to each element on the other screen (Figure 9.3.11). The distance between array elements is half a wavelength. Distances are varied to simulate various values of F_{cor} and F_{pin}.

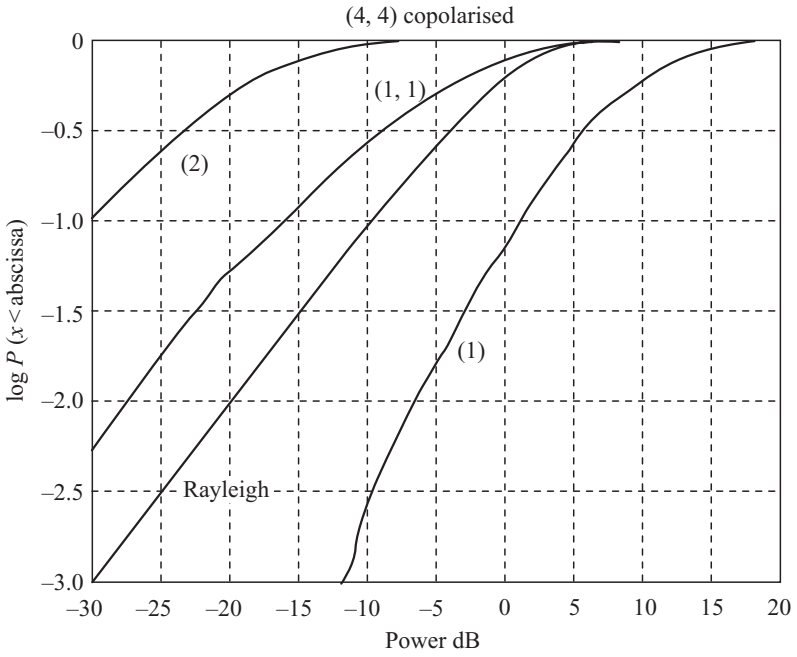


Figure 9.3.12 *Cumulative probability distribution for two four-element arrays.* $F_{\text{cor}} = 1$ *and* $F_{\text{pin}} = 0.1$. *The four curves from the right are the maximum eigenvalue channel, a reference Rayleigh distribution, a reference case of one antenna at each end, and the next largest eigenvalue channel. The two remaining channels have a gain which is too small to be shown*

An illustrative example is shown in Figure 9.3.12 for partly correlated four-element arrays and some pinhole effect ($F_{\text{cor}} = 1$, $F_{\text{pin}} = 0.1$). The (1, 1) case is for one-element arrays, and it is seen that it is worse than Rayleigh, since it approaches a so-called double-Rayleigh, a product of two Rayleigh distributions. There should be four channels (eigenvalues), but only two have sufficient gain to appear in the graph. The mean gain for the maximum eigenvalue is close to 12 dB (6 dB gain at each end) as it should be relative to the (1, 1) case, and some diversity is noted, since the slope of the curve is higher than for the (1, 1) case.

9.3.6.2 Examples of correlations and pinholes

The variation with F_{cor} and F_{pin} is indicated in Figures 9.3.13 and 9.3.14. Figure 9.3.13 shows the case of $F_{\text{pin}} = 10$ (highly interactive screens) as a function of F_{cor}. The $SNR = 10$ dB, and the figure shows the mean number of active channels, N_{channel}, the gain in dB of the best channel at the 10 per cent level, and the mean capacity in bits s^{-1} Hz^{-1}. At the left side of Figure 9.3.13, where F_{cor} is small, there is only one channel, no diversity gain and the basic capacity of 7 bits s^{-1} Hz^{-1} for the one channel with 12 dB mean gain. In this region there is no reason to spread the signal over the antennas since there is only one channel. At the right side we get the four

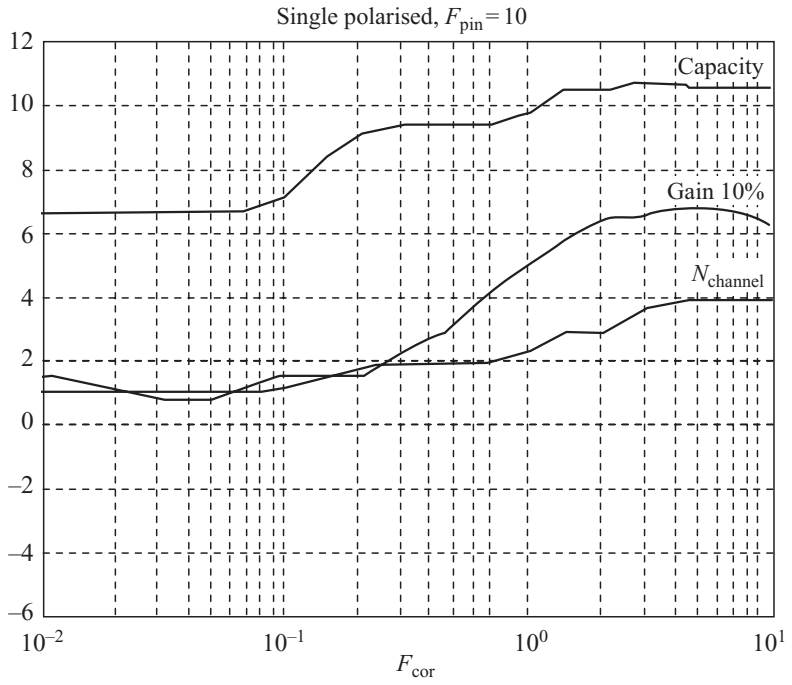


Figure 9.3.13 *Two four-element arrays with $F_{pin} = 10$, and $SNR = 10\,dB$. $N_{channel}$ is the mean number of active parallel channels, Gain 10 per cent is the gain in dB at the 10 per cent outage level, and Capacity is the capacity in bits $s^{-1}Hz^{-1}$*

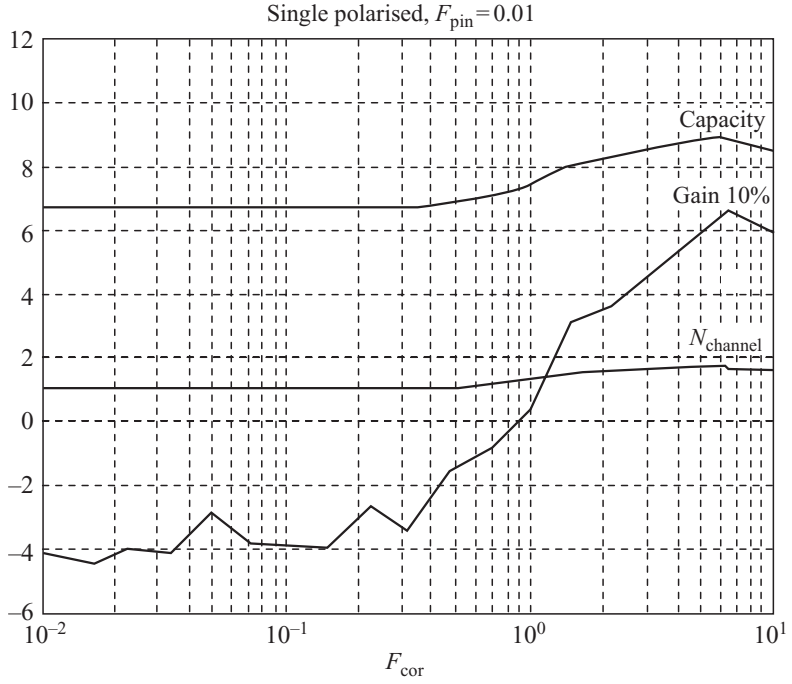


Figure 9.3.14 *As in Figure 9.3.13, but with $F_{pin} = 0.01$*

channels with some diversity gain, and the capacity increases to 11 bits s^{-1} Hz^{-1}. The crossover point is between values of F_{cor} of 0.1–1.

Figure 9.3.14 represents the pinhole effect with $F_{pin} = 0.01$. There is a moderate increase in capacity with maximally two channels, but a fair diversity gain since the −4 dB at the left derives from the double-Rayleigh distribution.

9.4 Outage for optimally combined receiving arrays with many interferers

9.4.1 Introduction: analogue and digital outage

The performance of an array in multi-user digital mobile communications is complicated to assess because of the many different transmission and signal processing processes. The basic analogue measure of the antenna array receiving performance is the SINR of the signal after combination at the array, in which it is assumed that all other components of the reception are working ideally. This SINR, as a system performance measure, therefore does not depend on the coding and modulation schemes, etc., which of course are a part of the system and its performance. Aspects of these schemes are often used in the implementation of the signal combining algorithm for the array.

The accuracy of the signal combining algorithm, under certain conditions, relates to the SINR. Specifically, the mean square error (MSE, defined below) of the combiner is given by Monzingo and Miller (1980) as $MSE = 1/(1 + SINR)$, and various performance measures can be based on this.

Measures related to the digital channel use the bit-error ratio (BER). For example, the digital link outage probability can be defined as

$$\mathrm{P}(BER_0) = \mathrm{Prob}(BER \geq BER_0). \tag{9.4.1}$$

A bound, for static channels, is given by Foschini and Salz (1983) and Winters *et al.* (1994) as

$$BER \leq \alpha \mathrm{e}^{-a \cdot SINR}, \tag{9.4.2}$$

where α and a are dependent on the modulations. For example, for DPSK, $(\alpha, a) = (0.5, 1)$ (Jakes, 1974; Winters and Salz, 1994) and for erfc function formulations of the BER, see Smith (2002). For the case of DPSK, the outage is therefore bounded by

$$\mathrm{P}(BER_0) \leq \mathrm{Prob}\left(SINR \leq \frac{\ln(\alpha/BER_0)}{a}\right), \tag{9.4.3}$$

so a distribution for the *SINR* allows bounds on the outage of a digital link.

9.4.1.1 Performance of optimum combining

A maximum ratio combiner for an M-branch antenna with uncorrelated noise and Rayleigh fading will result in the signal-to-noise ratio (SNR) being distributed as a chi-square with $2M$ degrees of freedom (Appendix C). In optimum combining where

the number of interferers is less than the number of antenna branches, the N interferers can be, in principle, nulled out and the resulting SINR becomes the same as the SNR, and this is distributed as a chi-square with $2(M - N)$ degrees of freedom.

If the number of interferers becomes asymptotically large and the interferers are uncorrelated and power-limited, with no dominant contributions, then it is intuitive that their contribution cannot be distinguished from the uncorrelated noise. So optimum combining becomes the same as maximum ratio combining in this case, and the distribution of the SINR becomes the same as that of the SNR, a chi-square random variable with $2M$ degrees of freedom, denoted χ^2_{2M}. This is the same as a gamma with mean M, with its random variable denoted g_M, in the sense that $\chi^2_{2M} = 2g_M$.

When the number of uncorrelated interferers is finite but greater than the number of antenna branches (a realistic situation), the analysis for the performance becomes complicated. This section presents cdf results for canonical situations and an approximation for a more general situation using a standard, purely statistical model of the array.

9.4.2 Statistical model of receiving array with interferers

The model is depicted in Figure 9.4.1, and has sources representing the user's terminals radiating into a multipath environment, which then illuminates a receiving array antenna with a combiner and additive noise. The notation used for this section follows.

There are $N + 1$ source signals, transmitted from the $N + 1$ terminals, and these signals are written here as $\boldsymbol{x} = (x_0, x_1, \ldots, x_N)^{\mathrm{T}}$, where x_0 is the wanted source, and the N others are interferers. The sources are modelled with independent random variables with zero mean. The sources are therefore assumed uncorrelated with each

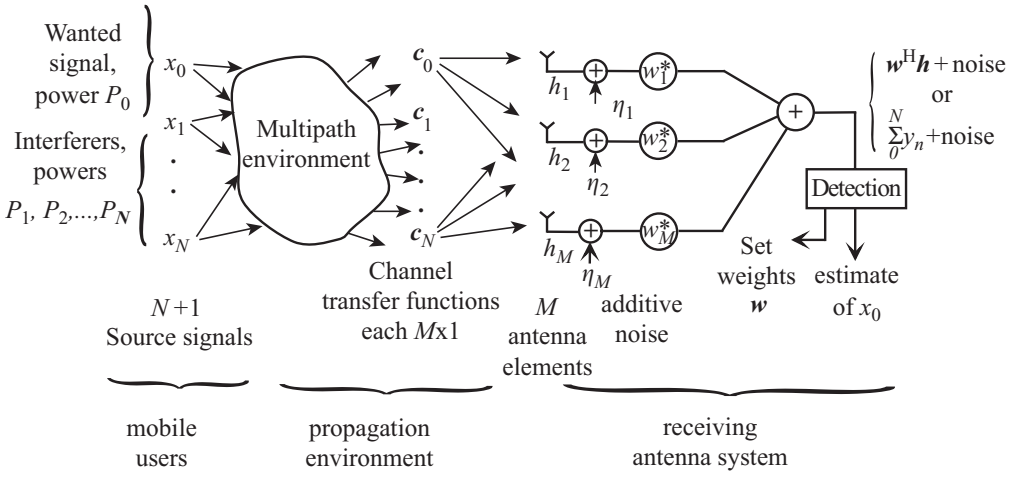


Figure 9.4.1 The model and notation for the multiuser mobile signals incident on a receiving array. The combined signal can be expressed as a sum of the received signals from the individual sources (y_n) plus the noise contribution, or as the sum of the weighted branch signals, each of which comprises contributions from all of the sources, $\boldsymbol{w}^H \boldsymbol{h}$ plus noise

other, i.e. $E\{x_i^* x_j\} = P_i \delta_{i,j}$, and the source powers, which are also random variables, are denoted by P_i. The total interferer power is constrained in the sense that the average transmitted power is finite, namely $1/N \sum_{n=1}^{N} P_n = \bar{P}$.

Further below, the form $x_n = P_n^{1/2} z_n$ where $z_n \sim \mathcal{N}(0, 1)$ and here z is real, is used for generality. The usual Rayleigh fading case is where z_n has a complex Gaussian distribution whereby x_n is also complex Gaussian.

9.4.2.1 Transfer function and its covariance

The notation for the general statistical model for the SINR requires modification from that of Section 9.2. Here, the transfer functions of each of the N sources to each of the M antenna elements, are assembled in the $M \times 1$ vectors $\boldsymbol{c}_0, \boldsymbol{c}_1, \ldots, \boldsymbol{c}_N$. These are zero-mean, complex Gaussians with covariance matrix denoted here $E\{\boldsymbol{c}_n \boldsymbol{c}_n^H\} = \sigma_c^2 \boldsymbol{I}_M$, where $\boldsymbol{I}_M$ is the $M \times M$ identity matrix, and σ_c^2 is a channel gain expressed as a variance (see below). The distribution of the $\boldsymbol{c}_n$ is expressed

$$\boldsymbol{c}_n \sim C\mathcal{N}_M(\boldsymbol{0}, \sigma_c^2 \boldsymbol{I}_M), \tag{9.4.4}$$

(the $C\mathcal{N}$ abbreviates complex normal) where $\boldsymbol{0}$ is the $M \times 1$ vector of zeros. The power transfer functions, $|\boldsymbol{c}_n|^2$, are exponentially distributed, i.e. chi square with two degrees of freedom, or a gamma, denoted $\chi_2^2/2 = g_1$. The gamma variable is convenient here and used for expressing the cdf approximation result below.

The transfer functions can be assembled into an $M \times (N + 1)$ matrix as $\boldsymbol{c} = [\boldsymbol{c}_0\, \boldsymbol{c}_1 \cdots \boldsymbol{c}_N]$, where $\boldsymbol{c}$ is called the transmission matrix. Each of the transfer functions has variance σ_c^2, which can be interpreted as the gain of the channel and array elements, and the covariance matrix is

$$E\{\boldsymbol{c}\boldsymbol{c}^H\} = \sigma_c^2 (N + 1) \boldsymbol{I}_M; \tag{9.4.5}$$

σ_c^2 can be set to unity without any loss of generality in the statistics; it is included to help trace it through the succeeding equations.

The signals in the M diversity branches are written as the $M \times 1$ vector $\boldsymbol{h} = \boldsymbol{c}\boldsymbol{x}$. For example, the first branch has the signal given by the first element of $\boldsymbol{h}$, denoted h_1 (note that h_1 is a scalar), and $h_1 = c_{01}x_0 + c_{11}x_1 + \cdots + c_{N1}x_N$, where c_{nm} is the nmth element of $\boldsymbol{c}$.

The sample covariance matrix (i.e. a 'snapshot' estimate of its expectation value) of the source signals received at the antenna elements, $\boldsymbol{R}_{\boldsymbol{h}}$, is from a snapshot of the $\boldsymbol{h}$. It is expressed in terms of the transmission matrix as $\boldsymbol{R}_{\boldsymbol{h}} = \boldsymbol{c}\boldsymbol{x}\boldsymbol{x}^H\boldsymbol{c}^H$. In practice, the snapshot may comprise some time averaging over l samples, often expressed in the form $\boldsymbol{R}_h = \sum_l \boldsymbol{h}_l \boldsymbol{h}_l^H$.

9.4.2.2 Array weights and output power from the sources

The complex weights of the antenna are denoted by the $M \times 1$ vector $\boldsymbol{W}$ and the weights are constrained to have finite gain, for example, $|\boldsymbol{W}|^2 = \boldsymbol{W}^H \boldsymbol{W} = M$.

At the receiver, the combined signal featuring contributions from all the sources (without the noise) is $y = \boldsymbol{W}^H \boldsymbol{u}$. It is convenient to define $M \times 1$ vectors describing

the signal from the nth source in the M antenna branches as

$$\boldsymbol{u}_n = \boldsymbol{c}_n x_n. \tag{9.4.6}$$

For example, the signal after the combiner received from the nth source only (no noise is included) is $y_n = \boldsymbol{W}^H \boldsymbol{c}_n x_n$, and the signal resulting from all the sources is $y = \sum_{n=0}^{N} y_n$. The power of the received source signals (again, no noise is included) after the combiner is $|y|^2 = \boldsymbol{W}^H \boldsymbol{h}\boldsymbol{h}^H \boldsymbol{W}$; cf. equation (9.2.3) for a single source.

9.4.2.3 Noise

The additive noise at each antenna element (denoted η_M in Figure 9.4.1) is complex Gaussian and has variance σ^2 in each receiving branch. (Note that the σ^2 relates to the noise and the σ_c^2 relates to the channel.) The noise vector is $\eta = [\eta_1, \eta_2, \ldots, \eta_M]^{\mathrm{T}}$ and so the noise signal after the combiner is $\boldsymbol{W}^H \eta$.

At cellular communication frequencies, the noise is dominated by thermal noise from the front end of each receiver, and so is Gaussian and independent between the branches of the array. In interference-limited systems, this noise does not play an important role of course, except for the situation of a very large number of interferers, or very weak interferers, whose contributions may appear, or be consigned mathematically, as noise.

9.4.2.4 Sample covariance matrix

The sample covariance matrix of the interference plus noise requires some explanation. We take samples of the signals in the antenna branches, after the additive noise, and form a sample covariance matrix by averaging the samples. The channel and the signals are considered as unchanging during this sampling, but the noise is changing. This averaging process therefore samples a snapshot of the channel but the covariance matrix of the noise contribution becomes (essentially) diagonal owing to its uncorrelated and changing nature during the sampling and averaging. So at the antenna elements, the covariance matrix of the noise contribution is written

$$\overline{R_{\mathrm{noise}}} = \overline{\eta\eta^H} \approx E\{\eta\eta^H\} = \sigma^2 \boldsymbol{I}_M \tag{9.4.7}$$

(the overbar indicates the averaging) and the approximation becomes an equality as the number of samples becomes asymptotically large.

9.4.2.5 Optimum weights for maximum SINR and minimum MSE

The optimal weight, $\boldsymbol{W}_0$, maximises performance in the sense $SINR_{\max} = \left(\boldsymbol{W}_0^H \boldsymbol{R}_{\mathrm{S}} \boldsymbol{W}_0\right) / \left(\boldsymbol{W}_0^H \boldsymbol{R}_{\mathrm{NI}} \boldsymbol{W}_0\right)$, where

$$\boldsymbol{R}_{\mathrm{S}} = \boldsymbol{u}_0 \boldsymbol{u}_0^H = (\boldsymbol{c}_0 x_0)(\boldsymbol{c}_0 x_0)^H \tag{9.4.8}$$

is the $M \times M$ sample covariance matrix of the wanted signal only. The covariance matrix of the interference-plus-noise signals only (no wanted signal) is

$$\begin{aligned} \boldsymbol{R}_{\mathrm{NI}} &= \boldsymbol{R}_{\mathrm{I}} + \overline{\boldsymbol{R}_{\mathrm{Noise}}} \\ &= \sum_{n=1}^{N} \boldsymbol{u}_n \boldsymbol{u}_n^H + \sigma^2 \boldsymbol{I}_M \\ &= \sum_{n=1}^{N} |x_n|^2 \boldsymbol{c}_n \boldsymbol{c}_n^H + \sigma^2 \boldsymbol{I}_M, \end{aligned} \tag{9.4.9}$$

where σ^2 is recalled as the variance, or power, of the noise.

The optimum weights are (cf. equation (9.2.5)),

$$\boldsymbol{W}_0 = \left(\frac{\boldsymbol{W}_0^H \boldsymbol{R}_{\mathrm{NI}} \boldsymbol{W}_0}{\boldsymbol{W}_0^H \boldsymbol{u}_0} \right) \boldsymbol{R}_{\mathrm{NI}}^{-1} \boldsymbol{u}_0 \tag{9.4.10}$$

and the term in brackets is a complex constant. This optimum weight vector can also be found as the minimum mean square error solution, which minimises the square difference between the estimate of the received wanted signal, and the true value of the wanted signal, i.e. $MSE = |\boldsymbol{W}_0^H \boldsymbol{u}_0 - x_0|^2$.

In some notations, the $\boldsymbol{u}_0$ (or $\boldsymbol{c}_0$) in equation (9.4.10) has a complex conjugate, and this arises from defining the covariance in the form $\boldsymbol{u}^* \boldsymbol{u}^{\mathrm{T}}$ instead of $\boldsymbol{u}\boldsymbol{u}^H$, used here. The weights appear with conjugates in Figure 9.4.1 because of the choice of notation here.

In practice, it may not be straightforward to estimate the above covariance matrices because the signals from the various sources need to be separated and the performance of the required multi-user detection process may become a limiting factor. For example, to estimate $\boldsymbol{R}_{\mathrm{NI}}$ requires, in principle, complete removal of the wanted signal from the received signal. But if the interference-plus-noise and the wanted signal covariance matrices can be separated, namely $\boldsymbol{R} = \boldsymbol{R}_{\mathrm{S}} + \boldsymbol{R}_{\mathrm{NI}}$, then the solution can be expressed as the well known result above, $\boldsymbol{W}_0 \propto \boldsymbol{R}^{-1} \boldsymbol{u}_0$. The necessary condition for being able to separate the covariance matrices is that the interferers are uncorrelated with the wanted signal. Note that $\boldsymbol{c}_0$ can be used instead of $\boldsymbol{u}_0$ in the weights equation.

Finally, the power of the received signal after combination at the receiving array (this is wanted signal-plus-noise-plus-interference) is $|y + \boldsymbol{W}^H \eta|^2 = \boldsymbol{W}_0^H \boldsymbol{R} \boldsymbol{W}_0 \propto \boldsymbol{u}_0^H \boldsymbol{R}^{-1} \boldsymbol{u}_0$.

9.4.3 SINR for a larger number of interferers than array elements

The maximum SINR has a similar quadratic form to that for the total signal power, but with the difference that the covariance matrix now contains only the interference-plus-noise (cf. equation (3.94) in Monzingo and Miller, 1980), which is written (the

subscript 'max' is dropped from now on)

$$\begin{aligned} SINR &= \boldsymbol{u}_0^H \boldsymbol{R}_{\mathrm{NI}}^{-1} \boldsymbol{u}_0 \\ &= \boldsymbol{u}_0^H \left(\sum_{n=1}^{N} |x_n|^2 \boldsymbol{c}_n \boldsymbol{c}_n^H + \sigma^2 \boldsymbol{I}_M \right)^{-1} \boldsymbol{u}_0. \end{aligned} \tag{9.4.11}$$

The distribution of this maximum SINR is the quantity of interest. Rearranging,

$$N \cdot SINR = \boldsymbol{u}_0^H \left(\frac{\boldsymbol{R}_{\mathrm{NI}}}{N} \right)^{-1} \boldsymbol{u}_0, \tag{9.4.12}$$

and

$$\frac{\boldsymbol{R}_{\mathrm{NI}}}{N} = \bar{\boldsymbol{X}} + \frac{\sigma^2}{N} \boldsymbol{I}_M, \tag{9.4.13}$$

where a mean interference covariance matrix is defined as

$$\bar{\boldsymbol{X}} = \frac{1}{N} \sum_{n=1}^{N} X_n \tag{9.4.14}$$

in which

$$X_n = |x_n|^2 \boldsymbol{c}_n \boldsymbol{c}_n^H. \tag{9.4.15}$$

As the number of interferers, N, becomes large, $\bar{\boldsymbol{X}}$ approaches its true mean value, $E\{\bar{\boldsymbol{X}}\}$, and also $E\{\bar{\boldsymbol{X}}\}$ approaches $\sigma_c^2 E\{\bar{P}\} \boldsymbol{I}_M$ as defined in the statistical model above. Using these results gives

$$N \cdot SINR \approx \frac{\boldsymbol{u}_0 \boldsymbol{u}_0^H}{\sigma_c^2 E\{\bar{P}\}} = g_M \frac{|x_0|^2}{E\{\bar{P}\}} \quad (\text{large } N). \tag{9.4.16}$$

This result is independent of the distributions of the powers of the interfering signals except for the mean of the total powers, $\bar{P}$. It is the same as the case of *no interferers* in the sense that the formulation is exact for an infinite number of interferers with finite total power and no dominant interferers, and this is the same statistical situation as the assumed noise. The noise term has dropped out because it is negligible, and it is 'replaced' with the interferer average power term.

Another way to write this result is

$$\begin{aligned} SINR &\approx \left(N E\{\bar{P}\} \sigma_c^2 \right)^{-1} \boldsymbol{u}_0^H \boldsymbol{u}_0 \quad (\text{large } N), \\ &= \frac{1}{N} \frac{1}{E\{\bar{P}\}} \left(\frac{|x_0|^2}{\sigma_c^2} \right) \boldsymbol{c}_0^H \boldsymbol{c}_0. \end{aligned} \tag{9.4.17}$$

From the assumption that the quantity $\boldsymbol{c}_0^H \boldsymbol{c}_0$ is distributed as a χ_{2M}^2 (i.e. as the gamma-distributed random variable, $2g_M$), it follows that the SINR, with an expectation taken over the changing channel and interferers, follows the same distribution. Note that the expectation does not include the wanted source, x_0; recall that the wanted source is not in $\boldsymbol{R}_{\mathrm{NI}}$.

9.4.4 Cumulative distribution function for SINR

9.4.4.1 Case for fixed interferer powers

In the case where interferer powers are fixed, then the SINR can be conditioned on the powers, $P = \{P_1, P_2, \ldots, P_n\}$. This allows a solution (Gao *et al.*, 1998) for the cumulative distribution for the SINR, here showing the conditioning on P,

$$\text{Prob}(SINR|_P < \beta) = 1 - \mathrm{e}^{-\beta/\gamma_0} \sum_{m=1}^{M} \frac{A_m(\beta)}{\Gamma(m)} \left(\frac{\beta}{\gamma}\right)^{m-1}, \tag{9.4.18}$$

where $\gamma = P_0\sigma_c^2/\sigma^2$ is a signal-to-noise ratio in the array branches before the combiner, and

$$A_m(\beta) = \begin{cases} 1, & M \geq N + m \\ \dfrac{1 + \sum_{i=1}^{M-m} C_i \beta^i}{\prod_{n=1}^{N}(1 + (P_n/P_0)\beta)}, & M < N + m \end{cases} \tag{9.4.19}$$

where C_i is the coefficient of β^i in $\prod_{n=1}^{N}(1 + (P_n/P_0)\beta)$. This coefficient can be expressed as

$$C_i = \sum_{n_1=1}^{N} \sum_{n_2=n_1}^{N} \cdots \sum_{n_i=n_{i-1}}^{N} \left(\frac{P_1}{P_0}\right)\left(\frac{P_2}{P_0}\right)\cdots\left(\frac{P_i}{P_0}\right). \tag{9.4.20}$$

The quick calculation of this term is a major problem for conveniently assessing large arrays in the presence of many interferers. Gao *et al.* give a recursion method for the $A_m(\beta)$ allowing evaluation of equation (9.4.18).

9.4.4.2 Common, fixed interferer powers and no noise

For the special case of all interferer powers the same, and no noise, there is a simplified solution. Putting $P_n \equiv P_1$, and $\sigma^2 = 0$, results in (see also equation (9.4.42) for a further simplification for the no-noise case)

$$\text{Prob}(SINR|_{P_n \equiv P_1} < \beta) = 1 - \frac{\sum_{i=1}^{M-1} \binom{N}{i} ((P_1/P_0)\beta)^i}{(1 + (P_1/P_0)\beta)^N}. \tag{9.4.21}$$

9.4.4.3 Case for variable interferer powers

With ever-evolving spectral usage and allocations, cellular-type systems sometimes have to coexist in the same band with other services, such as point-to-point links which in the future will increasingly use multiple antenna techniques and adaptive modulations. Variable interferer powers may arise from terminals which vary or adapt their modulation power. For cases when the interferer powers are variable, the solution for the SINR cdf becomes more complicated. For the case of a large number of interferers, and the ratio of interferers to receiving array elements is large, $M \ll N$, a

series expansion due to C.S. Withers is possible which can be applied to approximate the cdf of the SINR. This is of the form of power terms of $1/N$, and is, here showing the conditioning on the wanted signal, x_0, (Withers and Vaughan, 2000)

$$\text{Prob}(SINR|_{x_0} < \beta) = \sum_{i=0}^{\infty} A_i(\beta, N) N^{-i} \quad (N \gg M). \tag{9.4.22}$$

This result is independent of the distributions of the powers of the interfering signals except for the mean of the total interferer power, $\bar{P}$. The expansion arises from the distribution of the gamma variable, g_M, here using x as the cumulative density value,

$$G_M(x) = \text{Prob}(g_M < x) = 1 - \mathrm{e}^{-x} \sum_{j=0}^{M-1} \left(\frac{x^j}{j!}\right) \left(= \int_0^x \frac{\mathrm{e}^{-y} y^{M-1}}{(M-1)!}\, \mathrm{d}y\right) \tag{9.4.23}$$

which leads to the form (Withers and Vaughan, 2000)

$$\text{Prob}(ZN < x) = G_M(x) + \sum_{i=1}^{\infty} \frac{1}{N^i} G_{M.i}(x) \tag{9.4.24}$$

where the $G_{M.i}$ are a linear combination of G_Ms and here Z is a dummy parameter which contains the SINR; see equation (9.4.26).

The remainder of this section gives the initial terms of the expansion which can be used to rapidly compute the cdf for the SINR in a given interference situation. The first term of the series expansion, $A_0(\beta, N)$, is the distribution of a multiple of the gamma random variable, g_M, from equation (9.4.17),

$$\begin{aligned} \text{Prob}\left(SINR|_{x_0} < \beta\right) &\approx \text{Prob}\left(g_M \frac{1}{N}\left(\frac{|x_0|^2}{E\{\bar{P}\}}\right) < \beta\right) \quad \text{(large } N\text{)} \\ &= \text{Prob}\left(g_M < \frac{N E\{\bar{P}\}}{|x_0|^2}\right) \\ &= A_0(\beta, N). \end{aligned} \tag{9.4.25}$$

So to order $1/N$, the SINR is distributed as a chi-square with $2M$ degrees of freedom (a gamma distributed variable), i.e. the same distribution as resulting from maximum ratio combining with no interferers, as noted above.

Setting $\nu = E\{\bar{P}\} + (\sigma^2/N\sigma_c^2)$, and $\beta = (|x_0|^2/N\nu)$, then the expansion can be written as

$$\text{Prob}(SINR|_{x_0} < \beta b) = G_M(b) + \sum_{i=1}^{\infty} \frac{1}{N^i} G_{M.i}(b) \tag{9.4.26}$$

where

$$G_{M.i} = \sum_{k=i+1}^{2i} \nu^{-k} \sum_{l=0}^{k} G_{M+l}(b) K_{lki}\, p_k. \tag{9.4.27}$$

K_{lki} is a row vector determined by M, with property $\sum_{l=0}^{k} K_{lki} = 0$, and p_k is a column vector determined by the moments of the instantaneous power of the sources, $Q_n = |x_n|^2 = P_n|z_n|^2$, up to order $2k$; z_n is zero mean and unit variance, and $E\{Q_n\} = E\{P_n\}$. The components of p_k are of the form

$$q(i^r) = \frac{1}{N}\sum_{n=1}^{N}\left(E\{Q_n^i\}\right)^r, \tag{9.4.28}$$

$$q(21) = \frac{1}{N}\sum_{n=1}^{N}\left(E\{Q_n^2\}\right)(E\{Q_n\}) \tag{9.4.29}$$

or, in terms of P_n and z_n,

$$q(1^r) = \frac{1}{N}\sum_{n=1}^{N}(E\{P_n\})^r, \tag{9.4.30}$$

$$q(i) = \frac{1}{N}\sum_{n=1}^{N}(E\{P_n^i\})(E\{|z_n|^{2i}\}), \tag{9.4.31}$$

$$q(21) = \frac{1}{N}\sum_{n=1}^{N}(E\{P_n^2\})(E\{|z_n|^4\})E\{P_n\}. \tag{9.4.32}$$

For the Rayleigh fading case, z_n is complex Gaussian as noted above, so $|z_n|^2$ is exponentially distributed with mean one and $E\{|z_n|^{2i}\} = i!$.

The next two terms of the expansion are given below. Typically, these are sufficient for most calculations, although convergence should be checked when $N \lesssim 2M$.

All the terms are a linear combination of a set of the distributions of the gamma variables, $\boldsymbol{G}_{2i}$, which is written

$$\boldsymbol{G}_r(b) = (G_M(b), G_{M+1}(b), \ldots, G_{M+r}(b))^{\mathrm{T}}. \tag{9.4.33}$$

The parameter

$$M_r = E\{g_M^r\} = M(M+1)\cdots(M+r-1) \tag{9.4.34}$$

is also used.

The term $i = 1$ of the expansion is

$$G_{M.1}(b) = \frac{1}{2\nu^2}\boldsymbol{G}_2^{\mathrm{T}}(b)K_{1.1}p_2, \tag{9.4.35}$$

with

$$p_2 = \left(q(2), q(1^2)\right)^{\mathrm{T}}, \tag{9.4.36}$$

$$K_{1.1} = \begin{pmatrix} 0 & -2M_2 & 2M_2 \\ M - M^2 & 2M^2 & -M_2 \end{pmatrix}^{\mathrm{T}}. \tag{9.4.37}$$

The $i = 2$ term is

$$G_{M.2}(b) = \frac{1}{3!\nu^3}\boldsymbol{G}_3(b)^{\mathrm{T}}K_{2.1}p_3 + \frac{1}{4!\nu^4}\boldsymbol{G}_4(b)^{\mathrm{T}}K_{2.2}p_4, \tag{9.4.38}$$

with

$$p_3 = \left(q(3), q(21), q(1^3)\right)^{\mathrm{T}}, \quad p_4 = \left(q(2)^2, q(2)q(1^2), q(1^2)^2\right)^{\mathrm{T}}, \tag{9.4.39}$$

$$K_{2.1} = \begin{pmatrix} 0 & 0 & 6M_3 & -6M_3 \\ 0 & 6M_2(M-1) & -6M_2(2M+1) & 6M_3 \\ 2(M-2)_3 & -6M^2(M-1) & 6M^2(M+1) & -2M_3 \end{pmatrix}^{\mathrm{T}}, \tag{9.4.40}$$

$$K_{2.2} = \begin{pmatrix} 0 & 0 & 12(M-1)_4 & -24M_3(M+1) & 12M_4 \\ 0 & 12(M-2)_4 & -12(M-1)_3(3M+2) & 12M_3(3M+1) & -12M_4 \\ 3(M-3)_4 & -12(M-2)_3M & 18M^2(M^2+1) & -12M_3M & 3M_4 \end{pmatrix}^{\mathrm{T}}. \tag{9.4.41}$$

For the no-noise case, and fixed interferer powers, the expressions can be simplified to

$$\mathrm{Prob}(SINR < \beta) = I_\nu(M, N-M+1) \quad \text{(all interferer powers equal, no noise)} \tag{9.4.42}$$

where

$$\nu = \frac{\beta(P_1/P_0)}{1+\beta(P_1/P_0)} \tag{9.4.43}$$

and I_ν is the incomplete beta function, a result originally due to R. Willink (private communication). This is useful for checking other calculated results, and may be applied to situations where the noise can be neglected, as happens with interference-limited systems. It gives the same result as equation (9.4.21).

9.4.4.4 Example

An example of the cdf is given in Figure 9.4.2. The wanted power is $P_0 = 1$, and all the interferers are the same, with $P_n = 1$. The $SNR = 3$ dB. This is unrealistically low, but serves to demonstrate the graphical results better. There are $N = 8$ interferers and $M = 4$ receiving antennas, also low numbers since the formulation is really for larger arrays with larger numbers of interferers, but again these numbers demonstrate the results conveniently. The exact result, given by the circles, is from Gao *et al.*'s result given by equation (9.4.18), and the exact no-noise result, plotted with the fine dots, is from equations (9.4.21) or (9.4.42). The first term (dashed) and the first two terms (dots) of the expansion are to the left and right of the exact result respectively, and the first three terms (solid) give a fairly accurate approximation. This oscillating nature of the series is helpful for checking convergence.

For the case $M/N = 5/8$ case, the three term approximation has good accuracy for the lower tail but departs significantly from the correct result in the range about ±4 dB of SINR. For this case, the first term alone is good to 1 to 2 dB. This sounds

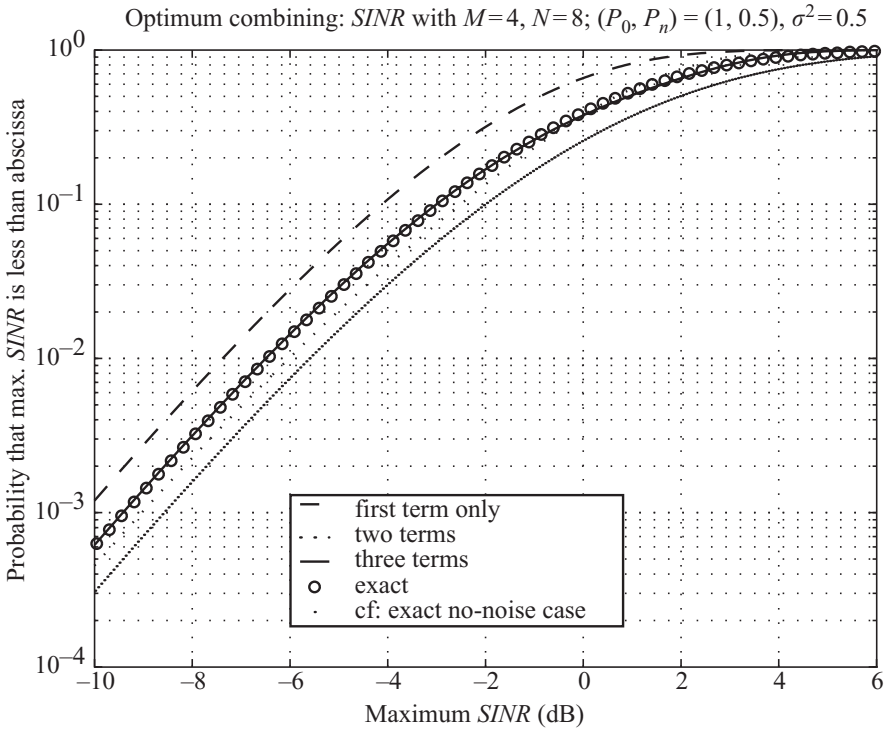


Figure 9.4.2 *The cdf of the maximum SINR for optimum combining with $M = 4$ antennas and $N = 8$ interferers, each with half the power of the wanted signal. The first term curve is the chi-square with $2M = 8$ degrees of freedom, corresponding to maximum ratio combining with no interferers. The expansion using the first three terms is sufficiently accurate for most applications, departing from the exact solution in this example at about the SINR = 1 dB point*

a small error in terms of the SINR, or analogue outage, but this can have quite an impact on the digital outage owing to the sensitivity of the BER to SINR, shown by equation (9.4.2), and plotted in Figure 9.4.3. As a rule of thumb, the three-term approximation seems accurate for these flat-fading cases where $N \gtrsim 2M$.

9.5 Capacity outage for transmit and receive arrays

9.5.1 Introduction

The recent formulations for the capacity of parallel channels have drawn more attention to the advantages of multiple antenna systems. In MIMO (multiple-input-multiple-output) systems, the capacity efficiency (or just capacity, in bits $s^{-1}Hz^{-1}$) is proportional to the minimum of the number of antennas at the transmitter or receiver, as shown in Section 9.3. For the case when the channel is known at the transmitter, the water filling solution gives the Shannon capacity directly. This section concerns the case when the channel is unknown at the transmitter.

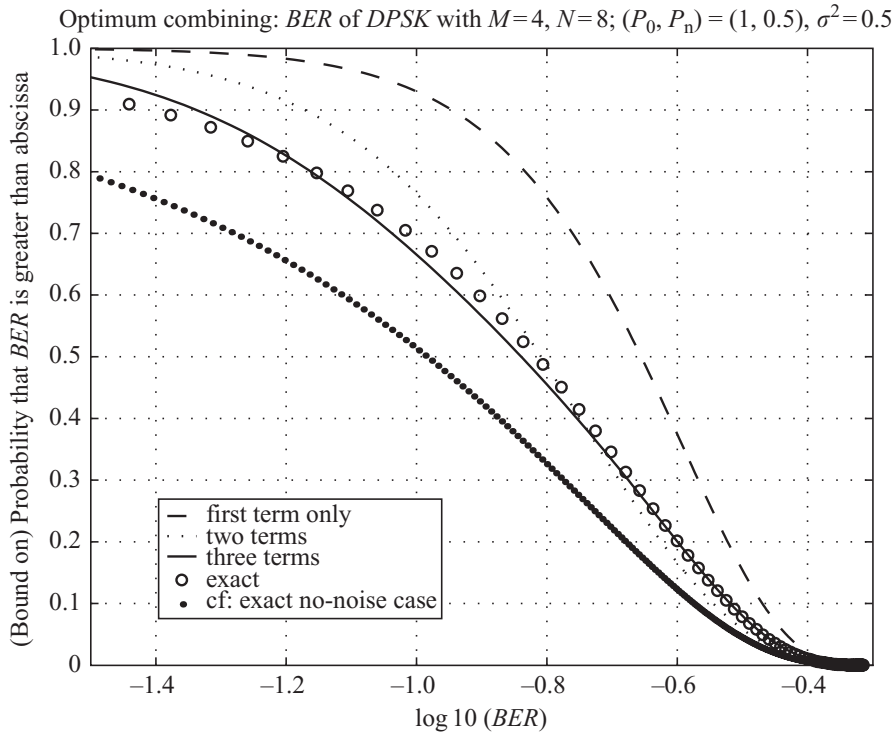


Figure 9.4.3 Digital outage: the BER bound as calculated from the various terms in the example, which has a low SINR. The BER bound behaviour is sensitive to inaccuracies in the SINR distribution

The capacity reveals limiting performance in the sense that some coding has to be applied and the block lengths of the code become large. The coding here refers to the interpretation of the basic Shannon capacity $C = \log_2(1 + SNR)$bits $s^{-1}Hz^{-1}$ defining an upper capacity limit in which it is assumed that there are infinitely long codes applied over a constant *SNR* channel. The capacity distribution is from many realisations (for example, different physical locations for the antennas and the propagation environment) of the channel.

The mean capacity has a known accurate solution (see below), and an approach is also given here for the efficient approximation of the percentile capacity outage. It is based on an expansion, due to C.S. Withers, of the asymptotic capacity (in the sense that the number of transmitter elements, N, approaches infinity), and is therefore best suited to large numbers of transmit antennas. The asymptotic mean gives optimistic values for practical antenna numbers; however, the percentile formulas tend to be accurate away from the extreme tails. If only the mean capacity is required, the approximation by Rapajic and Popescu (2001) is simple to use and accurate.

The performance measure of MIMO type systems is often described as the capacity in bits $s^{-1}Hz^{-1}$, rather than related directly to the antenna array aperture size (which nevertheless remains important), as with classical point-to-point links. The digital outage relates to the cumulative distribution of the capacity and an example for

the percentiles of the capacity is given for the Rayleigh channel. A purely statistical model is applied, as in Section 9.4, in which all the communications signal processing is assumed to be performing ideally. The realised capacity in a real-world system will of course be reduced because of the real-world performance of the signal processing. In this sense, 'capacity' is really 'potential capacity'.

Although the point-to-point capacity is the focus of the formulas, the potential impact of MIMO antennas on the spectral efficiency and implementation of cellular systems is profound. The implementation of cellular systems with MIMO affects most basic system issues such as the cellular layout, networking/data distribution, handoff, coding and modulation. The communication system, signal processing and coding issues are not discussed further here. The following draws from Withers and Vaughan (2001).

9.5.1.1 Basic MIMO capacity formula for unknown channel at the transmitter

The starting point is Foschini and Gans's (1998) formula for the capacity of a static channel, N-transmit, M-receive system, (cf. Section 9.3.5),

$$C = \log_2 \left| \boldsymbol{I}_M + \frac{1}{N} \frac{P_\mathrm{T}}{Q} \sum_{n=1}^{N} \boldsymbol{H}_{n0} \boldsymbol{H}_{n0}^H \right| \quad (\text{bits s}^{-1}\text{Hz}^{-1}). \tag{9.5.1}$$

An equivalent formula to this is also given by Winters (1987). Here, $\boldsymbol{I}_M$ is the $M \times M$ identity matrix, P_T is the average total (across all transmit elements) transmitted power, Q is the noise power at each receiving branch, and $\boldsymbol{H}_{n0}$ is the $M \times 1$ matrix of channel transfer functions between the nth transmit antenna and all the receive antennas. The $\{\boldsymbol{H}_{n0}\}$ are distributed as independent complex Gaussians, described below. The summation of vectors can be replaced by an $M \times N$ matrix $\boldsymbol{H}$ which contains all the scalar transfer functions between the N transmit antennas and M receive antennas. In the context of the statistical model of Section 9.4, the channels between multiple users and all the receiving elements are contained in a matrix $\boldsymbol{c}$ (see Figure 9.4.1). Here, the $\boldsymbol{H}$ contain the same antenna element-to-element transfer functions as $\boldsymbol{c}$, but the $\boldsymbol{H}$ are from the transmit array elements whereas the $\boldsymbol{c}$ are from the multiple users or interferers.

The above formula is for the unknown channel case, and has equal division of the transmit power between the N transmit antennas. The maximum capacity is therefore greater than this, corresponding to the case of known channels and using the water-filling algorithm to allocate the transmitted power levels between the eigenchannels; see Section 9.3.5.

Interpreting Figure 9.4.1 in Section 9.4 as a point-to-point (but typically non-line-of-sight) link, the $\boldsymbol{x}$ can be considered as transmitting antenna branches with block coded signals, instead of a wanted signal plus interferers. (In the MIMO model case the transmission matrix $\boldsymbol{c}$ of Figure 9.4.1 corresponds to $\boldsymbol{H}$ in the Foschini capacity formula.) In a communications system, the codes would be applied to the multiplexed data streams from the transmission data source and transmitted through their individual antenna elements.

In terms of measuring or simulating a real-world link and applying the capacity formula, the scaling of the transfer functions can be as follows. Denote the individual complex transfer functions between the nth transmit antenna branch and the mth receive branch as $h(m, n)$. For example, for perfectly matched and co-polarised antenna elements, the Friis formula, for single path propagation, is $|h(m, n)|^2 = G_m G_n / L_{mn}$, where the element gains are G_m, G_n from the transmit and receive arrays respectively, and L_{mn} is the path loss between the elements; see Appendix A. In the multipath case, which is the normal situation, there must be a summation of Friis equations over the multiple paths, which results in the $(\boldsymbol{H}_{n0})_m$.

9.5.2 *Statistical model including correlated, Rician channels with different SNRs*

The $\{\boldsymbol{H}_{n0}\}$ of Foschini and Gans are taken as independent complex Gaussians, with vectors containing the means, $\boldsymbol{\mu}_0$, and covariance matrix, $\boldsymbol{V}_0$ denoted $CN_M(\boldsymbol{\mu}_0, \boldsymbol{V}_0)$. For Rayleigh channels, the means are all zero, $\boldsymbol{\mu}_0 = \boldsymbol{0}$. If the channels are uncorrelated and of equal gain, then the covariance matrix is diagonal, $\boldsymbol{V}_0 = \boldsymbol{I}_M$. When all the channels are the same (correlated), corresponding to $(F_{\text{cor}}, F_{\text{pin}}) \ll (1, 1)$, or a 'pinhole' channel (Figure 9.3.11), then $\boldsymbol{V}$ contains all ones.

To cater for Rician, correlated channels with different signal-to-noise ratios, the capacity formula can be written

$$C = \log_2 \left| \boldsymbol{I}_M + \frac{1}{N} \sum_{n=1}^{N} P_n \boldsymbol{H}_n \boldsymbol{H}_n^H \right| \tag{9.5.2}$$

where independent noise levels, Q_n, are introduced into the channels in the form $H_n = \boldsymbol{D}_Q^{-1/2} \boldsymbol{H}_{n0}$ with $\boldsymbol{D}_Q = \text{diag}\{Q_1, Q_2, \ldots, Q_N\}$, and the transmit power for each element is $P_n = P_{\text{T}}/N$ where P_{T} is recalled as the total transmitted power. This means that the channels (with noise included), $\{\boldsymbol{H}_n\}$, are now modelled with independent complex Gaussians, written $C\mathcal{N}_M(\boldsymbol{\mu}, \boldsymbol{V})$ where the means are in the M-vector $\boldsymbol{\mu} = \boldsymbol{D}_Q^{-1/2} \boldsymbol{\mu}_0$ and the covariances are in the $M \times M$ matrix $\boldsymbol{V} = \boldsymbol{D}_Q^{-1/2} \boldsymbol{V}_0 \boldsymbol{D}_Q^{-1/2}$.

9.5.3 *Method for computing the approximate cdf of capacity*

The distribution of C can be approximated by transforming it to a smooth function of the weighted sample mean (the second term in the determinant of the capacity formula), and applying Cornish-Fisher expansions (Abramowitz and Stegun, 1972, p.935) for the distribution and percentiles of C. The results are summarised here so that calculations can be undertaken; further detail of the series and its application is given in Withers (1984) and Withers and Vaughan (2001).

The smooth function is

$$C_0(\hat{\boldsymbol{w}}) = \log |\boldsymbol{I}_M + \hat{\boldsymbol{w}}| \tag{9.5.3}$$

where the natural log is used for convenience and the weighted mean is the transmitted power from each of the N transmitting elements applied the transfer functions to the

M receive elements,

$$\hat{\boldsymbol{w}} = N^{-1} \sum_{n=1}^{N} P_n \boldsymbol{H}_n \boldsymbol{H}_n^H. \tag{9.5.4}$$

The log of the characteristic function of C_0 is expressed as a series with the cumulants κ_r as (Abramowitz and Stegun, 1972, p.928)

$$\log \phi(t) = \log(E\{e^{\mathrm{j}tC_0}\}) = \sum_{r=1}^{\infty} \kappa_r(C_0) \frac{(\mathrm{j}t)^r}{r!} \tag{9.5.5}$$

The mean is the first ($r = 1$) cumulant and the variance is the second cumulant. Under certain conditions, the cumulants can be expressed as a series in a parameter N,

$$\kappa_r(C_0) = \sum_{i=r-1}^{\infty} N^{-i} a_{ri}, \quad r > 1, \tag{9.5.6}$$

where the a_{ri} are called the cumulant coefficients. In this case, a random variable Y_N is distributed as a standardised normal, for N approaching infinity, when (Withers, 1984),

$$Y_N = \sqrt{\frac{N}{a_{21}}}(C_0 - a_{10}) \quad (\rightarrow N(0, 1) \text{ as } N \rightarrow \infty). \tag{9.5.7}$$

The cumulative density of Y_N is expressed as an expansion below. This first order approximation (in equation 9.5.7) shows how the capacity is approximately Gaussian (as do the higher order approximations). The capacity seems not to be Gaussian in the limiting case of fixed (M/N), with both $M, N \rightarrow \infty$, although here it is well approximated by a Gaussian distribution (Smith and Shafi, 2002).

The expectation of the channel transfer functions from the statistical model is written

$$E\{\boldsymbol{H}_n \boldsymbol{H}_n^H\} = \boldsymbol{V} + \boldsymbol{\mu}\boldsymbol{\mu}^H \tag{9.5.8}$$

so

$$\hat{\boldsymbol{w}} = P_{\mathrm{T}} E\{\boldsymbol{H}_n \boldsymbol{H}_n^H\} = P_{\mathrm{T}}(\boldsymbol{V} + \boldsymbol{\mu}\boldsymbol{\mu}^H) \tag{9.5.9}$$

and the asymptotic mean of the capacity is

$$c = \log_2 |\boldsymbol{I}_M + P_{\mathrm{T}} E\{\boldsymbol{H}_n \boldsymbol{H}_n^H\}| = \log_2 |\boldsymbol{I}_M + P_{\mathrm{T}}(\boldsymbol{V} + \boldsymbol{\mu}\boldsymbol{\mu}^H)|. \tag{9.5.10}$$

This shows that the asymptotic mean corresponds to the capacity using the mean of the *SNR*. This is larger than the mean capacity for finite N.

As the number of transmitters, N, becomes asymptotically large, the distribution of the capacity approaches Gaussian as noted above:

$$\begin{aligned} C &= c + \sqrt{\frac{\nu}{N}} Y_N \\ &= a_{10} + \sqrt{\frac{a_{21}}{N}} Y_N \end{aligned} \tag{9.5.11}$$

for

$$\begin{aligned} c &= \log_2 |\boldsymbol{I}_M + P_{\mathrm{T}}\boldsymbol{V}| + \log_2(1 + P_{\mathrm{T}}\boldsymbol{\mu}^H(\boldsymbol{I}_M + P_{\mathrm{T}}\boldsymbol{V})^{-1}\boldsymbol{\mu}); \\ \nu &= a_{21}/(\log 2)^2; \\ a_{21} &= (tr((\boldsymbol{I}_M + \boldsymbol{w})^{-1}\boldsymbol{V})^2 + 2\boldsymbol{\mu}^H(\boldsymbol{I}_M + \boldsymbol{w})^{-1}\boldsymbol{V}(\boldsymbol{I}_M + \boldsymbol{w})^{-1}\boldsymbol{\mu})P_{2.N}; \\ a_{10} &= c\log 2. \end{aligned} \tag{9.5.12}$$

$P_{2.N}$ is given below. The first term in c corresponds to the Rayleigh channel and the second term corresponds to a non-zero mean component, and the sum is the Rician capacity. Here, the Rician channel provides a greater capacity for a given number of antennas. This is because the non-zero mean term, or specular component of the Rician channel, is acting to increase the *SNR*. For a fixed *SNR* (which includes both the diffuse and specular components), the Rayleigh channel has a greater capacity than the Rician, which in turn is greater than the purely line-of-sight, for this form of MIMO system.

In summary, we set $C_0 = C\log 2$ and this is distributed asymptotically ($N \to \infty$) as $\mathcal{N}(a_{10}, a_{21}/N)$. The capacity is then $C = C_0(\log 2)^{-1}$ and this is distributed as $\mathcal{N}(c, \nu/N)$ for $c = a_{10}/\log 2$ and $\nu = a_{21}/(\log 2)^2$. c is called the asymptotic mean capacity and ν/N is the asymptotic variance.

9.5.4 Asymptotic expansion for cdf

The transform of the capacity from a zero mean, normalised variance (standardised) variable, Y_N, is $C = c + (\nu/N)^{1/2}Y_N$. An expansion approximates the cumulative probability of Y_N which can then transform to an approximation for the probability of the capacity C.

The standardised Gaussian density, $\mathcal{N}(0, 1)$, is $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ with the cdf given by the standard normal function $\Phi(x)$; see Appendix C. For example, if $x = -2.326$ then $\Phi(x) = 0.01$, corresponding to the one-percentile. Denoting the cdf of Y_N

$$\mathrm{Prob}_{YN}(x) = \mathrm{Prob}(Y_N \leq x) \tag{9.5.13}$$

then asymptotic expansions, called Edgeworth expansions (cf. Abramowitz and Stegun, 1972, p.935) can be written

$$\mathrm{Prob}_{YN}(x) = \Phi(x) - \phi(x)\sum_{r=1}^{\infty} N^{-r/2}h_r(x); \tag{9.5.14}$$

and the inverse is

$$\mathrm{Prob}_{YN}^{-1}(\Phi(x)) = x + \sum_{r=1}^{\infty} N^{-r/2}g_r(x) \tag{9.5.15}$$

where h_r and g_r are polynomials in x and standardised cumulant coefficients (see below) given in Withers (1984).

The infinite sum is replaced with a sum to R terms. For $R = 1$ (first order approximation) the sum term is zero and the central limit theorem applies, giving $\text{Prob}_{YN}(x) \approx \Phi(x)$. For example, again using $x = -2.326$ with $\Phi(x) = 0.01$, the one-percentile of Y_N is $\text{Prob}_{YN}^{-1}(\Phi(x))$, and the one-percentile of C_0 is, for the first order approximation,

$$\text{Prob}_{C_0}^{-1}(0.01) = a_{10} + \sqrt{\frac{a_{21}}{N}}\text{Prob}_{YN}^{-1}(0.01) \tag{9.5.16}$$

where a_{10} and a_{21} are given below.

9.5.5 Asymptotic capacity and variance for Rayleigh channels

9.5.5.1 Rayleigh uncorrelated case

For Rayleigh fading and uncorrelated antennas, we have $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{V}_0 = \boldsymbol{I}_M$, and $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{V} = \text{diag}(\boldsymbol{Q}_m^{-1})$, giving the asymptotic ($N \to \infty$) capacity

$$c = \sum_{m=1}^{M} \log_2\left(1 + \frac{P_\text{T}}{Q_m}\right) \tag{9.5.17}$$

and variance

$$v = \frac{P_{2\cdot N}}{(\log 2)^2} \sum_{m=1}^{M} (P_{1\cdot N} + Q_m)^{-2}, \tag{9.5.18}$$

where

$$P_{r\cdot N} = \frac{1}{N} \sum_{n=1}^{N} P_n^r. \tag{9.5.19}$$

For the case of common signal powers and noise powers, so that the signal-to-noise ratio $SNR = P/Q$ is common to all the branches (cf. Section 9.3.5), then

$$c = M \log_2(1 + SNR), \quad v = \frac{M}{((\log 2)(1 + SNR^{-1}))^2}. \tag{9.5.20}$$

These equations show how adding more receivers (larger M) has an additive effect on the asymptotic mean and variance of the capacity.

9.5.5.2 Rayleigh, correlated case

For Rayleigh fading and fully correlated antennas, we have $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{V}_0 = \mathbf{1}_M \mathbf{1}_M^H$, ($\mathbf{1}_M$ is a vector of M ones), and the formulation results in

$$c = \log_2\left(1 + P_\text{T} \sum_{m=1}^{M} Q_m^{-1}\right) \tag{9.5.21}$$

and variance

$$\nu = \frac{P_{2\cdot N}}{(\log 2)^2}\left(P_{1.N} + \frac{1}{\sum_{m=1}^{M} Q_m^{-1}}\right)^{-2} \tag{9.5.22}$$

and for common *SNRs*,

$$c = \log_2(1 + M\ SNR), \quad \nu = \frac{1}{((\log 2)(1 + M^{-1}SNR^{-1}))^2}. \tag{9.5.23}$$

So in the fully correlated case the capacity and variance behave as in the uncorrelated case but with a single receiving element, i.e. $M = 1$, and with the *SNR* multiplied by M, and this is consistent with line-of-sight array considerations.

9.5.6 *Coefficients for capacity cdf with percentile outage for Rayleigh, uncorrelated channels*

The coefficients are given here for the second and third order expansions for the Rayleigh, uncorrelated case. Recall that the 95-percentile of the capacity, for example, is

$$\text{Prob}_C^{-1}(0.95) = c + \sqrt{\frac{\nu}{N}}\text{Prob}_{YN}^{-1}(0.95). \tag{9.5.24}$$

For general percentiles,

$$\begin{aligned} Y_N &= x, && (R = 1) \\ &= x + N^{-1/2}g_1(x), && (R = 2) \\ &= x + N^{-1/2}g_1(x) + N^{-1}g_2(x), && (R = 3) \end{aligned} \tag{9.5.25}$$

where

$$g_1(x) = A_{11} + A_{32}\frac{(x^2 - 1)}{6}; \tag{9.5.26}$$

and for the third order approximation,

$$g_2(x) = A_{22}\frac{x}{2} + A_{43}\frac{(x^3 - 3x)}{24} + A_{32}^2\frac{(2x^3 - 5x)}{36}; \tag{9.5.27}$$

and the standardised cumulant coefficient is

$$A_{ri} = \frac{a_{ri}}{a_{21}^{r/2}}. \tag{9.5.28}$$

With $\nu = 1 + SNR$, and with

$$\nu_r = \left(\frac{P}{Q\nu}\right)^r = \left(\frac{1}{1 + SNR^{-1}}\right)^r, \tag{9.5.29}$$

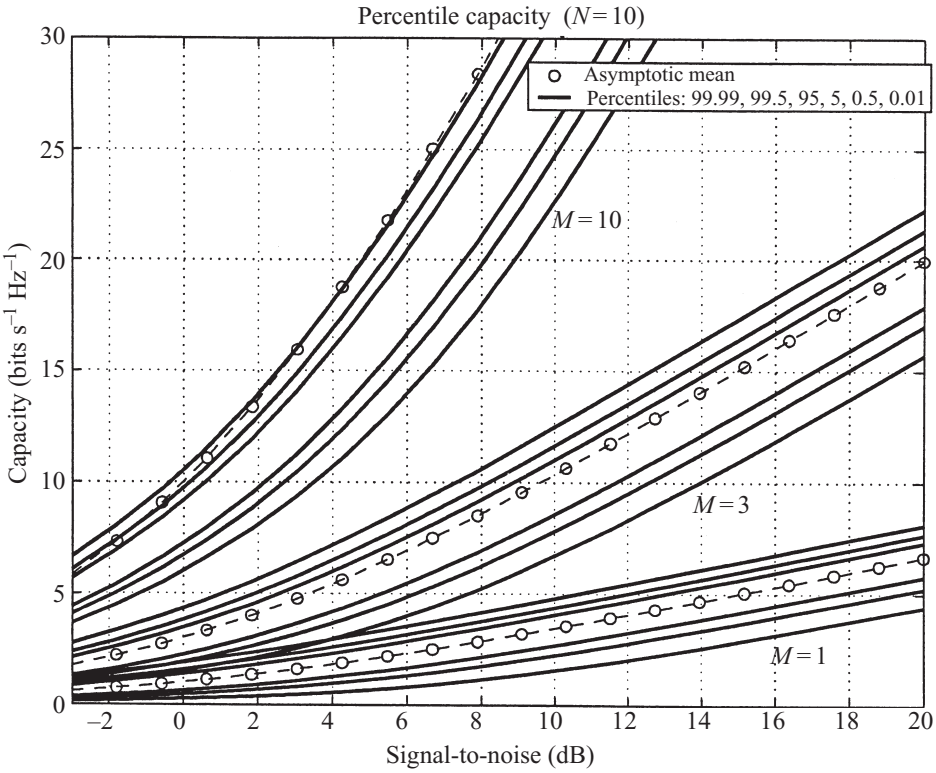


Figure 9.5.1 *The asymptotic mean and percentile capacity for the case of fixed total transmit power. There are $N = 10$ transmit antennas and $M = 1$, 3 and 10 receive antennas. The asymptotic mean capacity (dotted) is inaccurate unless $N \gg M$, although the percentiles for outage are well behaved in this example*

the cumulant coefficients are

$$\begin{aligned} a_{10} &= M \log v, \\ a_{21} &= M\nu_2, \\ a_{11} &= -M^2\nu_2/2, \\ a_{32} &= M\left(2\nu_3 - 3\nu_2^2\right), \\ a_{22} &= M^2\left(5\nu_2^2/2 - 2\nu_3\right), \\ a_{43} &= M\left(20\nu_2^3 - 24\nu_2\nu_3 + 6\nu_4\right). \end{aligned} \tag{9.5.30}$$

9.5.6.1 Example

An example calculation for the one-percentile for $N = 10$ transmitting antennas and $M = 1$, 3 and 10 receiving antennas indicates that the second and third order approximations are essentially the same, i.e. the second order approximation is sufficiently

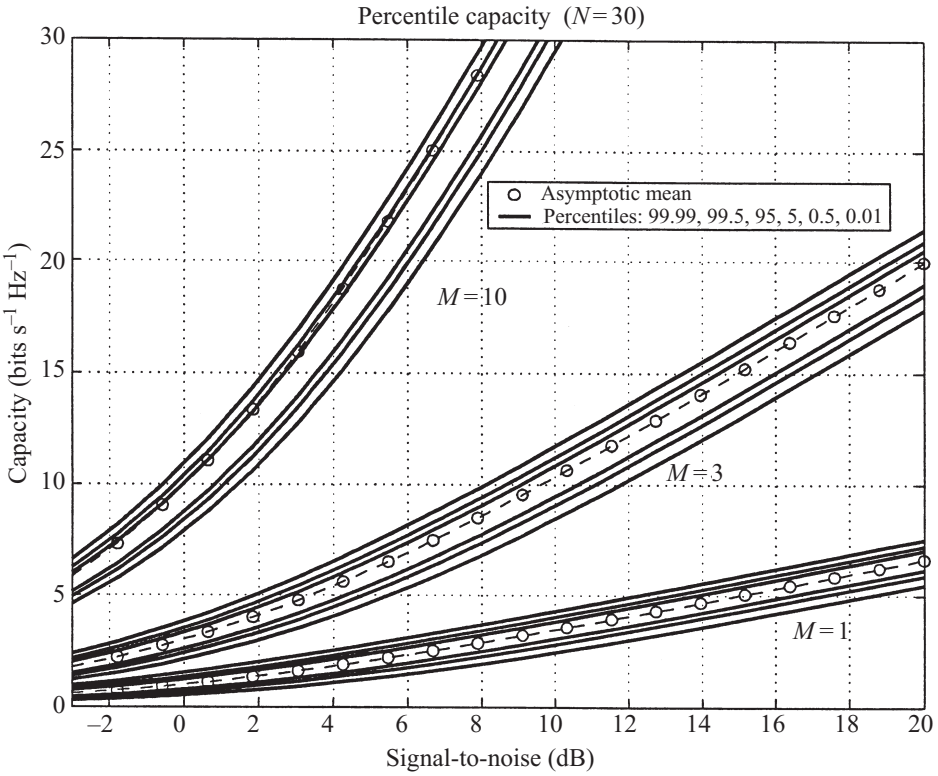


Figure 9.5.2 The asymptotic mean and percentile capacity for the case of transmit power fixed for each transmit antenna, for $N = 30$, which reduces the variance relative to the $N = 10$ case

accurate for this example. Using the third order approximation, the asymptotic mean and percentiles for 10 transmit elements and 30 transmit elements are given respectively in Figures 9.5.1 and 9.5.2 using $x = \{3.891, 2.576, 1.6449, -1.6449, \ldots\}$ for percentiles $\{99.99, 99.5, 95, 5, 0.5, 0.01\}$. These curves indicate the sensitivity to the percentile value of the outage, in terms of the signal-to-noise ratio. Note the sharply decreased variance with the larger number of transmit antennas.

9.5.6.2 Mean capacity expression

In Figures 9.5.1 and 9.5.2 the asymptotic mean capacity is not accurate and sits too high relative to the percentiles, although the percentiles themselves seem well behaved after convergence. For practicable finite antenna array sizes, a better approximation is given by Rapajic and Popescu (2001). This approximation is known to be very accurate (Smith and Shafi, 2002). The approximation can be written, for $y = M/N$, $M \leq N$; or $y = N/M$, $N \leq M$; as

$$c \approx \min(N, M) \frac{1}{\log 2} \left\{ \log(w_y SNR) + \left(\frac{1-y}{y} \right) \log \left(\frac{1}{1-v_y} \right) - \frac{v_y}{y} \right\} \quad (9.5.31)$$

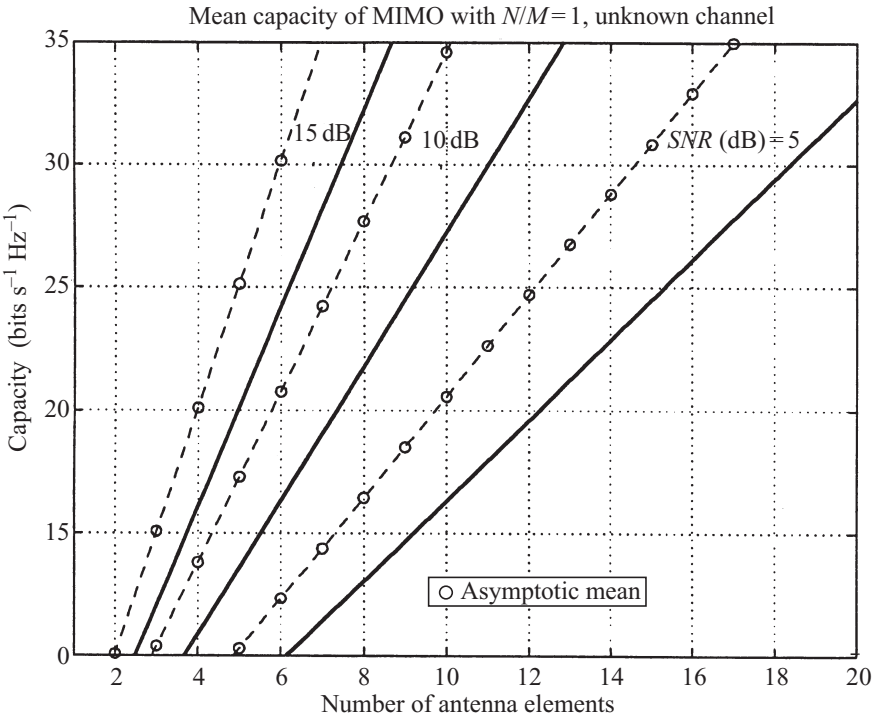


Figure 9.5.3 The mean capacity for equal numbers of transmit and receive antennas, and the corresponding value for the asymptotic mean, which is inaccurate for small N

where

$$w_y = \frac{1}{2}\left(\left(1 + y + \frac{1}{SNR}\right) + \sqrt{\left(1 + y + \frac{1}{SNR}\right)^2 - 4y}\right) \tag{9.5.32}$$

$$v_y = \frac{1}{2}\left(\left(1 + y + \frac{1}{SNR}\right) - \sqrt{\left(1 + y + \frac{1}{SNR}\right)^2 - 4y}\right) \tag{9.5.33}$$

An example calculation for $N = M = 10$ shows that the approximations become the same when N/M is about 40. Some values are given in Figure 9.5.3, where the asymptotic capacity is compared with the accurate approximation of Rapajic and Popescu. Note that this accurate mean value cannot be used in place of the asymptotic mean in the formulas for calculating the percentiles.

As a final note, these capacities are theoretical limits, and it is not obvious how closely the limits can be approached in an economical, real-world system. Extensive signal processing is required, and the implementation of compact, uncorrelated antennas is necessary.

9.6 References

ABRAMOWITZ, M. and STEGUN, I.A.: *Handbook of Mathematical Functions*, 9th edition, Dover, 1972.

ALAMOUTI, S.: 'A simple transmitter diversity technique for wireless communications', *IEEE Journal on Selected Areas in Communications, Special Issue on Signal Processing in Wireless Communications*, 1998, **16** (8), pp. 1451–1458.

ANDERSEN, J. BACH: 'Antenna arrays in mobile communications', *IEEE Antennas and Propagation Magazine*, April 2000, **42** (2), pp. 12–16.

ANDERSEN, J.BACH: 'Array gain and capacity for known random channels with multiple element arrays at both ends', *IEEE Journal on Selected Areas in Communications*, November 2000, **18** (11), pp. 2172–2178.

CHIZHIK, D., FOSCHINI, G.J., and VALENZUELA, R.A.: 'Capacities of multi-element transmit and receive antennas: correlations and keyholes', *Electronics Letters*, 22 June 2000, **36** (13), pp. 1099–1100.

FOSCHINI, G.J.: 'Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas', *Bell Laboratories Technical Journal*, Autumn 1996, pp. 41–59.

FOSCHINI, G.J. and GANS, M.J.: 'On the limits of wireless communications in a fading environment when using multiple antennas', *Wireless Personal Communications*, (6), Kluwer, 1998, pp. 311–335.

FOSCHINI, G.J. and SALZ, J.: 'Digital communications over fading radio channels', *Bell System Technical Journal*, February 1983, **62**, pp. 429–456.

GALLAGER, R.G.: *Information Theory and Reliable Communications*, Wiley, 1968.

GAO, H., SMITH, P.J., and CLARK, M.V.: 'Theoretical reliability of MMSE linear diversity combining in Rayleigh-fading additive interference channels', *IEEE Transactions*, May 1998, **Com-46** (5), pp. 666–673.

GEMAN, S.: 'A limit theorem for the norm of random matrices', *Annals of Probability*, 1980, **8**, pp. 252–261.

GESBERT, D., BOLCSKEI, H., GORE, D., and PAULRAJ, A.: 'MIMO wireless channels: capacity and performance prediction', *Proceedings of the Global Telecommunications Conference*, *2000*, (GLOBECOM '00), 2000, **2**, pp. 1083–1088.

HAAGERUP, U. and THORBJØRNSSON, S.: 'Random matrices and K-theory for exact C^*-algebras', Centre for Mathematical Physics and Stochastics, Aarhus University, Denmark, Research Report (14), September 1998.

HAYKIN, S.: *Adaptive Filter Theory*, Prentice Hall, 1991.

JAKES, W.C. (Ed.):, *Microwave Mobile Communications*, AT&T Press, 1974; also IEEE Press, 1996.

JAMES, A.T.: 'Distributions of matrix variates and latent roots derived from normal samples', *Annals of Mathematical Statistics*, 1964, **35**, pp. 475–501.

MONZINGO, R.A. and MILLER, T.W.: *Introduction to Adaptive Arrays*, Wiley, New York, 1980.

RAPAJIC, P. B. and POPESCU, D.: 'Information capacity of a random signature multiple-input, multiple-output channel', *IEEE Transactions on Communications*, August 2001, **48** (8), pp. 1245–1248.

SCHARF, L.L.: *Statistical Signal Processing*, Addison-Wesley, 1991.

SILVERSTEIN, J.W.: 'The smallest eigenvalue of a large dimensional Wishart matrix', *Annals of Probability*, 1985, **13**, pp. 1364–1368.

SMITH, P.J.: 'Exact performance analysis of optimum combining with multiple interferers in flat Rayleigh fading', *IEEE Transactions on Communications*, 2002.

SMITH, P.J. and SHAFI, M.: 'On a Gaussian approximation to the capacity of wireless MIMO systems', *IEEE International Conference on Communications*, (ICC 2002), New York, April/May 2002.

TAROKH, V., SESHADRI, N., and CALDERBANK, A.R.: 'Space-time codes for high data rate wireless communication: performance criterion and code construction', *IEEE Transactions*, March 1998, **IT-44** (2), pp. 744–765.

TELATAR, I.E.: 'Capacity of multi-antenna Gaussian channels', Technical Note, AT&T Bell Laboratories, 1996.

VAUGHAN, R.G.: 'On optimum combining at the mobile', *IEEE Transactions*, November 1988, **VT-37** (4), pp. 181–188.

WINTERS, J.H.: 'On the capacity of radio communications systems with diversity in a Rayleigh fading environment', *IEEE Journal of Selected Areas on Communications*, June 1987, **SAC-5** (5), pp. 871–878.

WINTERS, J.H.: 'The diversity gain of transmit diversity in wireless systems with Rayleigh fading', *IEEE Transactions on Vehicular Technology*, February 1998, **47** (1), pp. 119–123.

WINTERS, J.H. and SALZ, J.: 'Upper bounds on the bit error rate of optimum combining in wireless systems', *Proceedings of the IEEE Vehicular Technology Conference*, Stockholm, Sweden, June 1994, pp. 942–946.

WINTERS, J.H., SALZ, J., and GITLIN, R.D.: 'The impact of antenna diversity on the capacity of wireless communications systems', *IEEE Transactions*, February/March/April 1994, **Com-42** (2/3/4), pp. 1740–1750.

WITHERS, C.S.: 'Asymptotic expansions for distributions and quantiles with power series cumulants', *Journal of the Royal Statistical Society*, B, 1984, (46), pp. 389–396.

WITHERS, C.S. and VAUGHAN, R.G.: 'Distribution of signal-to-interference-plus-noise ratio for optimum combining', *Proceedings of the 10th Virginia Tech/MPRG Symposium on Wireless Communications*, June 2000, pp. 483–494.

WITHERS, C.S. and VAUGHAN, R.G.: 'Distribution and percentiles of channel capacity for multiple arrays', *Proceedings of the 11th Virginia Tech/MPRG Symposium on Wireless Communications*, June 2001, pp. 141–152.

Appendix A

Field strength and path loss

In general, the path loss, as a deterministic parameter, is defined in the following way. A transmitter transfers P_T (W) into the port of an antenna with gain G_T, and the perfectly co-polarised receiving antenna with gain G_R delivers power P_R (W) into a perfectly matched load at its port. The path loss L is defined from the Friis transmission equation, which reduces to

$$\frac{P_R}{P_T} = G_T G_R \frac{1}{L}. \tag{A.1}$$

It is clear that the 'path gain' is simply the reciprocal of the path loss. A few simple equations show the important relations between path loss and field strength.

At a distance d (m) of free space from the phase centre of the transmitting antenna, the power density S is

$$S = \frac{P_T G_T}{4\pi d^2} \text{ W m}^{-2}. \tag{A.2}$$

Sometimes the field strength at the receiving antenna, E_R (V m^{-1}), is wanted instead of the path loss. The relation between E_R and S is

$$S = \frac{|E_R|^2}{Z_0} = \frac{|E_R|^2}{120\pi}, \tag{A.3}$$

and so

$$|E_R|^2 = 30\frac{P_T G_T}{d^2} \text{ (V m}^{-1})^2. \tag{A.4}$$

At the receiving antenna, the incident power density S is multiplied by the antenna effective cross section A_R, defined by

$$A_R = \frac{\lambda^2}{4\pi} G_R, \tag{A.5}$$

to give the received power as

$$P_R = S\frac{\lambda^2}{4\pi} G_R. \tag{A.6}$$

The power ratio can now be written

$$\frac{P_{\mathrm{R}}}{P_{\mathrm{T}}} = G_{\mathrm{T}} G_{\mathrm{R}} \left(\frac{\lambda}{4\pi d} \right)^2, \tag{A.7}$$

and thus the free-space path loss is given by the expression

$$L_0 = \left(\frac{4\pi d}{\lambda} \right)^2 = \left(\frac{4\pi f d}{c} \right)^2. \tag{A.8}$$

The free-space path loss is also well known in its logarithmic form, where the frequency is expressed in MHz and the distance is in km:

$$L_0(\mathrm{dB}) = 10 \log_{10}(L_0) = 32.4 + 20 \log_{10} f_{\mathrm{MHz}} + 20 \log_{10} d_{\mathrm{km}}. \tag{A.9}$$

Note that for frequency-independent antenna gains, the path loss is proportional to the frequency squared, in contrast to the field strength which is independent of frequency.

Substituting equation (A.8) into equation (A.4), the relation between path loss and field strength is

$$|E_{\mathrm{R}}|^2 = 30 \left(\frac{4\pi}{\lambda} \right)^2 P_{\mathrm{T}} G_{\mathrm{T}} \frac{1}{L} = 30 \left(\frac{4\pi f}{c} \right)^2 P_{\mathrm{T}} G_{\mathrm{T}} \frac{1}{L}. \tag{A.10}$$

Note from equation (A.1) that the path loss is a well-defined quantity and its use is not restricted to the inverse square law of equation (A.9) corresponding to free space.

Appendix B

Basic statistics for mobile communications

B.1 Probability and statistical independence

This appendix reviews some basic statistics which are commonly used in mobile communications and signal processing, and is associated with Chapter 2. The results are well known and are drawn from several sources such as the classical text by Papoulis (1965) and more recent texts such as Kay (1993).

The probability of the real random variable z being less than or equal to a value Z, is written

$$\mathrm{P}(z \leq Z) = \int_{-\infty}^{Z} p_z(x)\,\mathrm{d}x, \tag{B.1}$$

where P_Z is the cumulative density function (cdf), also called the cumulative probability function (cpf), and $p_z(z)$ is the probability density function (pdf); k random variables are statistically *independent* if their joint pdf, $p_z(\boldsymbol{x})$, factors into the product of the individual random variable pdfs:

$$p_z(\boldsymbol{x}) = p_{z_1}(x_1)p_{z_2}(x_2)\cdots p_{z_k}(x_k) \tag{B.2}$$

where $\boldsymbol{z}$ and $\boldsymbol{x}$ are the set of z_i and x_i, respectively.

B.2 Probability density function: fundamental theorem and transformation

If the pdf of z, $p_z(z)$, is known, then the pdf of a function $g(z)$ is found (Papoulis, 1965) from

$$p_g(g) = \sum_n \frac{p_z(z^{(n)})}{|g'(z^{(n)})|} \tag{B.3}$$

where the prime denotes differentiation with respect to the argument and $z^{(n)}$ are the real roots of $g(z)$.

A change in variables, say from $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)$ to $\boldsymbol{w} = (w_1, w_2, \ldots, w_p) = \boldsymbol{w}(\boldsymbol{z})$, for a pdf, is found from the function transformation theorem,

$$p_w(\boldsymbol{w}) = p_w(w_1, w_2, \ldots, w_p) = \frac{p_z(z_1, z_2, \ldots, z_p)}{|\boldsymbol{J}|} \tag{B.4}$$

where the Jacobian is

$$|\boldsymbol{J}| = \det(\boldsymbol{w}'(\boldsymbol{z})) = \begin{vmatrix} \dfrac{\partial w_1}{\partial z_1} & \dfrac{\partial w_2}{\partial z_1} & . \\ \dfrac{\partial w_1}{\partial z_2} & \dfrac{\partial w_2}{\partial z_2} & . \\ . & . & . \end{vmatrix} = \begin{vmatrix} \dfrac{\partial w_1}{\partial z_1} & \dfrac{\partial w_1}{\partial z_2} & . \\ \dfrac{\partial w_2}{\partial z_1} & \dfrac{\partial w_2}{\partial z_2} & . \\ . & . & . \end{vmatrix}. \tag{B.5}$$

B.3 Expectation and covariance

The expectation is also called the *statistical average* and is denoted using $E\{\cdot\}$ although both the overbar $\bar{\cdot}$ and the angular brackets $\langle\cdot\rangle$ are also used. In mobile communications, the angle brackets and overbar are normally used for indicating a time or *sample average* (see Section 2.2.1.2), but the statistical average shares this notation according to context and historical use.

The expected value of a function $g(\boldsymbol{x})$ is defined from the joint pdf as

$$E\{g(\boldsymbol{x})\} = \int_{-\infty}^{\infty} g(\boldsymbol{x}) p_z(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{B.6}$$

If $\boldsymbol{x}$ is complex, then the integral is for both the real and imaginary parts going from minus to plus infinity.

The covariance is a measure of the degree of correlation between two complex numbers around their mean values,

$$C_{z_i z_j} = E\{(z_i - m_{z_i})(z_j - m_{z_j})^*\} \tag{B.7}$$

where $m_z = E\{z\}$ denotes the mean of z and the $*$ superscript is the complex conjugate. The covariance of complex random variables is complex in general. A single subscript $C_z = C_{zz}$ is used to denote the variance of z, often written as σ_z^2.

The normalised covariance is the *correlation coefficient*,

$$\rho_{z_i z_j} = \frac{C_{z_i z_j}}{\sigma_{z_i} \sigma_{z_j}}, \tag{B.8}$$

which is complex, and in which the standard deviation is the real quantity

$$\sigma_{z_i} = \sqrt{E\{|z_i - m_{z_i}|^2\}}. \tag{B.9}$$

For a general stochastic process $z(t)$, the *autocovariance function* is defined as

$$\begin{aligned} C_z(t_1, t_2) &= E\{(z(t_1) - m_z(t_1))(z(t_2) - m_z(t_2))^*\} \\ &= E\{z(t_1)z^*(t_2)\} - m_z(t_1)m_z^*(t_2), \end{aligned} \tag{B.10}$$

where the time-dependent mean is properly defined through equation (B.6), which requires the time-dependent pdf of z. In practice, ergodicity in the mean is assumed in a local sense so that an estimate of $m(t)$ can be defined over some finite interval $2T$, i.e.

$$m_z(t) \approx \hat{m}_z(t) = \frac{1}{2T} \int_{t-T}^{t+T} z(t)\, \mathrm{d}t. \tag{B.11}$$

This local mean is also called the *running mean*.

Either the covariance function or the correlation function (below), comprises the basic second-order (i.e. depending on the first and second moments) quantity in random processes. Second-order statistics are needed to derive power quantities and spectra. Strictly speaking, the covariance is the fundamental statistical quantity, but the correlation is used directly in communications signal processing.

B.4 Correlation (autocorrelation) and ergodicity

Correlation is defined similarly to covariance, but without subtraction of the running means. As with the covariance, the autocorrelation is denoted with a single subscript. Thus the autocorrelation of z is written

$$R_z(t_1, t_2) = E\{z(t_1)z^*(t_2)\} \tag{B.12}$$

so that

$$R_z(t_1, t_2) = C_z(t_1, t_2) + m_z(t_1)m_z^*(t_2) \tag{B.13}$$

and the *normalised autocorrelation function* is

$$r_z(t_1, t_2) = \frac{C_z(t_1, t_2) + m_z(t_1)m_z^*(t_2)}{C_z(t_1, t_1) + |m_z(t_1)|^2}, \tag{B.14}$$

in which we can write $C_z(t_1, t_1) = \sigma_z^2(t_1)$. The autocorrelation of the centred process $z(t) - m_z(t)$ is the same as the autocovariance of $z(t)$. So for a zero-mean process, the correlation and the covariance functions are the same and the normalised correlation becomes the same as the correlation coefficient.

Cross-correlations are denoted in a similar way. The definition of moments such as the cross-correlation is, from equation (B.6),

$$R_{xy}(t_1, t_2) = E\{xy^*\} = \int\int xy^* p_{xy}(x, y)\, \mathrm{d}x\, \mathrm{d}y, \tag{B.15}$$

where $p_{xy}(x, y)$ is the joint pdf of $x(t_1)$ and $y(t_2)$. The integration is from minus infinity to plus infinity, and from here on, unlabelled integral limits are taken to be

over the full range of the integration variable. For *correlation–ergodic* processes, the cross-correlation may be expressed as

$$R_{xy}(\tau) = \lim(T \to \infty)\frac{1}{2T}\int_{-T}^{T} x(t)y^*(t+\tau)\,\mathrm{d}t. \tag{B.16}$$

The correlation coefficient function for ergodic processes can be expressed in terms of the correlation function as

$$\rho(\tau) = \frac{R_{xy}(\tau) - m_x m_y^*}{R_{xy}(0) - m_x m_y^*}. \tag{B.17}$$

Ergodicity is the property that formally allows estimates of the moments, etc., to be taken from finite samples of the random signals. In mobile communications, signals representing channels are seldom truly ergodic, but are considered ergodic over a period of interest. Ergodicity implies stationarity (see below), but not vice versa.

Uncorrelated signals or variables implies, strictly speaking, that their correlation is zero. However, in practice, uncorrelated is taken to be when the correlation coefficient is 'low', say $|\rho| \leq 0.5$. Independence implies (strictly) uncorrelated, but not vice versa unless the statistics of the variables are Gaussian.

B.5 Correlation functions and a deterministic channel

The need for correlation functions is illustrated by sending a random signal through a deterministic, time-invariant linear channel as illustrated in Figure B.1.

In the case of a deterministic input signal $x(t)$ there results a deterministic output signal $y(t)$ which is the convolution with the channel impulse response $h(\tau)$ and the input signal, i.e.

$$y(t) = x(t) * h(t) = \int x(\alpha)\,h(t-\alpha)\,\mathrm{d}\alpha \tag{B.18}$$

or in the frequency domain,

$$Y(\omega) = H(\omega)X(\omega). \tag{B.19}$$

Since the channel itself is deterministic, moments of the input and output signals such as the means also relate through the convolution, i.e.

$$m_y(t) = \int m_x(\alpha)\,h(t-\alpha)\,\mathrm{d}\alpha. \tag{B.20}$$

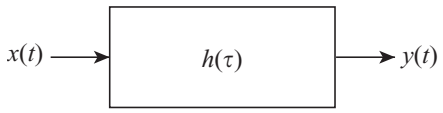


Figure B.1 Stochastic signals through a deterministic channel

This result is quickly found from equation (B.18) using an interchange of integration and expectation. Similarly, for the correlation functions, the convolution relation is

$$R_y(t_1, t_2) = \int\int R_x(\alpha, \beta)h(t_1 - \alpha)h^*(t_2 - \beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta. \tag{B.21}$$

From equations (B.10) and (B.18), identical relations hold for the covariance, i.e.

$$C_y(t_1, t_2) = C_x(t_1, t_2) * h(t_1) * h^*(t_2). \tag{B.22}$$

B.6 Wide-sense stationarity (WSS)

Wide-sense stationarity (WSS) is a particularly useful form of stationarity for stochastic processes and refers to the covariance being dependent only on the time difference and not on absolute time. In terms of the correlation, this infers that the mean is constant and the correlations also depend only on the time difference, i.e.

$$\text{WSS}: \quad m_z(t) = \text{constant}; \quad \text{and} \quad R_x(t_1, t_2) = R_x(\tau), \quad \text{where } \tau = t_2 - t_1. \tag{B.23}$$

For the linear transfer function and WSS signals, equation (B.20) shows that the mean of the output is constant if the mean of the input is constant. The correlation function becomes

$$R_y(\tau) = \int\int R_x(\tau - \alpha + \beta)h(\alpha)h^*(\beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta, \tag{B.24}$$

showing a WSS output for a WSS input.

WSS is a weaker condition than *strict-sense stationarity* (SSS), which requires that all the pdfs are invariant with time, or in terms of signals, $\{x(t_1 + t_0), x(t_2 + t_0) \ldots x(t_p + t_0)\}$ behaves like $\{x(t_1), x(t_2) \ldots x(t_p)\}$, for all t_1, t_2, t_p. However, for a Gaussian process, which is specified completely by its statistics up to second order, i.e. by its mean and autocorrelation, then WSS implies SSS.

Virtually any WSS Gaussian random process can be represented as the output of the linear filter driven at the input by white Gaussian noise (see Section B.10).

Finally, a process can be called non-stationary if any of the moments or pdfs (or any functions of interest) vary with time. It follows that a process can be non-stationary but still have some stationary properties. To clarify the properties of processes, stationarity and ergodicity are often referred to a specific function or set of functions. For example, a process may be referred to as nth order stationary (all moments up to the nth order, but not necessarily those higher than nth order, are stationary), or distribution–ergodic, with reference to a particular distribution.

B.7 Power spectrum

The *spectrum* is the Fourier transform of a function. With deterministic processes, the time signal itself normally has a Fourier transform which is called the continuous

spectrum. With stochastic processes, the correlation function is required to describe the time properties, and it is the Fourier transforms of the correlations that provide the various (power) spectra. Thus the *power spectral density*, or power spectrum, of a wide-sense stationary process with correlation function $R(\tau)$ is defined as the Fourier transform of $R(\tau)$ (Wiener–Khinchin theorem), i.e.

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau) \mathrm{e}^{-\mathrm{j}\omega\tau} \mathrm{d}\tau. \tag{B.25}$$

In the context of a deterministic signal $x(t)$, $S_x(\omega) = |X(\omega)|^2$ is used to denote the power spectrum, and the total power is expressed

$$P_x = \int_{-\infty}^{\infty} S_x(\omega)\, \mathrm{d}\omega = \int_{-\infty}^{\infty} |X(\omega)|^2\, \mathrm{d}\omega. \tag{B.26}$$

The subscript x on S and P is often dropped, unless the context requires it for clarification.

The relation between the power spectrum and the pdf of the frequency can be elucidated by considering a cissoid where the random variable is the angular frequency $\omega = 2\pi f$, namely for $x(t) = \exp(\mathrm{j}\omega t)$, which is WSS and zero-mean;

$$R_x(\tau) = E\{\mathrm{e}^{\mathrm{j}\omega\tau}\} = \frac{1}{2\pi} \int \mathrm{e}^{\mathrm{j}\omega\tau}\, p_\omega(\omega)\, \mathrm{d}\omega. \tag{B.27}$$

$S_x(\omega) = p_\omega(\omega)$, i.e. the power spectrum of $x(t)$ is equivalent to the pdf of ω.

The inverse transform of equation (B.25) can give the correlation of the output of the channel, $R_y(\tau)$, in terms of the power spectrum at the input and the transfer function (the Fourier transform $H(\omega)$ of $h(t)$ of equation (B.18)) as

$$R_y(\tau) = \frac{1}{2\pi} \int S_x(\omega) |H(\omega)|^2 \mathrm{e}^{\mathrm{j}\omega\tau} \mathrm{d}\omega \tag{B.28}$$

since

$$S_y(\omega) = S_x(\omega) |H(\omega)|^2. \tag{B.29}$$

B.8 Cross-correlations

Similar Fourier transform relations hold for cross-correlations. The correlation between input and output has the form

$$R_{xy}(\tau) = \int h^*(\alpha) R_x(\tau - \alpha)\, \mathrm{d}\alpha, \tag{B.30}$$

which is the intermediate step in equation (B.24). The corresponding (cross-) power spectral density relationship is

$$S_{xy}(\omega) = S_x(\omega) H^*(\omega) \tag{B.31}$$

with obvious properties $S_{xy}(\omega) = S^*_{yx}(\omega)$ and $R_{xy}(-\tau) = R^*_{yx}(\tau)$, and which is the intermediate step in equation (B.29).

Finally, it is instructive to consider $S(\omega_2 - \omega_1)$ as an autocorrelation of a WSS frequency domain transfer function, with $R(\tau)$ the spectrum defined using an inverse Fourier transform. So the terms *autocorrelation* and *spectrum* can be interchanged.

B.9 Covariance spectrum

The covariance spectrum of a wide-sense stationary input signal,

$$S^{\mathrm{c}}_x(\omega) = \int C_x(\tau)\mathrm{e}^{-\mathrm{j}\omega\tau}\mathrm{d}\tau, \tag{B.32}$$

yields directly the variance of the output signal,

$$\sigma^2_y = C_y(0) = \frac{1}{2\pi}\int S^{\mathrm{c}}_x(\omega)|H(\omega)|^2\mathrm{d}\omega, \tag{B.33}$$

and this has the same Fourier property as the correlation function above, cf. equation (B.22);

$$\begin{aligned} C_y(\tau) &= C_{xy}(\tau) * h(\tau) \\ &= C_x(\tau) * h(\tau) * h^*(-\tau). \end{aligned} \tag{B.34}$$

B.10 Uncorrelated noise example

An important property of a process or signal, which allows much progress towards simplified results, is that of being uncorrelated. An uncorrelated noise signal, $n(t)$, has $n(t)$ and $n(s)$ with zero covariance for all $t \neq s$. Sending $n(t)$ through a channel is written

$$R_n(\tau) = N_0\delta(\tau) \quad \Leftrightarrow \quad S_n(\omega) = N_0, \tag{B.35}$$

($\Leftrightarrow$ indicates the Fourier transform) where $N_0 = \sigma^2_n$, leading to the output relations

$$S_y(\omega) = N_0|H(\omega)|^2 \tag{B.36}$$

from equation (B.29), or

$$\begin{aligned} R_y(\tau) &= \frac{N_0}{2\pi}\int |H(\omega)|^2\cos(\omega\tau)\,\mathrm{d}\omega \\ &= \frac{N_0}{2\pi}\int h(\alpha-\tau)h^*(\alpha)\,\mathrm{d}\alpha \end{aligned} \tag{B.37}$$

from equation (B.24). Note that uncorrelated time signals are always correlated after passing through a channel (unless the channel is ideal and of infinite bandwidth). For example, a limit to the resolution of measuring equipment such as an antenna or other sensor with a transfer function $h(\tau)$ and a given input signal, is available from $R_y(\tau)$.

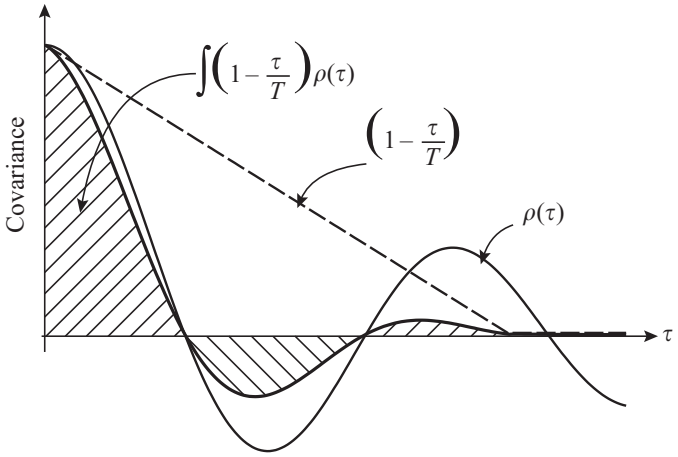


Figure B.2 The variance of the estimate of the mean using averaging over an interval $2T$ is proportional to the area under the solid line

B.11 Variance of mean estimate

The mean estimate of equation (B.11) corresponds to the impulse response in Figure B.1 being a rectangular window, of duration $2T$. Then

$$|H(\omega)|^2 = \left(\frac{\sin \omega T}{\omega T}\right)^2 \quad \Leftrightarrow \quad h(\tau) * h^*(-\tau) = 1 - \frac{|\tau|}{2T}, \tag{B.38}$$

and from equation (B.33), the variance of the mean estimate is

$$\begin{aligned}\sigma^2_{\hat{m}_y} &= \frac{1}{2T}\int_{-2T}^{2T}\left(1-\frac{|\alpha|}{2T}\right)C_x(\alpha)\,\mathrm{d}\alpha \\ &= \frac{2}{T}\int_0^T\left(1-\frac{\alpha}{T}\right)C_x(\alpha)\,\mathrm{d}\alpha,\end{aligned} \tag{B.39}$$

i.e. a triangle-weighted integral of the input covariance function. The averaging time required for a given accuracy in the estimate of the mean of a random process is found using equation (B.39). The relations are shown in Figure B.2. This result is also used as a test for ergodicity in the mean since the variance should approach zero as T approaches infinity.

B.12 Maximum likelihood and the Cramér–Rao bound for the variance of an estimate

Parameters to be estimated can usually be expressed as functions of the distribution moments (often confined to just the mean and variance).This allows estimation of the parameter using the usual population sample estimates of the moments. However,

maximum likelihood estimation offers a direct route to parameter estimation and is optimum in the sense that the estimate has a minimum variance as the number of samples becomes asymptotically large. The size of the variance for a finite sample size indicates the accuracy of the estimate and this can be given in terms of confidence intervals.

If the pdf of a random variable is known and if independent samples of the variable can be obtained, then it is possible to calculate lower bounds for the variance of any parameter estimate. One bound is called the Cramér–Rao bound, which uses maximum likelihood and this is the basic metric of accuracy for an estimate.

The parameter being estimated is denoted $\theta(x)$, and $p_x(x)$ is the pdf of the random variable x. Taking N samples of x, the joint density is

$$p_x(\boldsymbol{x}, \theta) = \prod_{n=1}^{N} p_x(x_n, \theta), \tag{B.40}$$

which is often written in the form of the conditional pdf, $p_x(x|\theta)$. If this is instead considered as a function of θ with x fixed, then it is called the likelihood function. When dealing with pdfs containing exponents, it is more convenient to deal with summations by using the log likelihood function,

$$L(\boldsymbol{x}, \theta) = \sum_{n=1}^{N} \ln p_x(x_n, \theta). \tag{B.41}$$

The likelihood function and the log likelihood function for θ have a maximum when $\theta = \theta_{\mathrm{ML}}$, and θ_{ML} is called the maximum likelihood estimate. The maximum occurs when the derivative of the log likelihood function is zero. This derivative is called Fisher's score function and can be expressed as

$$\frac{\partial L(\boldsymbol{x}, \theta)}{\partial \theta} = \frac{1}{p_x(\boldsymbol{x}, \theta)} \frac{\partial}{\partial \theta}(p_x(\boldsymbol{x}, \theta)). \tag{B.42}$$

The negative inverse of a further derivative, evaluated at the the peak (θ_{ML}), gives the curvature of the log likelihood function at its peak. When the log likelihood function is highly curved, i.e. the peak is sharp, then the variance about the peak is small. The minimised variance is the Cramér–Rao bound.

An example is using a single sample, x, from a Gaussian distribution, to estimate the mean, θ_{G}, i.e.

$$\hat{\theta}_{Mx} = x. \tag{B.43}$$

The Gaussian pdf is

$$p_{\mathrm{G}}(x, \theta_{Mx}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x - \theta_{Mx})^2}{2\sigma^2}\right\} \tag{B.44}$$

and so the score function and its derviative are respectively

$$\frac{\partial L(x, \theta_{Mx})}{\partial \theta_{Mx}} = \frac{x - \theta_{Mx}}{\sigma^2}, \quad \frac{\partial^2 L(x, \theta_{Mx})}{\partial \theta_{Mx}^2} = -\frac{1}{\sigma^2}. \tag{B.45}$$

From the fact that the estimator is just the sample itself which has a variance of σ^2, the curvature of the log likelihood function is exactly the variance of the Gaussian pdf,

$$\sigma^2_{\hat{\theta}_{Mx}} = \mathrm{Var}\{\hat{\theta}_{Mx}\} = \sigma^2 = -\frac{1}{\partial^2 L(x, \theta_{Mx})/\partial\theta^2_{Mx}}. \tag{B.46}$$

The likelihood function is itself a random variable, and so to remove the dependence of the samples, the expectation is taken over the $\boldsymbol{x}$.

The expectation of the score function is zero,

$$\begin{aligned}\left\langle \frac{\partial L(x,\theta)}{\partial\theta} \right\rangle_x &= \int \frac{\partial L(x,\theta)}{\partial\theta} p_x(x,\theta)\,\mathrm{d}x \\ &= \frac{\partial}{\partial\theta}\int p_x(x,\theta)\,\mathrm{d}x \\ &= 0.\end{aligned} \tag{B.47}$$

This is called the regularity condition and it is seen to depend on the interchangeability of the differentiation and the integration. If the regulatory condition does not hold, then the Cramér–Rao bound does not hold. This occurs rarely, for example when the support of the pdf depends on the parameter to be estimated. An example is the estimation of $g(x, \theta)$ where the samples of x are drawn from a uniform distribution $U[0, \theta]$.

The expectation over x of the second derivative gives the Fisher information, here expressed as a matrix used for estimating M parameters, $\theta = \theta_1, \theta_2, \ldots, \theta_M$,

$$\boldsymbol{J}_{\mathrm{F}}(\theta) = -\begin{bmatrix} \left\langle \frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial\theta_1^2} \right\rangle & \left\langle \frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial\theta_1\partial\theta_2} \right\rangle & \cdot & \left\langle \frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial\theta_1\partial\theta_M} \right\rangle \\ \left\langle \frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial\theta_2\partial\theta_1} \right\rangle & \left\langle \frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial\theta_2^2} \right\rangle & \cdot & \cdot \\ \cdot & \cdot & \cdot & \left\langle \frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial\theta_M^2} \right\rangle \end{bmatrix}, \tag{B.48}$$

where the ijth element is

$$\left(\boldsymbol{J}_{\mathrm{F}}(\theta)\right)_{ij} = \left\langle \frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial\theta_i\partial\theta_j} \right\rangle_x = \int \left(\frac{\partial^2 L(\boldsymbol{x},\theta)}{\partial\theta_i\partial\theta_j} \right) p_x(\boldsymbol{x},\theta)\,\mathrm{d}\boldsymbol{x}. \tag{B.49}$$

The mth parameter estimate, $\hat{\theta}_m$, has a variance given by

$$\sigma^2_{\hat{\theta}_m} = \mathrm{Var}\left\{\hat{\theta}_m\right\} = \left(\boldsymbol{J}^{-1}_{\mathrm{F}}(\theta)\right)_{mm} \tag{B.50}$$

and this is the Cramér–Rao lower bound. It has the property that it is asmyptotically Gaussian, i.e. the variable $\{\hat{\theta} - \theta\}$ is approximately Gaussian for large N, and this enables the calculation of approximate confidence intervals for the estimate.

Finally, it is worth noting that an alternative formula for the Fisher information, in terms of just the score function, is

$$J_{\mathrm{F}}(\theta) = \left\langle \left(\frac{\partial L(x,\theta)}{\partial \theta} \right)^2 \right\rangle_x = \int \frac{1}{p_x(x,\theta)} \left(\frac{\partial p_x(x,\theta)}{\partial \theta} \right)^2 \mathrm{d}x. \tag{B.51}$$

This can be shown in the steps

$$\begin{aligned} \left\langle \frac{\partial^2 L(x,\theta)}{\partial \theta^2} \right\rangle_x &= \left\langle \frac{1}{p_x(x,\theta)} \frac{\partial^2 p_x(x,\theta)}{\partial \theta^2} - \frac{1}{p_x(x,\theta)^2} \left(\frac{\partial p_x(x,\theta)}{\partial \theta} \right)^2 \right\rangle_x \\ &= \int \frac{1}{p_x(x,\theta)} \frac{\partial^2 p_x(x,\theta)}{\partial \theta^2} p_x(x,\theta)\, \mathrm{d}x \\ &\quad - \int \left(\frac{1}{p_x(x,\theta)} \right)^2 \left(\frac{\partial p_x(x,\theta)}{\partial \theta} \right)^2 p_x(x,\theta)\, \mathrm{d}x \\ &= \frac{\partial^2}{\partial \theta^2} \int p_x(x,\theta) \mathrm{d}x - \int \left(\frac{1}{p_x(x,\theta)} \right) \left(\frac{\partial p_x(x,\theta)}{\partial \theta} \right)^2 \mathrm{d}x \\ &= 0 - J_{\mathrm{F}}(\theta). \end{aligned} \tag{B.52}$$

B.13 Chi-square test for pdf

The chi-square test is standard for testing the fit of data to a pdf. Samples of the random variable are assembled into a histogram of at least five bins in the usual way. The basis for the chi-square test is as follows. For N histogram bins containing $m_1, m_2, \ldots, m_N$ sample counts, corresponding to the pdf values $e_1, e_2, \ldots, e_N$, the variable

$$\sum_{n=1}^{N} \frac{(m_n - e_n)^2}{e_n} \tag{B.53}$$

is asmyptotically chi-square with $N - 1$ degrees of freedom. For testing a pdf with assumed parameters, the above variable is compared to the chi-square (with $N - 1$ dof) value at its cumulative probability function value of $1 - \alpha$, and α is called the significance level. This value is denoted $C_{1-\alpha,N-1}$ and so the relationship

$$\sum_{n=1}^{N} \frac{\left(m_n - e_n\right)^2}{e_n} < C_{1-\alpha,N-1} \tag{B.54}$$

holds for a good fit, and its failure indicates that the histogram data do not substantiate the assumed pdf at the α significance level.

Table B.1 Critical values of D_N^α in the Kolmogorov–Smirnov test

N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5	0.51	0.56	0.67
10	0.37	0.41	0.49
15	0.30	0.34	0.40
20	0.26	0.29	0.36
25	0.24	0.27	0.32
30	0.22	0.24	0.29
35	0.20	0.23	0.27
40	0.19	0.21	0.25
45	0.18	0.20	0.24
50	0.17	0.19	0.23
>50	$1.22/\sqrt{N}$	$1.36/\sqrt{N}$	$1.63/\sqrt{N}$

B.14 Kolmogorov–Smirnoff test for cumulative density function

The Kolmogorov–Smirnoff (K–S) test is standard for testing data to a cdf. For N samples $\{x_1, x_2, \ldots, x_N\}$ a stepwise cumulative function is formed in increasing order, namely

$$S_N(x) = \begin{cases} 0, & x < x_1; \\ n/N, & x \le x < x_{n+1}; \\ 1, & x \ge x_N. \end{cases} \tag{B.55}$$

For the assumed cdf, $P(x_0) = P(x < x_0)$, the maximum departure from the stepwise cdf, i.e.

$$D_N = \max_x \left| P(x) - S_N(x) \right| \tag{B.56}$$

is a random variable depending, in general, on N. The K–S test compares D_N with tabulated (Massey, 1951; Hoel, 1962) critical values D_N^α , which occur when its cdf is $1 - \alpha$, i.e. the critical value is defined by $P\left(D_n < D_N^\alpha\right) = 1 - \alpha$. If $D_N < D_N^\alpha$, then the assumed cdf is an acceptable fit to the data for a significance of α. Otherwise, the cdf must be rejected as a good fit.

The D_N^α can be difficult to find, so values are quoted here from Hoel (1962) in Table B.1.

B.15 References

KAY, S.M.: *Fundamentals of Statistical Signal Processing*, Prentice-Hall, 1993.

HOEL, P.G.: *Introduction to Mathematical Statistics*, Wiley, New York, 1962.

MASSEY, J.F.: 'The Kolmogorov-Smirnov test for goodness-of-fit', *Journal of the American Statistical Association*, 1951, pp. 46–70.

PAPOULIS, A.: *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1965.

Appendix C

Gaussian-derived distributions in mobile communications

This appendix reviews the basic results for many of the distributions which are useful in mobile communications and is associated mainly with Chapter 2. The distributions are motivated by the need for descriptions for the electrical and field signals and the signal coverage and communications outage, etc. Gaussian statistics form the basis of most of the distributions. No physical modelling is included here so that the statistics are confined to envelope and power, and some phase distributions. The distributions of phase derivatives, for example, from physical models, are treated in the text and make use of the results here. The overbar and the angle brackets are used interchangeably with the usual expectation operator for simplicity here. Similarly, whereas in classical statistical notation, upper case letters are normally used for a random variable, this is not adhered to in this text.

C.1 Gaussian

The Gaussian distribution is the most important probability density because it models more random processes than any other distribution. This is also evident from the central limit theorem which results in Gaussian statements for the sum or average of almost any random variables. The asymptotic sum of random variables is a Gaussian quantity. Thermal noise (sum of the effect of electrons in motion) is therefore Gaussian, an example which is particularly close to the heart of the communications engineer. In mobile and other multipath communications, the signals, both wanted and unwanted, are again summations of components. In linear signal combining, used to repair the multipath-impaired channel, the resulting signals are also sums of random variables. The Gaussian distribution is therefore central to the statistical results for mobile communications.

A complex zero-mean Gaussian process

$$x + \mathrm{j}y = r\mathrm{e}^{\mathrm{j}\theta} \tag{C.1}$$

is defined here as the special case of having independent Gaussian probability density functions (pdf) for the in-phase and quadrature components x and y, respectively, written as

$$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma}\, \mathrm{e}^{-x^2/2\sigma^2}, \quad p_y(y) = \frac{1}{\sqrt{2\pi}\sigma}\, \mathrm{e}^{-y^2/2\sigma^2} \tag{C.2}$$

where σ is the standard deviation common to each component. General formulas for the moments of the real Gaussian processes are (Papoulis, 1991, p. 110):

$$E\{x^n\} = 0, \qquad n \text{ odd}, \tag{C.3a}$$

$$= (n-1)!!\,\sigma^n, \qquad n \text{ even}; \tag{C.3b}$$

$$E\{|x|^n\} = E\{x^n\}, \qquad n \text{ even}, \tag{C.3c}$$

$$= 2^{(n-1)/2}\left(\frac{n-1}{2}\right)!\,\sigma^n\sqrt{\frac{2}{\pi}}; \quad n \text{ odd}; \tag{C.3d}$$

where $(n-1)!! = 1\times3\times5\cdots(n-1) = n!\,2^{-n/2}/(n/2)!$ (for n even). (A more general notation (Abramowitz and Stegun, 1965, p. 258) is: $(2n)!! = 2\times4\times6\cdots(2n) = 2^n n!$; and $(2n-1)!! = 1\times3\times5\times\cdots(2n-1) = \pi^{-1/2}2^n\Gamma(n+1/2)$; with $0!! = 1$ and $(-1)!! = 1$.)

For example, a useful result in dealing with power quantities is

$$E\{x^4\} = 3\big(E\{x^2\}\big)^2 = 3\sigma^4, \tag{C.4}$$

which gives the variance of the energy (i.e. of the averaged power), as

$$\sigma_{x^2} = 3\sigma^4 - \sigma^4 = 2\sigma^4. \tag{C.5}$$

The averaged power in a zero mean, complex Gaussian process is

$$E\{r^2\} = \overline{r^2} = \overline{x^2} + \overline{y^2} = 2\sigma^2. \tag{C.6}$$

If the process has non-zero mean m_x, then the variable x can be centred by the replacement $x = x - m_x$, and the moments become centralised.

The cumulative distribution function for a Gaussian process is,

$$\mathrm{Prob}(x \le x_0) = \int_{-\infty}^{x_0} p_x(x)\,\mathrm{d}x \tag{C.7a}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x_0} \mathrm{e}^{-(x-m_x)^2/2\sigma^2}\mathrm{d}x \tag{C.7b}$$

$$= \Phi\left(\frac{x_0 - m_x}{\sigma}\right) \tag{C.7c}$$

and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \mathrm{e}^{-t^2/2}\mathrm{d}t \tag{C.8}$$

is the standard normal function often denoted $\Phi(x) = P(x) = 1 - Q(x)$, and $Q(x) = \Phi(-x)$ is the complementary normal function, or tail-area function,

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \mathrm{e}^{-t^2/2}\,\mathrm{d}t. \tag{C.9}$$

The standard normal function can be expressed as

$$\Phi\left(\frac{x_0 - m_x}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x_0-m_x)/\sigma} \mathrm{e}^{-t^2/2}\,\mathrm{d}t \tag{C.10a}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \mathrm{e}^{-t^2/2}\,\mathrm{d}t + \frac{1}{\sqrt{\pi}} \int_0^{(x_0-m_x)/\sqrt{2}\sigma} \mathrm{e}^{-y^2}\,\mathrm{d}y \tag{C.10b}$$

$$= \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x_0 - m_x}{\sqrt{2}\sigma}\right) \tag{C.10c}$$

$$= 1 - \frac{1}{2}\,\mathrm{erfc}\left(\frac{x_0 - m_x}{\sqrt{2}\sigma}\right), \tag{C.10d}$$

where

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2}\mathrm{d}t \tag{C.11}$$

is the error function and

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \mathrm{e}^{-t^2}\,\mathrm{d}t \tag{C.12}$$

is the complementary error function. Note $\mathrm{erf}(-x) = -\mathrm{erf}(x)$ and that the standard normal function is related to the error function by

$$\Phi\left(\sqrt{2}x\right) = \tfrac{1}{2}(\mathrm{erf}(x) + 1), \qquad \mathrm{erf}(x) = 2\Phi\left(\sqrt{2}x\right) - 1, \tag{C.13}$$

and the complementary error function relates to the Q-function as

$$\mathrm{erfc}(x) = 2Q\left(\sqrt{2}x\right), \quad Q(x) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right). \tag{C.14}$$

The area under a non-central Gaussian function within a support of $2L$, as indicated in Figure C.1, is

$$\begin{aligned} P(|z| \le L) &= \int_{-L}^{+L} \frac{1}{\sqrt{2\pi}\sigma}\,\mathrm{e}^{-(x-x_0)^2/2\sigma^2}\,\mathrm{d}x \\ &= \Phi\left(\frac{L - x_0}{\sigma}\right) - \Phi\left(\frac{-L - x_0}{\sigma}\right) \\ &= \frac{1}{2}\left(\mathrm{erf}\left(\frac{L - x_0}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{-L - x_0}{\sqrt{2}\sigma}\right)\right). \end{aligned} \tag{C.15}$$

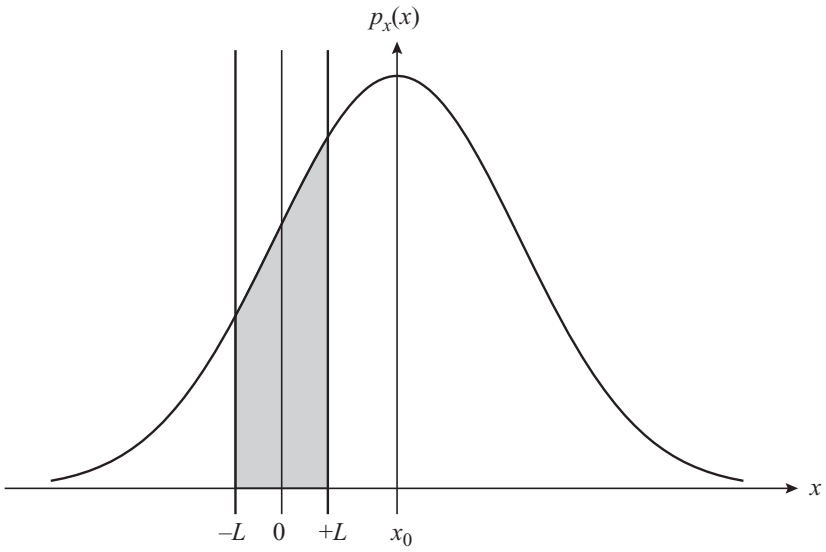


Figure C.1 The area between the limits of the non-central Gaussian pdf

When the Gaussian function is centralised on the support, the area becomes symmetrically shaped and the expression simplifies to

$$\Phi\left(\frac{L}{\sigma}\right) - \Phi\left(-\frac{L}{\sigma}\right) = \text{erf}\left(\frac{L}{\sqrt{2}\sigma}\right). \tag{C.16}$$

Another useful form for the centralised case is

$$\int_{-L}^{+L} \mathrm{e}^{-x^2/2\sigma^2}\mathrm{d}x = \sqrt{2\pi}\sigma\frac{1}{2}\left(\text{erf}\left(\frac{L}{\sqrt{2}\sigma}\right) - \text{erf}\left(-\frac{L}{\sqrt{2}\sigma}\right)\right) \tag{C.17a}$$

$$= \sqrt{2\pi}\sigma\ \text{erf}\left(\frac{L}{\sqrt{2}\sigma}\right), \tag{C.17b}$$

showing that when the variance approaches infinity, corresponding to a uniform distribution, this integral becomes $2L$, i.e.

$$\text{Lim}_{\sigma\to\infty}\left\{\sigma\,\text{erf}\left(\frac{L}{\sqrt{2}\sigma}\right)\right\} = \sqrt{\frac{2}{\pi}}L. \tag{C.18}$$

Finally, the mode (most probable value) is the same as the mean for the Gaussian distribution. For general pdfs, the mode must be found by differentiation of the pdf.

C.2 Rayleigh

The envelope r and phase θ of the complex process are related to the in-phase and quadrature components through $x = r\cos\theta$ and $y = r\sin\theta$. The Jacobian (see

Appendix B) is then

$$\left|\frac{\partial(x,\,y)}{\partial(r,\,\theta)}\right| = \begin{vmatrix} \dfrac{\partial x}{\partial r} & \dfrac{\partial x}{\partial \theta} \\ \dfrac{\partial y}{\partial r} & \dfrac{\partial y}{\partial \theta} \end{vmatrix} = r, \tag{C.19}$$

and r and θ are established as independent with Rayleigh and uniform distributions respectively, through the steps (see Appendix B)

$$\begin{aligned} p_{r,\theta}(r,\theta) &= p_{x,y}(x,\,y)\left|\frac{\partial(x,\,y)}{\partial(r,\,\theta)}\right| \\ &= \frac{2r}{\overline{r^2}}\mathrm{e}^{-r^2/\overline{r^2}}\cdot\frac{1}{2\pi} \\ &= \frac{r}{\sigma^2}\mathrm{e}^{-r^2/2\sigma^2}\cdot\frac{1}{2\pi}, \qquad r \geq 0; \\ &= p_r(r)\cdot p_\theta(\theta). \end{aligned} \tag{C.20}$$

The pdf of the phase is $1/(2\pi)$, so the mean phase is π and the standard deviation is $\pi/\sqrt{3}$.

The mean of the envelope is

$$\begin{aligned} E\{r\} = \bar{r} &= \int r p_r(r)\,\mathrm{d}r \\ &= \sqrt{\frac{\pi}{2}}\sigma, \end{aligned} \tag{C.21}$$

and with equation (C.6), $E\{r^2\} = \overline{r^2} = 2\sigma^2$, the variance follows as

$$\sigma_r^2 = 2\sigma^2 - \left(\sqrt{\frac{\pi}{2}}\sigma\right)^2 \tag{C.22a}$$

$$= \overline{r^2}\left(1 - \frac{\pi}{4}\right), \tag{C.22b}$$

so that the ratio of the envelope standard deviation to the Gaussian components standard deviation is

$$\frac{\sigma_r}{\sigma} \approx 0.655. \tag{C.23}$$

The mode of the distribution of r is readily found by differentiation to be σ.

The cumulative probability density of the envelope is

$$P(r \leq r_0) = \int_0^{r_0} p_r(r)\,\mathrm{d}r$$
$$= 1 - \mathrm{e}^{-r_0^2/\overline{r^2}} \tag{C.24}$$

and the median value r_{med}, follows from $P(r \leq r_{\text{med}}) = 0.5$, as

$$r_{\text{med}} = \sqrt{\overline{r^2}\ln 2}$$
$$\approx 1.77\sigma. \tag{C.25}$$

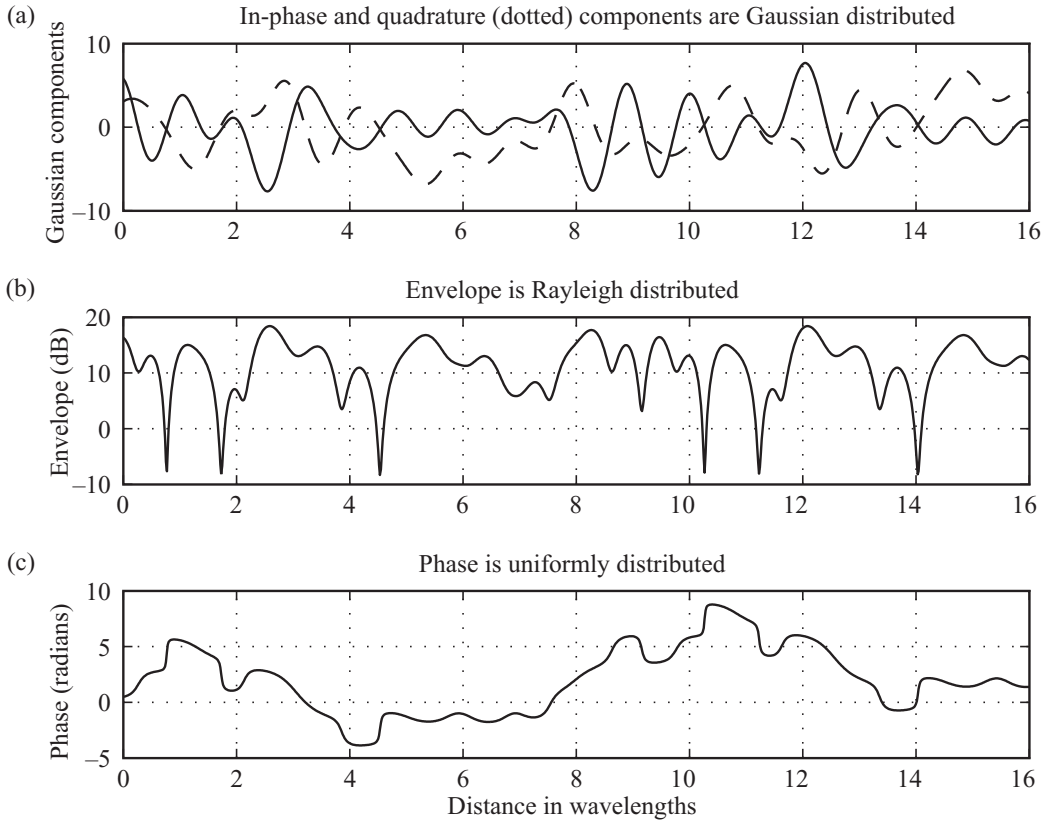


Figure C.2 Examples of: (a) the independent zero-mean Gaussian components (in-phase and quadrature); (b) envelope, and (c) phase, of a zero-mean, complex, Gaussian process, as a function of distance (or time, for a moving terminal). The signals correspond to the complex envelope variation of a carrier of wavelength λ in a narrowband multipath channel. The function depends on the Doppler spread which scales the signal in distance. The pdfs correspond to random samples of these types of waveforms

The instantaneous power of the complex Gaussian, $r^2 = x^2 + y^2$, is recognised as being a scaled chi-squared random variable with two degrees of freedom (exponential distribution).This can also be found from the envelope distribution through

$$\begin{aligned} p_{r^2}(r^2) &= p_r(r)\frac{\partial r}{\partial r^2} \\ &= \frac{1}{\overline{r^2}}\mathrm{e}^{-r^2/\overline{r^2}} \\ &= \frac{1}{2\sigma^2}\mathrm{e}^{-r^2/2\sigma^2}. \end{aligned} \tag{C.26}$$

The mean square of the power is related simply to the mean square of each component, namely

$$\begin{aligned} \overline{r^4} &= \int r^2 p(r^2)\,\mathrm{d}(r^2) \\ &= 2(\overline{r^2})^2 \\ &= 8\sigma^4, \end{aligned} \tag{C.27}$$

which gives the convenient expression for the variance of the power signal, r^2, as

$$\begin{aligned} \sigma_{r^2}^2 &= 2(\overline{r^2})^2 - (\overline{r^2})^2 \\ &= (\overline{r^2})^2 \\ &= 4\sigma^4, \end{aligned} \tag{C.28}$$

and so the standard deviation of the power is the same as the mean of the power.

Some pdfs are sketched in Figure C.3. The mode of the Rayleigh distribution is σ.

For higher moments of the Rayleigh envelope, the general formula is (Papoulis, 1991, p. 111)

$$E\{r^n\} = n!!\,\sigma^n\sqrt{\frac{\pi}{2}}, \qquad n \text{ odd}, \tag{C.29a}$$

$$= 2^{(n/2)}\left(\frac{n}{2}\right)!\,\sigma^n, \quad n \text{ even}. \tag{C.29b}$$

Cumulative density function plots for the Gaussian and Rayleigh are included in Section C.3.

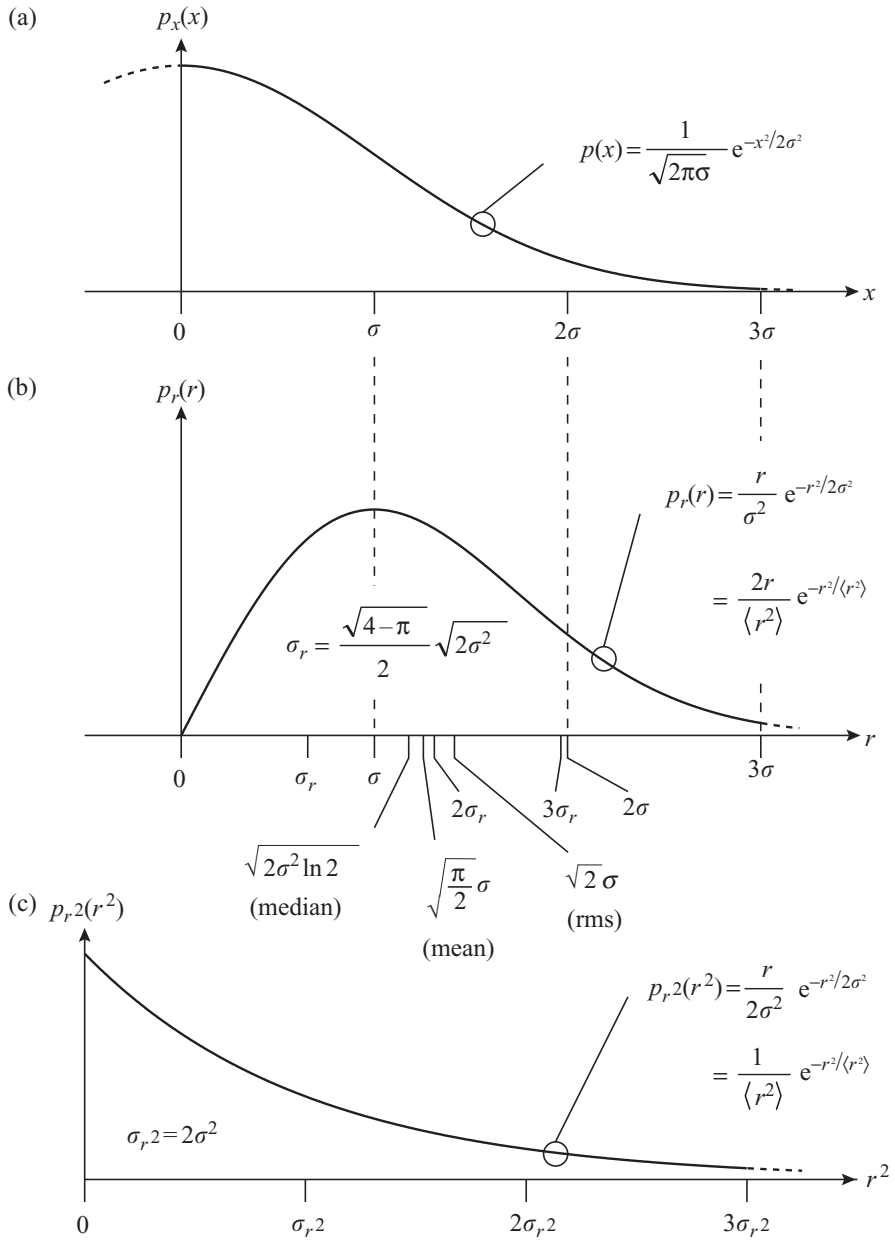


Figure C.3 The probability density distributions for a complex Gaussian process: (a) is a Gaussian component x, (b) is the Rayleigh envelope and (c) is the exponential instantaneous power

C.3 Rice

The Rice distribution results from one or both of the Gaussian processes having non-zero mean. These phase processes become

$$x_{\mathrm{Ri}} = x + x_{\mathrm{s}}, \qquad y_{\mathrm{Ri}} = y + y_{\mathrm{s}} \tag{C.30}$$

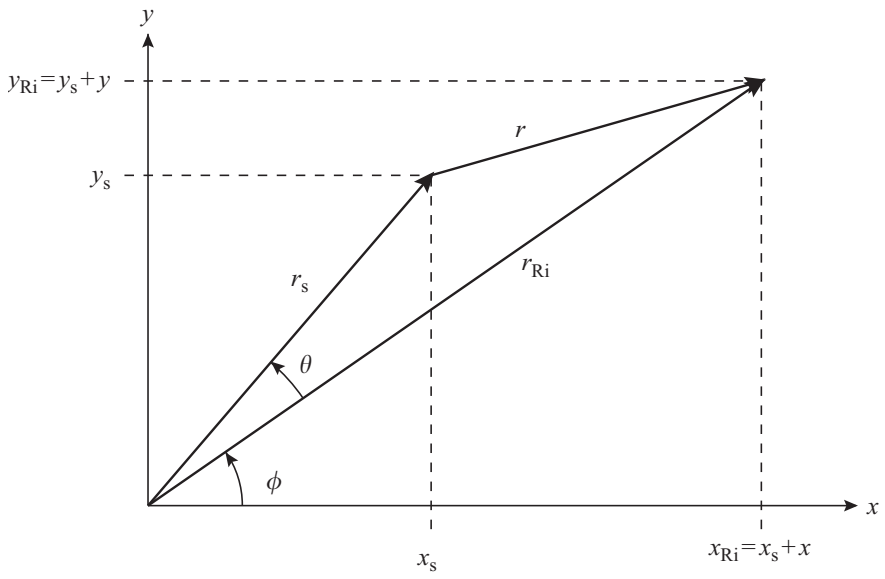


Figure C.4 The Rice process has envelope r_{Ri} comprising the additive constant r_s and the Rayleigh envelope r

where the x and y are zero-mean Gaussian and x_s and y_s are the respective means representing the dominant component (sometimes called the *specular*, or *coherent* component, with the x and y comprising the *diffuse*, or *incoherent* component) of the signal. The phasor combination is shown in Figure C.4 in which $\phi = \tan^{-1}(y_{Ri}/x_{Ri})$ is the absolute phase of the Rice envelope r_{Ri} and θ is the phase difference between r_{Ri} (Rayleigh component plus dominant component) and the dominant component r_s. The mean absolute phase of the process is $E\{\theta + \phi\}$. Since the specular component is fixed, a coordinate rotation allows the phase to be defined as just θ and so the mean becomes $E\{\theta\}$.

The Cartesian form of the pdf is

$$p_{x_{Ri},y_{Ri}}(x_{Ri}, y_{Ri}) = p_{x_{Ri}}(x_{Ri})p_{y_{Ri}}(y_{Ri}) = \frac{1}{2\pi\sigma^2}\exp\left[-\frac{(x_{Ri}-x_s)^2+(y_{Ri}-y_s)^2}{2\sigma^2}\right] \tag{C.31}$$

and using

$$p_{r_{Ri},\theta_{Ri}}(r_{Ri}, \theta) = p_{x_{Ri},y_{Ri}}(x, y)\left|\frac{\partial(x, y)}{\partial(r_{Ri}, \theta)}\right| \tag{C.32}$$

with

$$(x_{Ri}-x_s)^2+(y_{Ri}-y_s)^2 = r^2 \tag{C.33a}$$

$$= r_{Ri}^2 + r_s^2 - 2r_{Ri}r_s\cos\theta, \tag{C.33b}$$

there results

$$p_{r_{\mathrm{Ri}},\theta}(r_{\mathrm{Ri}},\theta) = \frac{r_{\mathrm{Ri}}}{2\pi\sigma^2}\exp\left[-\frac{r_{\mathrm{Ri}}^2 + r_{\mathrm{s}}^2 - 2r_{\mathrm{Ri}}r_{\mathrm{s}}\cos\theta}{2\sigma^2}\right]. \tag{C.34}$$

The envelope and phase are thus statistically dependent, unlike the Rayleigh case. As a time series signal, the $+\pi$ and $-\pi$ like transitions which occur in the phase of the Rayleigh signal as the locus passes near the origin (cf. Figure C.2), are now reduced to values which depend on the length of the envelope phasor component r_{Ri}.

The conventional Rice k-factor is the ratio of powers of the dominant component and the Rayleigh component,

$$k_{\mathrm{Rice}} = \frac{r_{\mathrm{s}}^2}{2\sigma^2}. \tag{C.35}$$

When the dominant component r_{s} approaches zero, k_{Rice} approaches 0, and the distribution reduces to Rayleigh. Similarly, when the dominant component becomes very large, the Rice distribution for the envelope approaches Gaussian with mean r_{s}. The Rice k-factor gives a convenient parameter defining the transition between the Rayleigh and Gaussian cases, but the envelope pdf cannot be written using k_{Rice} exclusively of the Gaussian standard deviation σ.

C.4 Rice envelope

For convenience, the envelope is here normalised by the Gaussian standard deviation:

$$r_{\mathrm{Ri}}^{(\mathrm{n})} = \frac{r_{\mathrm{Ri}}}{\sigma}. \tag{C.36}$$

and here the superscript '(n)' denotes the normalised form. The envelope pdf is

$$\begin{aligned} p_{r_{\mathrm{Ri}}}(r_{\mathrm{Ri}}) &= \frac{r_{\mathrm{Ri}}}{2\pi\sigma^2}\,\mathrm{e}^{-(r_{\mathrm{Ri}}^2+r_{\mathrm{s}}^2)/2\sigma^2}\int_0^{2\pi}\mathrm{e}^{r_{\mathrm{Ri}}r_{\mathrm{s}}\cos\theta/\sigma^2}\mathrm{d}\theta \\ &= \frac{r_{\mathrm{Ri}}}{\sigma^2}\,\mathrm{e}^{-\frac{r_{\mathrm{Ri}}^2+r_{\mathrm{s}}^2}{2\sigma^2}}\,I_0\left(\frac{r_{\mathrm{Ri}}r_{\mathrm{s}}}{\sigma^2}\right), \quad r_{\mathrm{Ri}} \geq 0; \end{aligned} \tag{C.37}$$

or

$$p_{r_{\mathrm{Ri}}^{(n)}}\left(r_{\mathrm{Ri}}^{(n)}\right) = \frac{1}{\sigma}r_{\mathrm{Ri}}^{(n)}\,\mathrm{e}^{-\left(\frac{1}{2}\left(r_{\mathrm{Ri}}^{(n)}\right)^2+k_{\mathrm{Rice}}\right)}I_0\left(r_{\mathrm{Ri}}^{(n)}\sqrt{2k_{\mathrm{Rice}}}\right), \quad r_{\mathrm{Ri}}^{(n)} \geq 0, \tag{C.38}$$

which is shown in Figure C.5. I_0 is the usual modified Bessel function.

The distribution is sometimes called Nakagami–Rice, in recognition of its independent development by Rice (1944) and Nakagami who reported it in English at a later stage (Nakagami, 1960).

The moment formula for the Rice envelope is (Proakis, 1983, p. 32),

$$E\left\{r_{\mathrm{Ri}}^n\right\} = (2\sigma^2)^{n/2}\,\mathrm{e}^{-k_{\mathrm{Rice}}}\Gamma\left(1+\frac{n}{2}\right)\,{}_1F_1\left(\left(1+\frac{n}{2}\right),1,k_{\mathrm{Rice}}\right), \quad n \geq 0, \tag{C.39}$$

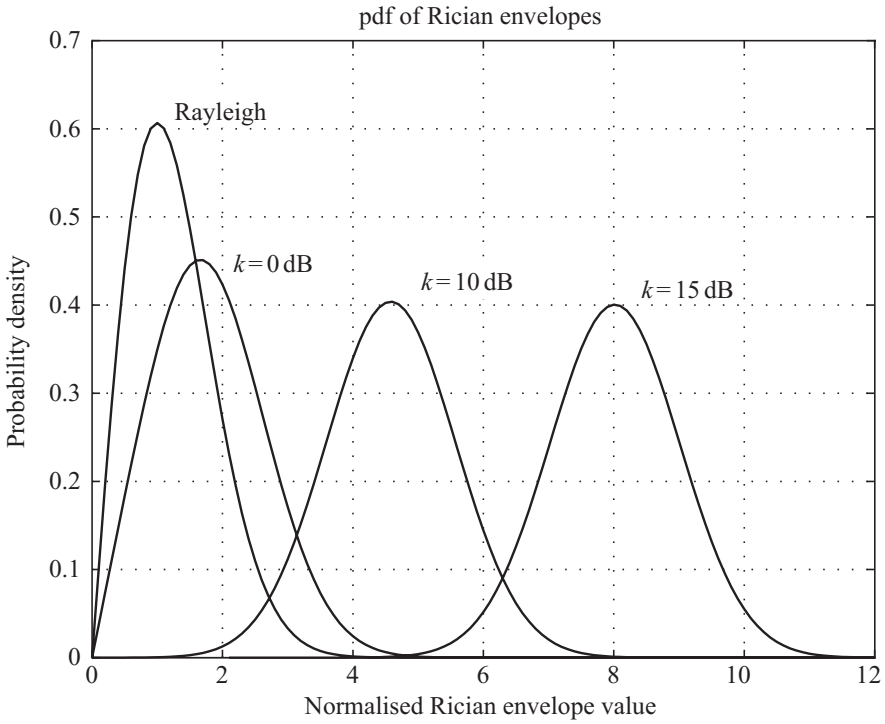


Figure C.5 *The Rice distribution is the transition between Rayleigh and Gaussian. The Rice factor k_{Rice} is the parameter which positions and shapes the distribution. The Gaussian-component distributions both have variance σ^2 and the Rayleigh distribution has variance $0.43\sigma^2$. The Rice variance is a complicated function of k_{Rice}. As k_{Rice} approaches infinity, the Rice pdf becomes a delta-like function, being a Gaussian with a variance approaching zero*

where Γ is the gamma function and

$$
\begin{aligned}
{}_1F_1(\alpha, \beta; x) &= \sum_{k=0}^{\infty} \frac{\Gamma(\alpha + k)\,\Gamma(\beta)\,x^k}{\Gamma(\alpha)\,\Gamma(\beta + k)\,k!}, \quad \beta \neq 0, -1, -2, \ldots, \\
&\to \frac{\Gamma(b)}{\Gamma(a)}\,\mathrm{e}^x x^{a-b}, \quad x \to \infty;
\end{aligned} \tag{C.40}
$$

is the confluent hypergeometric function, also known as Kummers function (Abramowitz and Stegun, 1965, p. 504). The mean of the envelope follows as

$$
E\{r_{\text{Ri}}\} = \overline{r_{\text{Ri}}} = \sigma\sqrt{\frac{\pi}{2}}\mathrm{e}^{-k_{\text{Rice}}}\,{}_1F_1\left(\frac{3}{2}, 1, k_{\text{Rice}}\right) \tag{C.41}
$$

and the mean square as

$$
E\{r_{Ri}^2\} = \overline{r_{\text{Ri}}^2} = 2\sigma^2\mathrm{e}^{-k_{\text{Rice}}}\,{}_1F_1(2, 1, k_{\text{Rice}})\,. \tag{C.42}
$$

Setting

$$F_1 = {}_1F_1\left(\tfrac{3}{2}, 1, k_{\text{Rice}}\right), \tag{C.43}$$

$$F_2 = {}_1F_1\left(2, 1, k_{\text{Rice}}\right); \tag{C.44}$$

(note that both F_1 and F_2 approach unity as k_{Rice} approaches zero) the variance of the Rice envelope can be written compactly as

$$\sigma^2_{r_{\text{Ri}}} = \sigma^2 \mathrm{e}^{-k_{\text{Rice}}} \left(2F_2 - \frac{\pi}{2}\mathrm{e}^{-k_{\text{Rice}}} F_1^2\right). \tag{C.45}$$

As the specular component r_{s} goes to zero (i.e. k_{Rice} goes to zero), the Rice mean and variance reduce to the Rayleigh mean and variance, as expected, from the identity ${}_1F_1(\alpha, \beta; 0) = 1$. As the specular component increases towards infinity, the mean approaches the specular value r_{s} and the variance approaches zero, and the pdf approaches a delta-like function.

The moments for the power, from the chi-square pdf, have the simpler formulas (see Section C.19, below, with $N = 2$ and $s_m = 2\sigma^2 k_{\text{Rice}}$):

$$E\left\{r^2_{\text{Ri}}\right\} = 2\sigma^2 \left(1 + k_{\text{Rice}}\right), \tag{C.46}$$

which is intuitive since it is the total power (sum of the diffuse and specular powers); and

$$E\left\{r^4_{\text{Ri}}\right\} = 4\sigma^4 \left(k^2_{\text{Rice}} + 4k_{\text{Rice}} + 2\right). \tag{C.47}$$

(Note that a comparison of the two expressions for $E\{r^2_{\text{Ri}}\}$ implies the relationship $F_2 = {}_1F_1(2, 1, k) = (1 + k)\exp(k)$.) The variance of the power signal r^2 is

$$\sigma^2_{r^2_{\text{Ri}}} = \left(2\sigma^2\right)^2 \left(1 + 2k_{\text{Rice}}\right). \tag{C.48}$$

Using the above chi-square expression for the envelope power, the envelope variance can be expressed without F_2 , as

$$\sigma^2_{r_{\text{Ri}}} = \left(2\sigma^2\right)^2 \left(1 + k_{\text{Rice}} - \frac{\pi}{4}\,\mathrm{e}^{-2k_{\text{Rice}}} F_1^2\right). \tag{C.49}$$

The Rice envelope cumulative density function is expressed as

$$P_{r_{\text{Ri}}}(r_{\text{Ri}}) = \text{Prob}(r \le r_{\text{Ri}}) = 1 - Q_1\left(\frac{r_{\text{s}}}{\sigma}, \frac{r_{\text{Ri}}}{\sigma}\right) \tag{C.50}$$

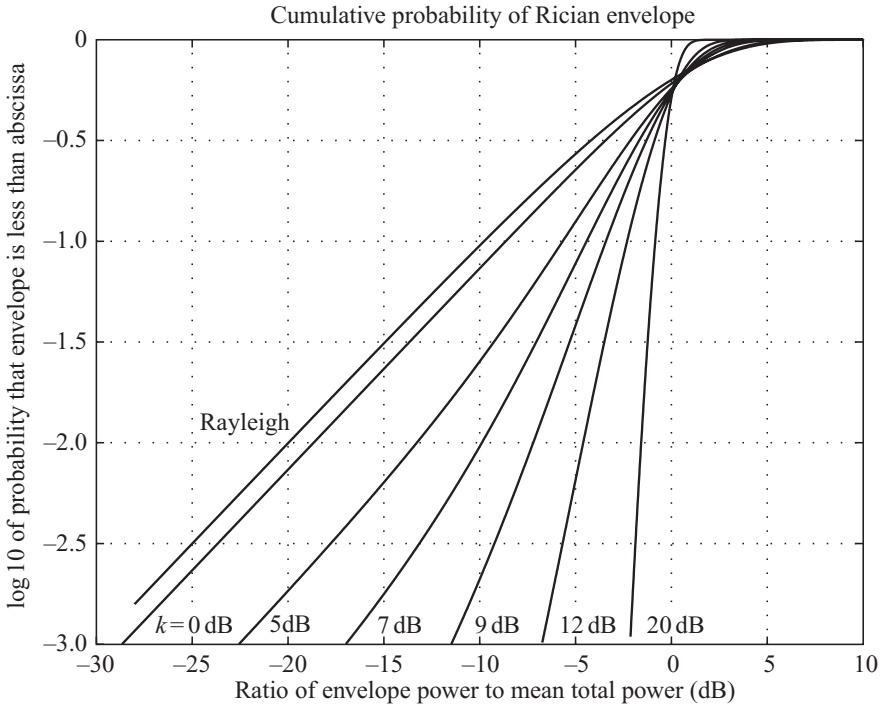


Figure C.6 *The Rice envelope cdf. For a zero specular component, the distribution is Rayleigh, and approaches Gaussian for a large specular component. The variance of the Gaussian distribution approaches zero as the specular component approaches infinity, so the pdf approaches a delta function*

where the Marcum (1960) Q function (see equation (C.53)) has the expansion (cf. Stein, 1966; Parl, 1980; Proakis, 1983), which is valid only for $a, b > 0$,

$$Q_1(a,b) = \mathrm{e}^{-(a^2+b^2)/2} \sum_{n=0}^{\infty} \left(\frac{a}{b}\right)^n I_n(ab) \tag{C.51a}$$

$$= \begin{cases} \mathrm{e}^{-\frac{(a-b)^2}{2}} \sum\limits_{n=0}^{\infty} \left(\frac{a}{b}\right)^n \mathrm{e}^{-ab} I_n(ab) & \text{if } a < b \\ 1 - \mathrm{e}^{-\frac{(a-b)^2}{2}} \sum\limits_{n=1}^{\infty} \left(\frac{b}{a}\right)^n \mathrm{e}^{-ab} I_n(ab) & \text{if } a \geq b. \end{cases} \tag{C.51b}$$

The second line is the more suitable for direct computation. For large values of the argument ab, the approximation $\mathrm{e}^{-ab} I_n(ab) \approx 1/\sqrt{2\pi ab}$ is used. An efficient algorithm is given below. Limiting values are

$$\begin{aligned} &Q_1(0,0) = 1,\ Q_1(a,0) = 1,\ Q_1(a,\infty) = 0; \\ &Q_1(\infty,b) = 1,\ Q_1(\infty,\infty) = 0.5,\ Q_1(0,b) = \exp(-b^2/2). \end{aligned} \tag{C.52}$$

The Rice envelope cdf is sketched in Figure C.6 for values of the Rice k-factor, including the Rayleigh case.

C.5 Calculating the Marcum Q function

The Marcum Q_1 function is the integral of the Rice envelope pdf,

$$Q_1(a, b) = \int_b^\infty x\, \mathrm{e}^{-(x^2+a^2)/2} I_0(ax)\, \mathrm{d}x \tag{C.53}$$

and can be expressed for a and b representing power quantities (although mathematically a and b are dimensionless) using

$$Q_1\left(\sqrt{2a}, \sqrt{2b}\right) = \int_b^\infty e^{-(x+a)} I_0\left(2\sqrt{ax}\right)\, \mathrm{d}x. \tag{C.54}$$

Useful results for computation are as follows. Using the expansion for large $|z|$,

$$I_0(z) \approx \frac{\mathrm{e}^z}{\sqrt{2\pi z}}, \tag{C.55}$$

there results

$$Q_1(a, b) \approx 0.5\, \mathrm{erfc}\left(\frac{b-a}{\sqrt{2}}\right), \quad b \gg 1,\ b \gg b-a. \tag{C.56}$$

From

$$I_0(z) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{e}^{z\cos\theta}\mathrm{d}\theta = \frac{1}{\mathrm{j}2\pi}\int_{c-\mathrm{j}\infty}^{c+\mathrm{j}\infty} \frac{1}{p}\mathrm{e}^{z(p+1)/(2p)}\mathrm{d}p, \quad c > 0,\ z > 0; \tag{C.57}$$

there results

$$Q_1(a, b) + Q_1(b, a) = 1 + \mathrm{e}^{-(a^2+b^2)/2} I_0(ab), \tag{C.58}$$

giving

$$Q_1(a, a) = 0.5\left(1 + \mathrm{e}^{-a^2/2} I_0(a^2)\right) \tag{C.59}$$

and also

$$Q_1(a, b) \approx 0.5\left(\mathrm{erfc}\left(\frac{b-a}{\sqrt{2}}\right) + \frac{\mathrm{e}^{-(b-a)^2/2}}{\sqrt{2\pi ab}}\right),\ b \gg 1,\ a \gg 1,\ b \gg b-a > 0. \tag{C.60}$$

The following algorithm (Parl, 1980) is a simple and efficient approximate computation of the Q_1 function. For a and b positive and non-zero,

$$Q_1(a, b) = \begin{cases} 0.5\dfrac{\alpha_N}{\beta_N}\, \mathrm{e}^{-0.5(a-b)^2} & \text{if } a < b \quad \text{(C.61a)} \\ 1 - 0.5\dfrac{\alpha_N}{\beta_N}\, \mathrm{e}^{-0.5(a-b)^2} & \text{if } a \geq b \quad \text{(C.61b)} \end{cases}$$

in which

$$\alpha_n = d_n + \frac{2n}{ab}\alpha_{n-1} + \alpha_{n-2} \tag{C.62}$$

$$\beta_n = 1 + \frac{2n}{ab}\beta_{n-1} + \beta_{n-2} \tag{C.63}$$

are iterated until $N = 5(1 + \sqrt{ab})$, using

$$d_{n+1} = d_n d_1, \quad \alpha_{-1} = 0, \quad \beta_{-1} = 0, \quad \beta_0 = 0.5; \tag{C.64}$$

$$\alpha_0 = 1, \quad d_1 = \frac{a}{b} \quad \text{if } a < b \tag{C.65a}$$

$$\alpha_0 = 0, \quad d_1 = \frac{b}{a} \quad \text{if } a \geq b. \tag{C.65b}$$

Figure C.7 shows the Marcum Q_1 function. Parl's algorithm can be modified to calculate also the generalised Q function, Q_m (see Section C.19, below).

Finally, simple bounding formulas, useful for quick and approximate calculations involving Q_1, are:

$$\mathrm{e}^{-(\beta+\alpha)^2/2} \leq Q_1(\alpha, \beta) \leq \mathrm{e}^{-(\beta-\alpha)^2/2}, \quad \beta > \alpha; \tag{C.66}$$

$$Q_1(\alpha, \beta) \geq 1 - \tfrac{1}{2}\left(\mathrm{e}^{-(\beta-\alpha)^2/2} - \mathrm{e}^{-(\beta+\alpha)^2/2}\right), \quad \alpha > \beta. \tag{C.67}$$

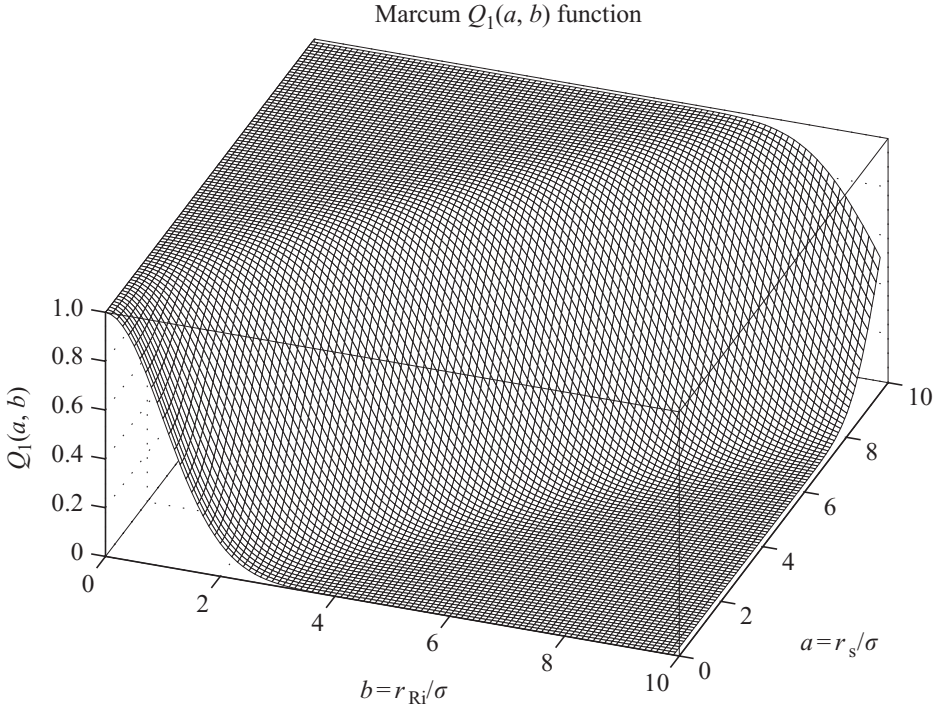


Figure C.7 Marcum Q function $Q_1(a, b)$ is the probability integral that a Rice envelope comprising specular envelope $a = r_s/\sigma$ and diffuse (Rayleigh envelope) power $2\sigma^2$, is greater than $b = r_{\mathrm{Ri}}/\sigma$; i.e. $Q_1(a, b) = Q_1(r_s/\sigma, r_{\mathrm{Ri}}/\sigma) = \mathrm{Prob}\,(r > r_{\mathrm{Ri}})$

C.6 Rice phase

The pdf of the Rice phase (Beckmann and Spizzichino, 1963, p. 127; Beckmann, 1967, p. 123),

$$p_{\theta_{\mathrm{Ri}}}(\theta) = \int_0^\infty p\,(r_{\mathrm{Ri}}, \theta)\ \mathrm{d}r_{\mathrm{Ri}}$$
$$= \frac{1}{2\pi}\mathrm{e}^{-k_{\mathrm{Rice}}}\left(1 + \sqrt{\pi k_{\mathrm{Rice}}}\cos\theta\ \mathrm{e}^{k_{\mathrm{Rice}}\cos^2\theta}\left\{1 + \mathrm{erf}\left(\sqrt{k_{\mathrm{Rice}}}\cos\theta\right)\right\}\right), \tag{C.68}$$

has a complicated dependency on the Rice k-factor, k_{Rice}. As the specular component reduces to zero ($k_{\mathrm{Rice}} \to 0$), the phase distribution becomes $1/2\pi$, i.e. uniform over 2π, the Rayleigh case. As k_{Rice} becomes very large, the phase approaches a delta-like function at zero (i.e. at the phase of the dominant component). The distribution is depicted in Figure C.8.

C.7 Rice generalisations: Hoyt and Beckmann distributions

If the variances of the two Gaussian components are different, say σ_x^2 and σ_y^2, then the distribution of the envelope becomes Hoyt (1947), with the envelope

$$p_{r_{\mathrm{H}}}(r_{\mathrm{H}}) = \frac{r_{\mathrm{H}}}{\sigma_x\sigma_y}\ \mathrm{e}^{-(\sigma_x^2+\sigma_y^2)r_{\mathrm{H}}^2/4\sigma_x^2\sigma_y^2}\, I_0\left(\frac{\left(\sigma_x^2 - \sigma_y^2\right)r_{\mathrm{H}}^2}{4\sigma_x^2\sigma_y^2}\right), \quad r \geq 0. \tag{C.69}$$

The Hoyt distribution crops up in rough surface scattering and also during transformations of joint envelope distributions. It reduces to Rayleigh when the variances are the same.

Beckmann (1959, 1967) gives a general distribution with $\sigma_x \neq \sigma_y,\ x_{\mathrm{s}},\ y_{\mathrm{s}} \neq 0$; which includes, and which conveniently reduces to, Rayleigh, Rice and Hoyt. The envelope phasors are drawn in Figure C.9.

The pdf of the normalised Hoyt envelope is

$$p_{r_{\mathrm{B}}}(r_{\mathrm{B}}) = \frac{K^2+1}{K}\ r_{\mathrm{B}}\exp\left[-\frac{K^2+1}{2}\left(k_{\mathrm{Hoyt}} + \frac{K^2+1}{2K^2}r_{\mathrm{B}}^2\right)\right]\ S_{\mathrm{B}}, \quad r_{\mathrm{B}} \geq 0, \tag{C.70}$$

where

$$S_{\mathrm{B}} = \sum_{m=0}^{\infty}(-1)^m\varepsilon_m I_m\left(\frac{\left(K^4-1\right)}{4K^2}r_{\mathrm{B}}^2\right) I_{2m}\left(\sqrt{k_{\mathrm{Hoyt}}}(K^2+1)r_{\mathrm{B}}\right), \tag{C.71}$$

and in which there are: an asymmetry parameter which gives the ellipticity on the x–y plane of the Rayleigh envelope distribution component (= 1 for Rayleigh and Rice),

$$K = \frac{\sigma_x}{\sigma_y}; \tag{C.72}$$

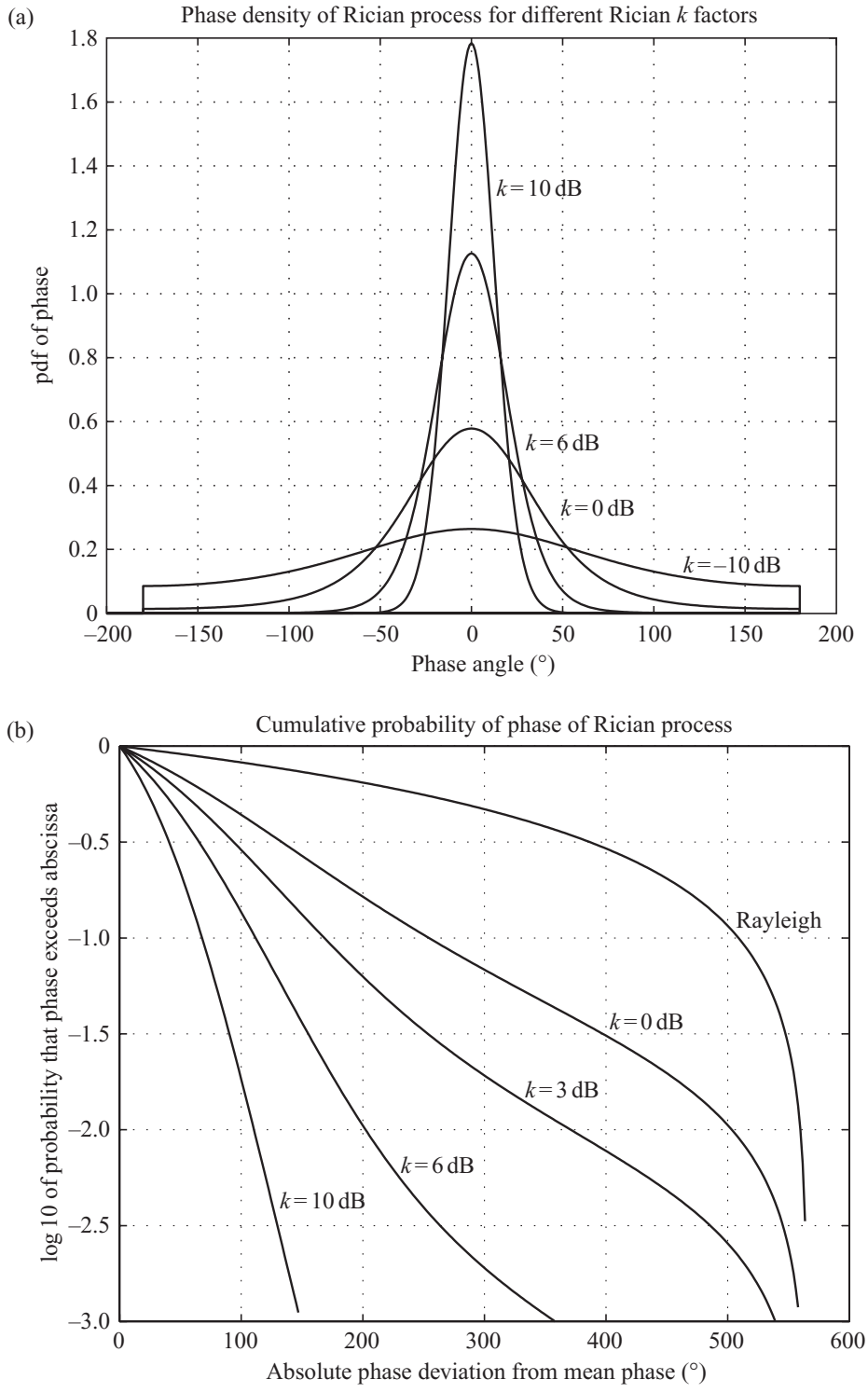


Figure C.8 *The Rice phase (a) pdf and (b) cdf with the Rice k-factor (ratio of powers of dominant component to random components) as parameter. The total power, i.e. the sum of the Rayleigh power ($2\sigma^2$) and the specular power (r_s^2) is the normalisation for the envelope. The phase signal is taken modulo $2\pi(360°)$ for the pdf, but not for the cdf*

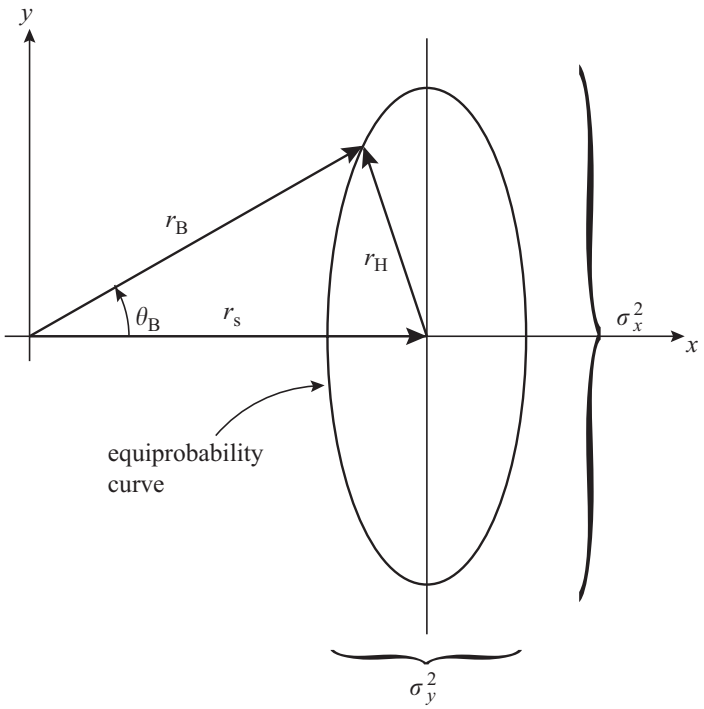


Figure C.9 A depiction of the Beckmann distribution phasors. The Gaussian components x and y have different variances and the Rayleigh part of the distribution takes on an elliptic-shaped density in the plane as indicated by the extent being proportional to the variances

a Hoyt k-factor (cf. the Rice k-factor),

$$k_{\text{Hoyt}} = \frac{r_s^2}{\sigma_x^2 + \sigma_y^2}; \tag{C.73}$$

and finally, in which the envelope is normalised as

$$r_B = \frac{r_H}{\sqrt{\sigma_x^2 + \sigma_y^2}}; \tag{C.74}$$

(note the nomalisation here differs by a factor of $\sqrt{2}$ from that in the Rice case above); and the factor $\varepsilon_m = 1$ for $m = 0$, and $\varepsilon_m = 2$ for $m \neq 0$.

The Beckmann phase distribution is

$$p_{\theta_B}(\theta_B) = \frac{K e^{-k_{\text{Hoyt}}(1+K^2)/2}}{2\pi\left(K^2 \cos^2\theta_B + \sin^2\theta_B\right)}\left(1 + G\sqrt{(\pi)}\, e^{G^2}(1 + \text{erf}(G))\right) \tag{C.75}$$

where

$$G = K \cos\theta_B \sqrt{\frac{k_{\text{Hoyt}}(1+K^2)}{2\left(K^2\cos^2\theta_B + \sin^2\theta_B\right)}}. \tag{C.76}$$

C.8 Random phasor plus Rayleigh

If the specular component in the Rice distribution is now made random, the resulting envelope is written

$$r_{\rm rr}{\rm e}^{{\rm j}\theta_{\rm rr}} = r_z{\rm e}^{{\rm j}\theta_z} + r{\rm e}^{{\rm j}\theta} \tag{C.77}$$

where subscript z denotes the random component, and no subscript denotes the Rayleigh phasor. Recalling from the Rice case, with r_z temporarily fixed,

$$p_{r_{\rm rr}}(r_{\rm rr}|r_z) = \frac{r_{\rm rr}}{\sigma^2}\,{\rm e}^{-(r_{\rm rr}^2+r_z^2)/2\sigma^2}\,I_0\left(\frac{r_{\rm rr}r_z}{\sigma^2}\right) \tag{C.78}$$

then the pdf of the random phasor plus Rayleigh is

$$p_{r_{\rm rr}}(r_{\rm rr}) = \frac{r_{\rm rr}}{\sigma}\int_0^\infty {\rm e}^{-(r_{rr}^2+r_z^2)/2\sigma^2}\,I_0\left(\frac{r_{\rm rr}r_z}{\sigma^2}\right)p_{r_z}(r_z){\rm d}r_z \tag{C.79}$$

where $p_{r_z}(r_z)$ is the pdf of the amplitude of the random phasor. The expression usually requires numerical integration. The limiting forms as the random phasor becomes small and large relative to the Rayleigh phasor, are the pdf of the phasor itself and the Rayleigh distribution, respectively (Beckmann, 1967, p. 124), i.e.

$$p_{r_{\rm rr}}(r_{\rm rr}) = \begin{cases} p_{r_z}(r_z) & \text{for } r_z \gg \sqrt{2\sigma^2} \quad \text{(random phasor)},\\ p_r(r) & \text{for } r_z \ll \sqrt{2\sigma^2} \quad \text{(Rayleigh)}. \end{cases} \tag{C.80}$$

If the specular component is normal, so that x, $x_{\rm s}$, y, $y_{\rm s}$, are normal and independent, then $x_{\rm Ri}$ and $y_{\rm Ri}$ are also normal and $r_{\rm Ri}^2$ has a non-central chi-square distribution as long as $\sigma_x = \sigma_y$ and $\sigma_{x{\rm s}} = \sigma_{y{\rm s}}$.

C.9 Lognormal

Whereas the Gaussian distribution results from the addition of many random variables, the lognormal distribution results from the product of many positive random variables. When a variable expressed in log units has a Gaussian distribution, then that variable follows a lognormal distribution. The transformation of variables between the distributions is denoted here with $z = {\rm e}^x$, or $\ln z = x$. So if x is Gaussian, then z is lognormal. Alternatively stated, if z is lognormal, then $\ln z$ is Gaussian. The pdf of the lognormal distribution is found as

$$p_l(z) = p_x(x)\left|\frac{\partial x}{\partial z}\right| = \frac{1}{\sqrt{2\pi}\sigma_{lz}z}{\rm e}^{-(\ln z - m_{lz})^2/2\sigma_{lz}^2}, \tag{C.81}$$

where the notation $m_x = E\{x\} = m_{lz} = E\{\ln z\}$ and similarly, $\sigma_x = \sigma_{lz} = \sqrt{{\rm Var}(\ln z)}$ are the mean and standard deviation respectively of the Gaussian $\ln z$.

C.9.1 Example: lognormal plus Rayleigh

As an example, a lognormal component plus Rayleigh distributed envelope, using the result of the previous section, has the pdf

$$p_{lR}(r_{lr}) = \frac{r_{lr}}{\sqrt{2\pi}\sigma^2\,\sigma_{lz}} \int_0^\infty \frac{1}{z} \exp\left[-\frac{(\ln z - m_{lz})^2}{2\sigma_{lz}^2} - \frac{r_{lr}^2 + z^2}{2\sigma^2} \right] I_0\left(\frac{r_{rl} z}{\sigma^2}\right) \mathrm{d}z. \tag{C.82}$$

This is equivalent to the Rice distribution with a lognormally varying specular component. The limiting cases of lognormal and Rayleigh for large and small r_{lz} respectively are easily established mathematically from the pdf.

The cdf is found by integration to be (cf. equation (C.7)),

$$\begin{aligned} \text{Prob}\,(z \le z_0) &= \Phi\left(\frac{\ln z_0 - m_{lz}}{\sigma_{lz}}\right) \\ &= \frac{1}{2} + \frac{1}{2}\text{erfc}\left(\frac{\ln z_0 - m_{lz}}{\sqrt{2}\sigma_{lz}}\right) \end{aligned} \tag{C.83}$$

and the median follows as

$$z_{\text{med}} = \mathrm{e}^{m_{lz}}, \tag{C.84}$$

which is always less than the mean, m_z. Some examples of the lognormal pdf are given in Figure C.10.

C.9.2 Relation between first moments

For the lognormal distribution, the mean and standard deviation are not the same as the Gaussian ones appearing in the pdf equation. By calculation of $m_z = E\{z\}$ and $\sigma_z^2 = E\{z^2\} - E\{z\}^2$ the relationship between the Gaussian and lognormal means and variances is (Ang and Tang, 1975, p. 104):

$$m_z = \mathrm{e}^{\left(m_{lz} + \frac{1}{2}\sigma_{lz}^2\right)}, \quad m_{lz} = \ln m_z - \tfrac{1}{2}\sigma_{lz}^2; \tag{C.85}$$

$$\sigma_z^2 = m_z^2\left(\mathrm{e}^{\sigma_{lz}^2} - 1\right), \quad \sigma_{lz}^2 = \ln\left(1 + \left(\frac{\sigma_z}{m_z}\right)^2\right). \tag{C.86}$$

For very small lognormal variances in the sense $\sigma_z \ll m_z$, then $\sigma_{lz} \approx \sigma_z/m_z$, the covariance of z.

C.9.3 Decibel units

In mobile communications, the motivation for the lognormal is for describing the long-term fading component of the narrowband fading envelope. The log of the mean of the envelope is postulated as being Gaussian, and the envelope mean follows as being lognormal. The decibel units used in many of the engineering applications prompts a formulation below using 20log(.), where log implies base 10. (Natural logs

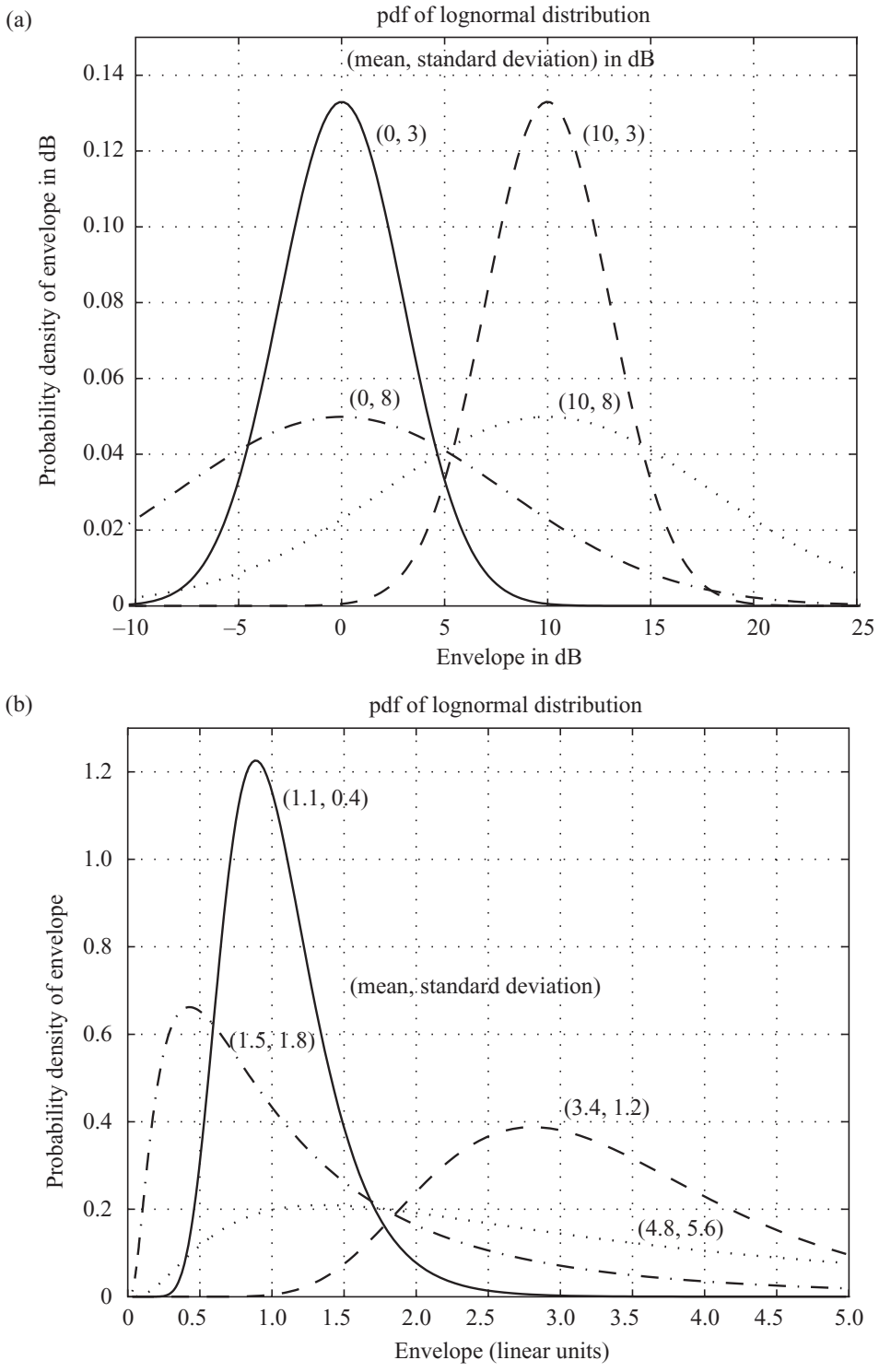


Figure C.10 Some pdf examples for the lognormal distribution showing (a) its Gaussian shape on the log abscissa and (b) its skew shape on the linear abscissa. The examples in (a) and (b) correspond and demonstrate the relationship between the lognormal and Gaussian means and variances

are denoted with ln as above.) Also, since the signal is in dB, it makes no difference whether the basic envelope quantity is power (r^2) or voltage (r). The envelope r in dB, with the normalisation understood, is

$$r_{\mathrm{dB}} = 20 \log (r) \tag{C.87}$$

and again the log implies base 10. The notation for the envelope moments in dB and the dB envelope moments, is as follows. The envelope mean expressed in dB is

$$\bar{r}_{\mathrm{dB}} = 20 \log (\bar{r}) \,, \tag{C.88}$$

which is 10 log ($\sigma \pi/2$) for Rayleigh envelopes. The notation for the dB moments of interest is, using the mean as an example:

$$\begin{aligned} E\{r_{\mathrm{dB}}\} = \overline{r_{\mathrm{dB}}} &= \int_{-\infty}^{\infty} r_{\mathrm{dB}} \; p_{r\mathrm{dB}} \, (r_{\mathrm{dB}}) \, \mathrm{d}r_{\mathrm{dB}} \\ &= \int_{0}^{\infty} 20 \log (r) \;\; p_{lr} \, (r) \, \mathrm{d}r; \end{aligned} \tag{C.89}$$

and the pdf of the decibel signal, $p_{r\mathrm{dB}}(r_{\mathrm{dB}})$, is Gaussian. The lognormal pdf for the envelope is now

$$p_{lr} \, (r) = p_{r\mathrm{dB}} \, (r_{\mathrm{dB}}) \left| \frac{\partial r_{\mathrm{dB}}}{\partial r} \right| = \frac{20 \log \mathrm{e}}{\sqrt{2\pi}\sigma_{\mathrm{dB}} r} \mathrm{e}^{-(r_{\mathrm{dB}} - \overline{r_{\mathrm{dB}}})^2 / 2\sigma_{\mathrm{dB}}^2} \tag{C.90}$$

where σ_{dB} is the standard deviation of the Gaussian process in dB units, given by

$$\sigma_{\mathrm{dB}}^2 = \overline{r_{\mathrm{dB}}^2} - (\overline{r_{\mathrm{dB}}})^2 \left(= E\{r_{\mathrm{dB}}^2\} - (E\{r_{\mathrm{dB}}\})^2 \right), \tag{C.91}$$

and the mean and variance of the log normal envelope are related to these by

$$m_r = \exp \left[\frac{m_{\mathrm{dB}}}{20 \log \mathrm{e}} + \frac{1}{2} \left(\frac{\sigma_{\mathrm{dB}}}{20 \log \mathrm{e}} \right)^2 \right] \tag{C.92}$$

and

$$\sigma_r^2 = m_r^2 \left(\mathrm{e}^{(\sigma_{\mathrm{dB}}/20 \log \mathrm{e})^2} - 1 \right). \tag{C.93}$$

In terms of the power (the envelope squared) the distribution is the same with the moment definitions and the multiplicative constant changed to 10 log e. For example, for the power signal, r^2, in dB, the lognormal distribution is written

$$p(r^2) = \left(\frac{10}{\ln 10} \right) \frac{1}{r^2 \sigma_{r^2} \sqrt{2\pi}} \mathrm{e}^{-(10 \log_{10} r^2 - m_{r^2})^2 / 2\sigma_{r^2}^2}. \tag{C.94}$$

Finally, it is noted that distributions other than Gaussian can be transformed to give a log form. The log-Rayleigh and log-Nakagami, for example, have been used for fitting to experimental results (Suzuki, 1977).

C.10 Suzuki

Combining the short-term Rayleigh and long-term lognormal distributions provides one model for the stochastic component of the path loss of a narrowband signal in mobile communications.

The lognormal distribution is over the mean of the envelope. This can be interpreted as Gaussian for the envelope mean in dB ($20 \log \bar{r}$) or the mean of the dB envelope ($\overline{20 \log r}$), since the quantities differ only by a multiplicative constant (additive dB constant). Recalling that the Rayleigh envelope mean is linearly related to the Gaussian standard deviation of the in-phase and quadrature components, namely $\bar{r} = \sqrt{\pi/2}\sigma$, the lognormal distribution can be applied to the σ, resulting in the Suzuki (1977) distribution,

$$p_{\mathrm{Su}}(r) = \int_0^\infty \frac{r}{\sigma^2} \, \mathrm{e}^{-r^2/2\sigma^2} \times \frac{1}{\sqrt{2\pi}\sigma\sigma_l} \, \mathrm{e}^{-(\ln\sigma - m_l)^2/2\sigma_l^2} \, \mathrm{d}\sigma, \tag{C.95}$$

in which the σ has been normalised. This is an example of a *compound*, or *mixture, distribution*.

To couch the expression in terms of the envelope and its mean in dB terms, the Rayleigh envelope pdf is written as

$$p_r(r) = p_r(r \mid \bar{r}) = \frac{\pi}{2} \frac{r}{\bar{r}^2} \, \mathrm{e}^{-(\pi/4)(r^2/\bar{r}^2)}. \tag{C.96}$$

From this and the above results, the Suzuki distribution can be written

$$p_{r_{\mathrm{Su}}}(r) = 20 \log \mathrm{e} \sqrt{\frac{\pi}{8\sigma_l^2}} r \int_0^\infty \frac{1}{\bar{r}^3} \exp\left[-\frac{\pi}{4}\left(\frac{r}{\bar{r}}\right)^2\right] \exp\left[-\frac{(20 \log r - \overline{20 \log r})^2}{2\sigma_l^2}\right] \mathrm{d}\bar{r}, \tag{C.97}$$

where σ_l is the standard deviation of the dB envelope. No closed form has been found for these integrals, which is a practical inconvenience when applying the Suzuki distribution. However, the distribution has the advantage of being based on a sound physical model for the envelope, and as such, offers good accuracy to experimental results for the envelopes of narrowband signals.

The integrand is well-behaved, so estimating the integral by numerical integration is normally straightforward as long as practical limits are chosen, for example several standard deviations above the mean depending on the accuracy required. The class of integral also lends itself to estimation by Hermite integration, in which

$$\int \mathrm{e}^{-x^2} f(x) \, \mathrm{d}x \approx \sum_{i=1}^{n} W_i f(x_i), \tag{C.98}$$

where x_i is the ith zero of the Hermite polynomial $H_n(x)$ and W_i are the corresponding weights (Abramowitz and Stegun, 1965, p. 890).

C.11 Rice with lognormal mean (mixture distribution)

The Rice distribution can also be mixed with the lognormal the same way as in the Suzuki. In mobile communications, the mix appeals from a physical mechanism viewpoint, but the extra parameter (the Rice k-factor, k_{Rice}) and the need for numerical integration of a complicated pdf make it numerically inconvenient to deal with. With the Rice distribution it is simpler to formulate with the mean power, $\overline{r_{\text{Ri}}^2}$, rather than the envelope mean. The pdf is then

$$\begin{aligned} p_{l\text{Ri}(r_{\text{Ri}})} &= \int p_{r_{\text{Ri}}}\left(r_{\text{Ri}} \mid \overline{r_{\text{Ri}}^2}\right) p_{\ln}\left(\overline{r_{\text{Ri}}^2}\right) \mathrm{d}\overline{r_{\text{Ri}}^2} \\ &= \frac{r_{\text{Ri}}}{\sigma^2}\, \mathrm{e}^{-(r_{\text{Ri}}^2+r_{\text{s}}^2)/2\sigma^2} I_0\left(\frac{r_{\text{Ri}} r_{\text{s}}}{\sigma^2}\right) \int \frac{1}{\sqrt{2\pi}\sigma_l \overline{r_{\text{Ri}}^2}} \mathrm{e}^{-\left(\ln \overline{r_{\text{Ri}}^2} - \overline{\ln r_{\text{Ri}}^2}\right)/2\sigma_l^2} \mathrm{d}\overline{r_{\text{Ri}}^2}, \end{aligned} \tag{C.99}$$

where σ_l^2 is the variance in log units of the distribution of the mean envelope power (also in log units).

The phase depends on the envelope, and the phase pdf can be calculated in the same way as for the Rice phase.

C.12 Nakagami

M. Nakagami (Nakagami, 1960) undertook a major measurement campaign for multipath propagation via the troposphere and ionosphere and developed distributions to fit, during the early 1940s. His fading distribution, developed to model the measured envelope results, is versatile and simple, requiring only a single parameter, m, which corresponds to the amount of fading in the sense that a larger m models decreasing fading. The distribution is written

$$p_{r_{\text{N}}}(r) = \frac{2}{\Gamma(m)} \left(\frac{m}{\overline{r^2}}\right)^m r^{(2m-1)}\, \mathrm{e}^{-\frac{mr^2}{\overline{r^2}}}, \quad r \geq 0; \tag{C.100}$$

where m is mathematically expressed as a ratio of moments (inverse normalised variance of envelope), here with $R = \overline{r^2}$

$$m = \frac{R^2}{\overline{\left(r^2 - R\right)^2}} \tag{C.101}$$

and $m = 0.5$ gives one-sided Gaussian or 'half-normal', and $m = 1$ gives Rayleigh. Note that m was only once measured by Nakagami to be less than $\frac{1}{2}$; however, mathematically m can be less than $\frac{1}{2}$.

Other distributions can be approximated by the Nakagami distribution. For example, for a Nakagami pdf with a given $m \geq 1$ and R (note R is not a shape-changing parameter), the equivalent Rice pdf (equationss C.37, C.38, and recall

$k_{Rice} = r_s^2/(2\sigma^2)$) can be found using

$$\begin{aligned} k_{Rice} &= m - 1 + \sqrt{m^2 - m}, \\ \sigma &= \sqrt{\frac{1}{2}\frac{R}{m}(m - \sqrt{m^2 - m})}. \end{aligned} \tag{C.102}$$

Conversely, for a Rice pdf with a given k_{Rice} and σ, an equivalent Nakagami pdf can be found using

$$\begin{aligned} m &= \frac{(1 + k_{Rice})^2}{1 + 2k_{Rice}}, \\ R &= 2\sigma^2 + r_s^2 = 2\sigma^2(1 + k_{Rice}). \end{aligned} \tag{C.103}$$

The cdf can be couched as the incomplete gamma function $P(a, x)$,

$$P_{\text{N}}(r < r_0) = P\left(m, \frac{mr_0^2}{R}\right) \tag{C.104}$$

where (see Section 6.5 in Abramowitz and Stegun, 1965)

$$P(a, x) = \frac{\gamma(a, x)}{\Gamma(a)} = \frac{1}{\Gamma(a)} \int_0^x \mathrm{e}^{-t} t^{a-1}\, \mathrm{d}t. \tag{C.105}$$

The main attraction of the Nakagami distribution is there being only the single parameter m to optimise when fitting experimental data or in setting up simulations. It also lends itself well to integrating with error expressions that have an exponential form (for example non-coherent and differential schemes) for finding bit error ratios over quasi-static fading channels including the Rayleigh case ($m = 1$).

The m parameter is worst case fading with smallest m, and $m = \infty$ being the flat channel. For a maximum ratio combination of m equal power, uncorrelated, Rayleigh envelopes, then the Nakagami-m gives the distribution of the combined signal envelope. Some Nakagami distributions are given in Figure C.11, showing the range of shapes that can be obtained.

C.13 Gamma (chi-squared)

For the envelope power, the Nakagami distribution transforms to the gamma distribution, i.e. the replacement of r^2 with s gives

$$p_{s\Gamma}(s) = \frac{m^m s^{(m-1)}\, \mathrm{e}^{-(ms/\overline{s})}}{\Gamma(m)\, \overline{s}^m}, \tag{C.106}$$

which, for $m = 1$, is the exponential distribution as expected. The gamma distribution is another name for the chi-squared distribution (for the case of two degrees of freedom), which is discussed below.

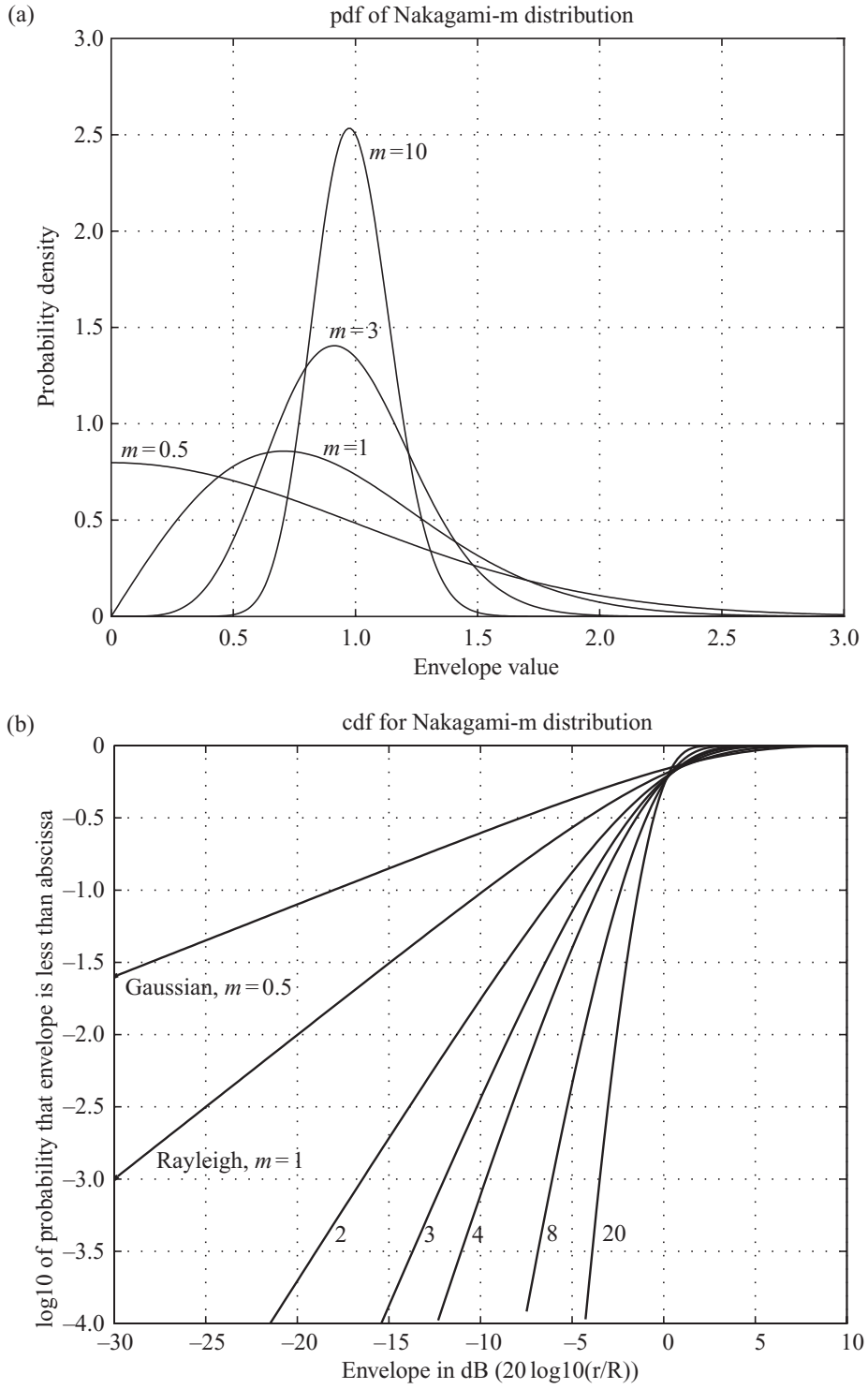


Figure C.11 *(a) The Nakagami-m pdf has a single parameter which adjusts the shape, making it a convenient and versatile distribution for describing the envelopes in multipath propagation. The $m = 0.5$ curve is one-sided Gaussian and the $m = 1$ is Rayleigh. As m approaches infinity, the pdf approaches a delta-like function describing a flat channel. (b) The cdf for the Nakagami-m distribution*

C.14 Generalised gamma

The skew, bell-shaped curves found experimentally for the envelope pdfs can be characterised with a power term for the behaviour near the origin (where the curve terminates), and some decaying exponential which describes the asymptotic behaviour. Griffiths and McGeehan (1982) proposed the pdf form of a power gamma

$$p(r) = Ar^a \, \mathrm{e}^{-r^b}, \tag{C.107}$$

which has the parameter b controlling the distribution shape for small values of the envelope variable r, and parameter a for the large values of r. This form has no physical basis, it is a convenient curve-fitting equation only. When the function is normalised by finding A in terms of the distribution parameters, it gives the generalised gamma (Stacy, 1962; Stacy and Mihram, 1965; Coulson *et al.*, 1998), here expressed using a parameter c (which has an analogous effect to b in the Griffiths–McGeehan form),

$$p_{\mathrm{GG}}(r) = \frac{cr^{c\alpha-1}}{\beta^{c\alpha}\Gamma(\alpha)} \, \mathrm{e}^{-(r/\beta)^c}; \tag{C.108}$$

a and c are primary parameters, and α and β are auxiliary parameters with

$$\alpha = \frac{(a+1)}{c}. \tag{C.109}$$

Calculating the distribution moments gives

$$\beta = E\{r\}\frac{\Gamma((a+1)/c)}{\Gamma((a+2)/c)} \tag{C.110a}$$

$$= \sqrt{E\{r^2\}\frac{\Gamma((a+1)/c)}{\Gamma((a+3)/c)}}, \tag{C.110b}$$

so the variance can be expressed in terms of the parameters. For the figures below, β is set to $\sqrt{2}$ so that the curves include the Rayleigh distribution.

The advantage of the generalised gamma for curve-fitting is the quasi-independent control over the shape of the distribution for small and large envelope values. The lower tail shape of the pdf is determined by $c\alpha$ (or more specifically by $c\alpha - 1$) and the upper tail by c. To get a feel for this effect, some pdfs for different parameters are sketched in Figure C.12 demonstrating, for example, the range (convex to concave) of forms for the small values of envelope. The cdf of Figure C.12(c) also demonstrates the control over the lower tail behaviour.

The many distributions that the generalised gamma reduces to are tabulated in Table C.1 (Griffiths and McGeehan, 1982). For example, with $c = 1$, $\beta = 2$, $a = N/2 - 1$, $\alpha = N$, the chi-squared distribution (see below) with N degrees of freedom is produced, where the variable r corresponds to the envelope squared

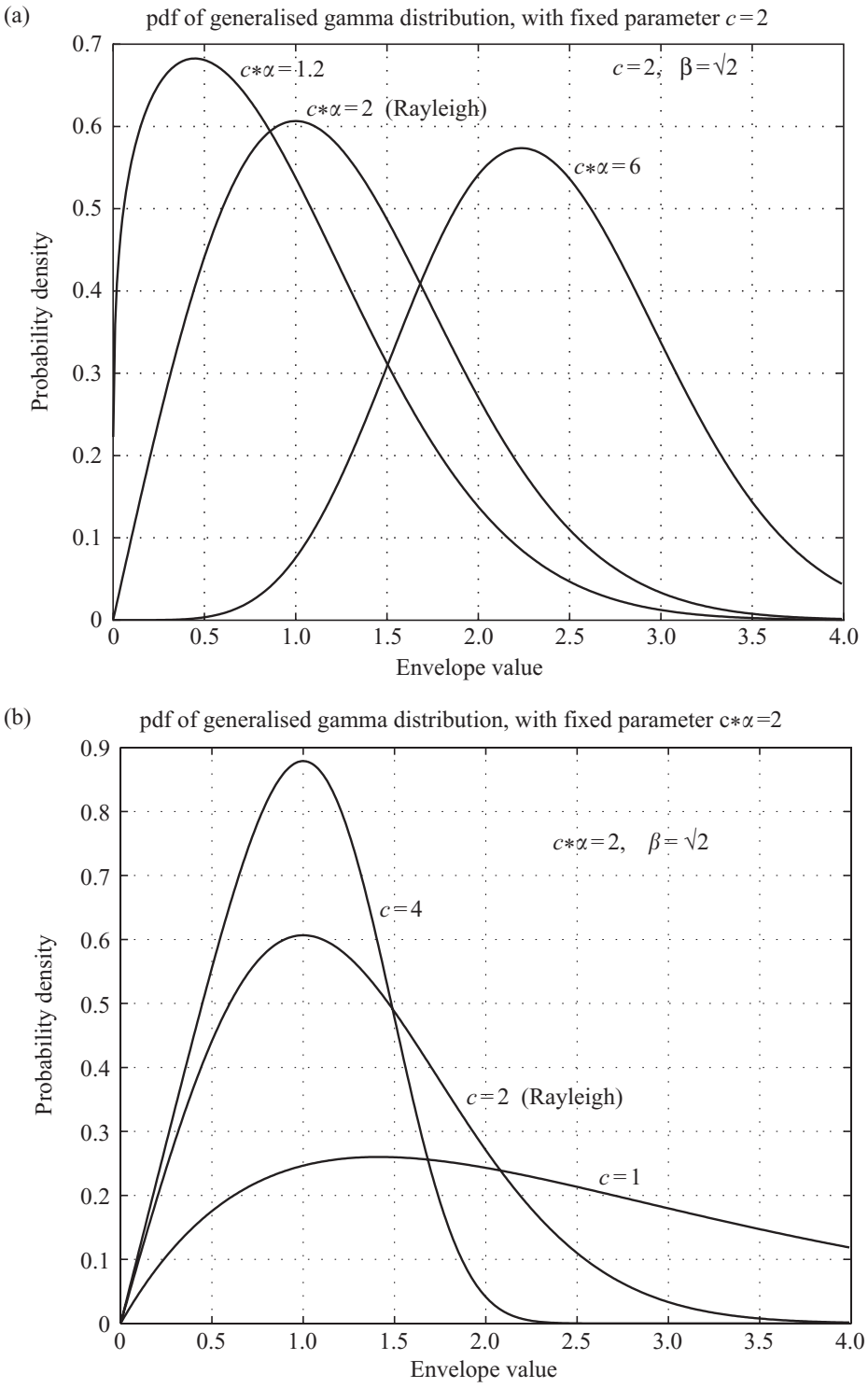


Figure C.12 The generalised gamma distribution showing the effect of the parameters: (a) varying cα with c fixed (lower tail behaviour); (b) varying c with cα fixed

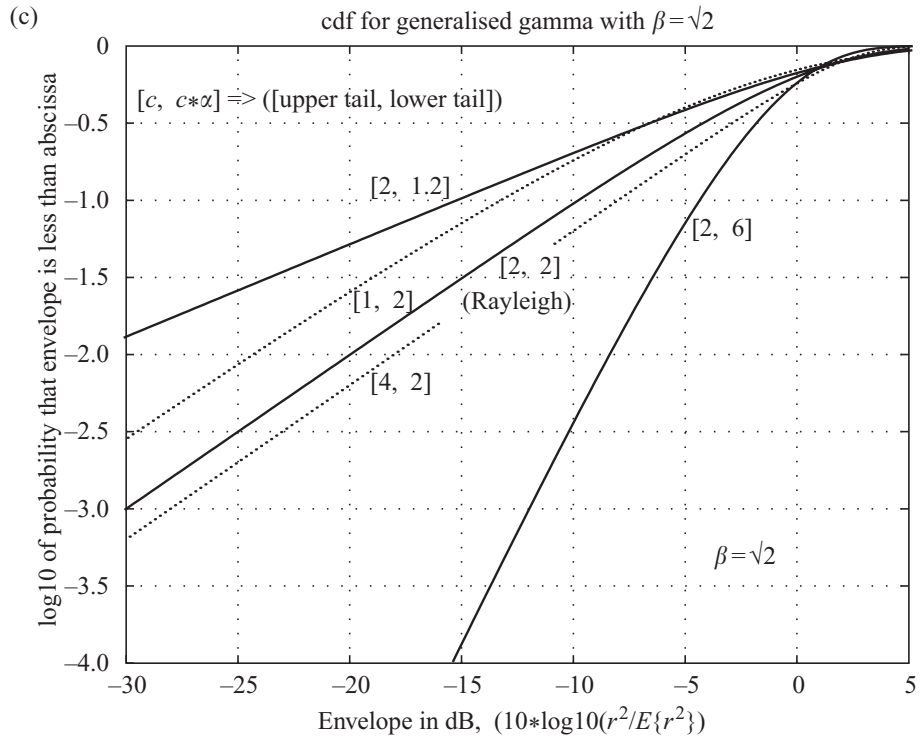


Figure C.12 The generalised gamma distribution showing the effect of the parmeters: (c) The cdf for the same parameters as in (a) and (b) showing the lower tail behaviour

normalised by σ^2. Although the Rice distribution has a different form to the generalised gamma, Nakagami (1960) shows that it is well approximated when the specular power and Rayleigh power are related to the distribution parameters as

$$\overline{r_s^2} = \overline{r^2}\sqrt{1-\frac{1}{\alpha}} \tag{C.111}$$

$$2\sigma^2 = \overline{r^2}\left(1-\sqrt{1-\frac{1}{\alpha}}\right), \tag{C.112}$$

respectively.

The cdf can be written as the incomplete gamma function (see Section C.12),

$$\text{Prob}\,(r < r_0) = P_{\text{GG}}\,(r_0) = \int_0^{r_0} p_{\text{GG}}(r)\,\mathrm{d}r = P\left(\alpha, \left(\frac{r_0}{\beta}\right)^c\right) \tag{C.113}$$

and $P(\alpha, x)$ is the incomplete gamma function given in equation (C.105).

Table C.1 Distributions that the generalised gamma can describe (Griffiths and McGeehan, 1982)

Parameter	Distribution name	pdf	Conditions
$c = 1$	gamma	$Ar^c \exp(-r/\beta)$	
$c = 1$	Erlang	$Ar^c \exp(-r/\beta)$	$(a + 1)$ positive integer
$c = 1$	chi-squared	$Ar^c \exp(-r/\beta)$	$\beta = 2$; $2(a + 1)$ is positive integer
$c = 2$	Nakagami-m	$Ar^c \exp(-r^2/\beta^2)$	
	Stacy	$Ar^c \exp(-r^c/\beta^c)$	
	Weibull	$Ar^{c-1} \exp(-r^c/\beta^c)$	$a = c - 1$
$a = 0$	exponential	$A \exp(-r/\beta)$	$c = 1$
$a = 0$	one-sided normal	$A \exp(-r^2/\beta^2)$	$c = 2$
$a = 0$	generalised exponential	$A \exp(-r^c/\beta^c)$	
$a = 1$	Rayleigh	$Ar \exp(-r^2/\beta^2)$	$c = 2$
$a = 1$	generalised Rayleigh	$Ar \exp(-r^c/\beta^c)$	

C.15 Additive mixtures with non-Gaussianity

A finite mixture density is the weighted sum of different pdfs, to give a mixed, or 'contaminated' pdf. The parameter ϵ is used here to give the amount of non-Gaussian contamination in

$$p_x(x) = (1 - \epsilon) p_x(x) + \epsilon p_{n\mathrm{G}}(x). \tag{C.114}$$

For example, the Gaussian-Laplace mix is (Miller and Thomas, 1976)

$$p_{\mathrm{GL}}(x) = \frac{1 - \epsilon}{\sqrt{2\pi}\sigma} \mathrm{e}^{-x^2/(2\sigma^2)} + \frac{\epsilon}{2\sigma_{\mathrm{L}}^2} \mathrm{e}^{-|x|/\sigma_l^2}, \quad 0 \le \epsilon \le 1. \tag{C.115}$$

By making the process complex, with independent components, the joint pdf of the envelope and phase can be found. For the Gaussian–Laplace, this is (Biyari and Lindsey, 1995)

$$p_{\mathrm{GL}}(r, \theta) = \theta \left(\frac{(1 - \epsilon)^2}{2\pi\sigma^2} u + \frac{\epsilon(1 - \epsilon)}{2\sigma_{\mathrm{L}}^2 \sqrt{2\pi}\sigma} v + \left(\frac{\epsilon}{2\sigma_l^2} \right)^2 w \right), \tag{C.116}$$

where

$$u = \mathrm{e}^{-r^2/(2\sigma^2)}, \tag{C.117}$$

$$v = \exp\left[-\left(\frac{r^2 \cos^2\theta}{2\sigma_{\mathrm{L}}^2} + \frac{|r \sin\theta|}{\sigma_{\mathrm{L}}^2} \right) \right] + \exp\left[-\left(\frac{r^2 \sin^2\theta}{2\sigma_{\mathrm{L}}^2} \times \frac{|r \cos\theta|}{\sigma_{\mathrm{L}}^2} \right) \right], \tag{C.118}$$

$$w = \exp\left[-\left(\frac{|r\cos\theta| + |r\sin\theta|}{\sigma_{\mathrm{L}}^2}\right)\right]. \tag{C.119}$$

The envelope and phase are seen to be dependent.

C.16 Middleton's Class A impulsive noise distribution

The impulsive nature in some noise measurements was modelled by Middleton (1977) by a mixture Gaussian and several interfering waveforms. The phase of the noise is assumed uniform and the envelope is the two-parameter mixture of Rayleighs with pdf

$$p_{\mathrm{M}}(r) = \mathrm{e}^{-A}\sum_{n=0}^{\infty}\frac{A^n r}{n!\sigma_n^2}\,\mathrm{e}^{-(r/\sigma_n)^2/2}, \quad r \geq 0; \tag{C.120}$$

where

$$\sigma_n = \frac{(n/A) + \Gamma_{\mathrm{M}}}{1 + \Gamma_{\mathrm{M}}} \tag{C.121}$$

and here Γ_{M} is a parameter.

A small number of terms usually gives a good approximation to the infinite expansion. The impulsive index, A, is the product of the received number of impulses per unit time and the impulse duration. It defines how non-Gaussian the distribution is, with small A highly impulsive and large A ($\sim$10) Gaussian. Γ_{M} is also a measure of the non-Gaussianity, being the mean ratio of the Gaussian noise component to non-Gaussian noise component. Small Γ_{M} implies large non-Gaussianity.

A final point about this distribution is that the in-phase and quadrature components would normally be well-correlated, rather than independent as assumed in other distributions discussed. This is a characteristic of impulsive noise. The distribution is used instead of purely Gaussian noise in the study of the BERs of various modulation schemes in the presence of impulsive-like noise or interference.

C.17 Diversity distributions

In a diversity system, the combined signal is in general a weighted sum of individual fading signals. To analyse the statistics of the combined signal, the joint distribution of the component signals is required. The joint pdf is then integrated under the appropriate combining rule to produce the pdf of the combined signal. The cornerstones of the statistical treatment are the multivariate Gaussian and the chi-squared distribution. In a way, these are the generalisations of the specific cases above, where two only independent Gaussians were added to give the Rice and Rayleigh distributions for the envelopes. There are few results available for multivariate cases of distributions other than Gaussian. Also, by being able to develop the distributions from the component Gaussians, the phase is included. Some of the previously discussed distributions are envelope-only.

C.18 Multivariate Gaussian

In this section, the N variables and their parameters are subscripted with n. The transpose is denoted with a superscript 'T'. The vector of Gaussian variables is denoted

$$\boldsymbol{X}^{\mathrm{T}} = (x_1, x_2, \ldots, x_N) \tag{C.122}$$

and similarly for the vector of means, $\boldsymbol{m}_x$. The $N \times N$ covariance matrix is symmetric,

$$\boldsymbol{C}_X = E\left\{(\boldsymbol{X} - \boldsymbol{m}_X)(\boldsymbol{X} - \boldsymbol{m}_X)^{\mathrm{T}}\right\} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \ldots & \sigma_1\sigma_N\rho_{1N} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & \ldots & \\ \ldots & \ldots & \ldots & \\ \sigma_1\sigma_N\rho_{1N} & & & \sigma_N^2 \end{bmatrix} \tag{C.123}$$

where ρ_{ij} is the correlation coefficient between the ith and jth variable. The joint pdf is then written as

$$p_X(x_1, x_2, \ldots, x_N) = \frac{1}{\left(\sqrt{2\pi}\right)^N |\boldsymbol{C}_X|^{1/2}} \mathrm{e}^{-(1/2)(\boldsymbol{X}-\boldsymbol{m}_x)^{\mathrm{T}}\boldsymbol{C}_X^{-1}(\boldsymbol{X}-\boldsymbol{m}_x)}. \tag{C.124}$$

For complex Gaussian random variables, $\boldsymbol{Z}^{\mathrm{T}} = (z_1, z_2, \ldots, z_N)$, where $z_i = x_i + \mathrm{j}y_i$, and in most electrical engineering cases we deal with the circular, uncorrelated case which means that the real and imaginary parts have the same mean and variance and are mutually independent. So we have $\boldsymbol{C}_X = \boldsymbol{C}_Y$, $\boldsymbol{C}_{XY} = -\boldsymbol{C}_{YX}$, and the covariance matrix

$$\boldsymbol{C}_Z = \boldsymbol{C}_X + \boldsymbol{C}_Y - \mathrm{j}(\boldsymbol{C}_{XY} - \boldsymbol{C}_{YX}) = 2(\boldsymbol{C}_X - \mathrm{j}\boldsymbol{C}_{XY}) \tag{C.125}$$

is Hermitian. For zero mean variables, i.e. $\boldsymbol{m}_Z = \boldsymbol{0}$, the pdf is written

$$p_Z(\boldsymbol{Z}) = \frac{1}{\pi^N |\boldsymbol{C}_Z|} \mathrm{e}^{-\boldsymbol{Z}^H \boldsymbol{C}_Z^{-1}\boldsymbol{Z}} \tag{C.126}$$

where the complex transpose is denoted with a superscript H.

C.19 Chi-squared

The non-central chi-squared distribution with N degrees of freedom describes the statistics of the sum of the squares (i.e. sum of powers) of N independent Gaussian variables, x_n, denoted

$$s_{Xn}^{(N)} = \sum_1^N x_n^2. \tag{C.127}$$

Each of the variables must have the same variance σ^2 but may have different means m_n. The sum of the squared means (the non-centrality parameter) is similarly denoted

$$s_\mu = \sum_{n=1}^{N} m_n^2 \tag{C.128}$$

and the pdf is then written as

$$p_{s_{Xn}}^{(N)}(s) = \frac{1}{2\sigma^2}\left(\frac{s}{s_\mu}\right)^{(n-2)/4} \mathrm{e}^{-(s+s_\mu)/2\sigma^2}\, I_{\left(\frac{N}{2}-1\right)}\left(\frac{\sqrt{s s_\mu}}{\sigma^2}\right), \quad s \geq 0. \tag{C.129}$$

The mean, mean square and variance are

$$E\left\{s_{Xn}^{(N)}\right\} = N\sigma^2 + s_\mu, \tag{C.130}$$

$$E\left\{\left(s_{Xn}^{(N)}\right)^2\right\} = 2\sigma^2\left(N\sigma^2 + 2s_\mu\right) + \left(N\sigma^2 + s_\mu\right)^2, \tag{C.131}$$

$$\sigma_{s_{Xn}^{(N)}}{}^2 = 2\sigma^2\left(N\sigma^2 + 2s_\mu\right), \tag{C.132}$$

respectively. For even N, the integral of the pdf is written:

$$P_{s_{X_n}^{(N)}}(s) = P_{s_{X_n}^{(N)}}(s \leq s_0) = 1 - Q_m\left(\frac{\sqrt{s_\mu}}{\sigma}, \frac{\sqrt{s_0}}{\sigma}\right) \tag{C.133}$$

where $m = N/2$ is an integer. $Q_m(a, b)$ is the generalised Marcum Q function,

$$Q_m(a, b) = \int_b^\infty x\left(\frac{x}{a}\right)^{m-1} \mathrm{e}^{-(x^2+a^2)/2} I_{m-1}(ax)\, \mathrm{d}x \tag{C.134}$$

which has the identities for $a, b > 0$;

$$Q_m(a, b) = Q_1(a, b) + \mathrm{e}^{-(a^2+b^2)/2} \sum_{n=1}^{m-1}\left(\frac{b}{a}\right)^n I_n(ab) \tag{C.135a}$$

$$= \begin{cases} \mathrm{e}^{-(a-b)^2/2} \sum\limits_{n=1-m}^{\infty} \left(\frac{a}{b}\right)^n \mathrm{e}^{-ab} I_n(ab) & \text{if } a < b \quad \text{(C.135b)} \\ 1 - \mathrm{e}^{-(a-b)^2/2} \sum\limits_{n=m}^{\infty} \left(\frac{b}{a}\right)^n \mathrm{e}^{-ab} I_n(ab) & \text{if } a \geq b. \quad \text{(C.135c)} \end{cases}$$

Limiting values are recalled as $Q_m(0, 0) = 1$, $Q_m(a, 0) = 1$, $Q_m(0, \infty) = 0$, $Q_m(\infty, 0) = 1$, $Q_m(\infty, \infty) = 0$,

$$Q_m(0, b) = \mathrm{e}^{-b^2/2} \sum_{k=0}^{m-1} \frac{1}{k!}\left(\frac{b^2}{2}\right)^k \tag{C.136}$$

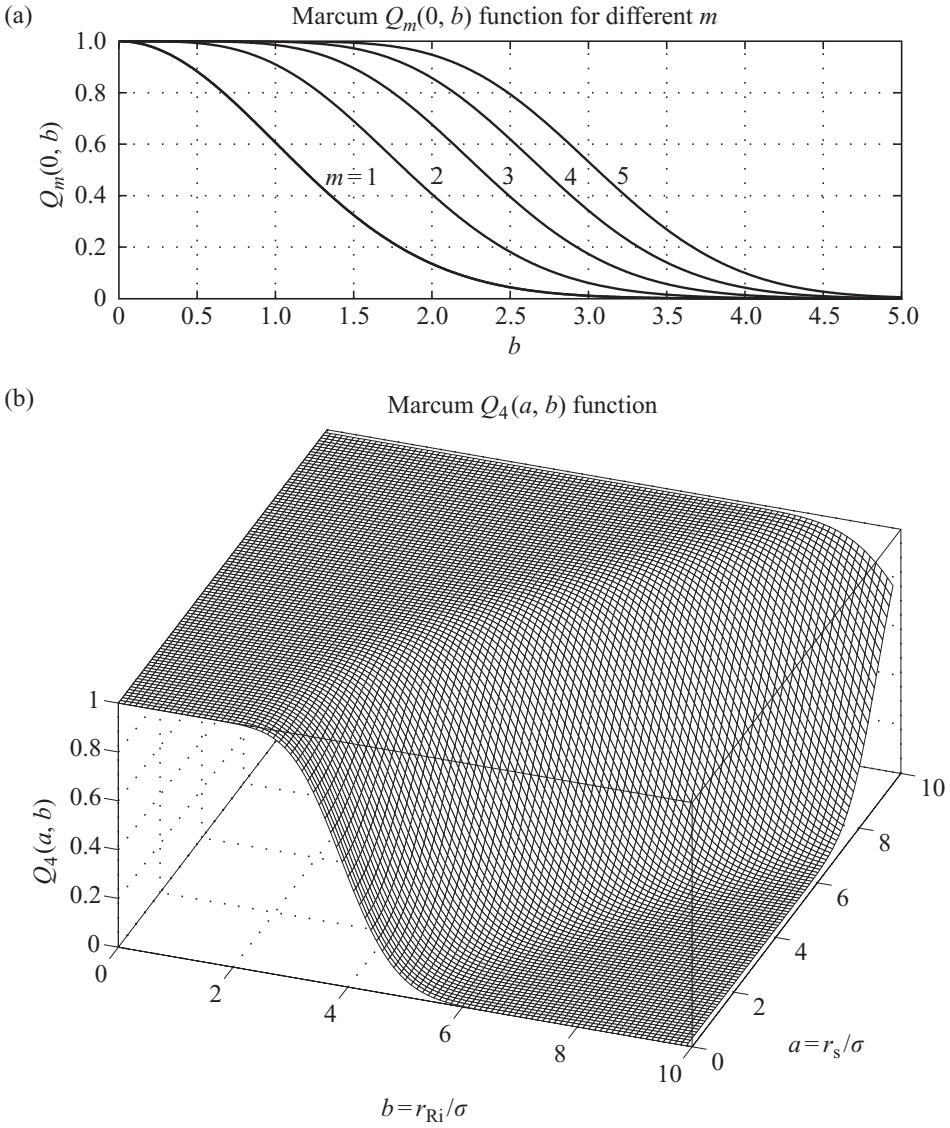
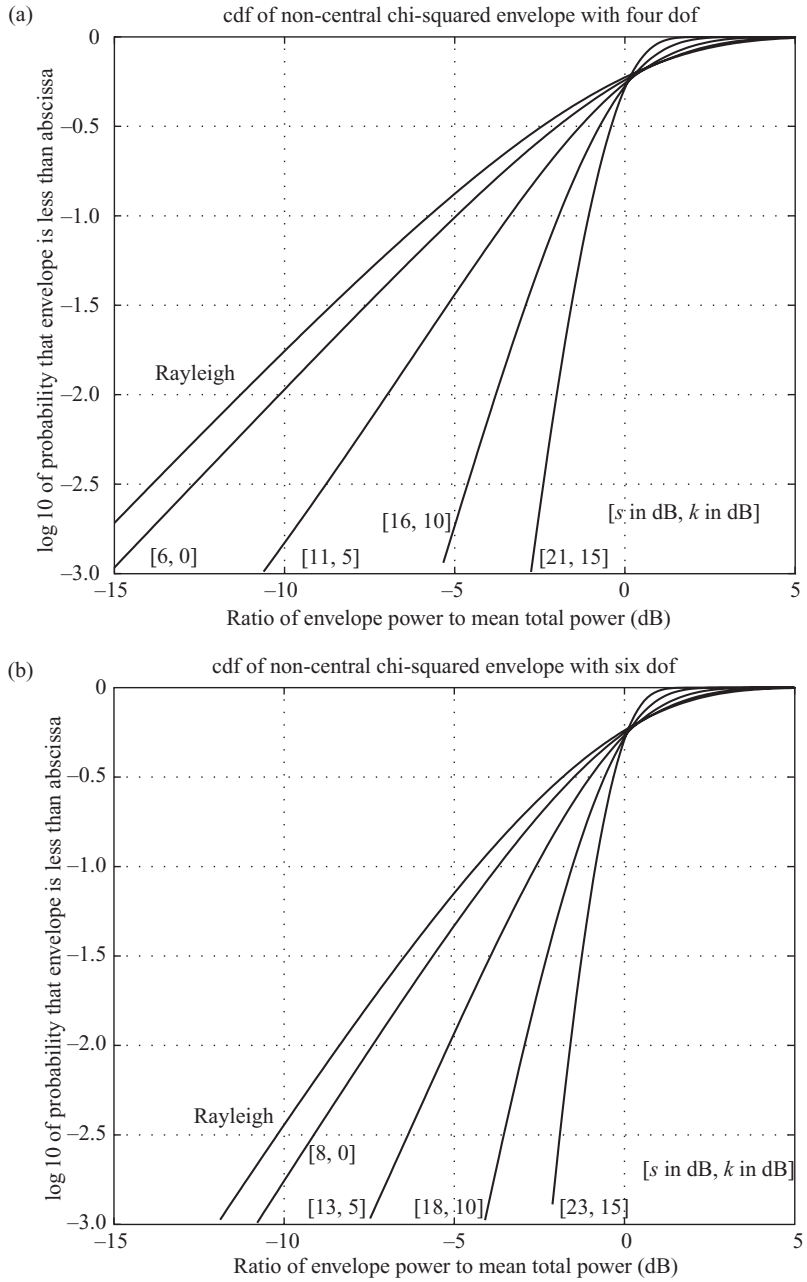


Figure C.13 *(a) The function $Q_m(0, b)$ for several m. The function is the probability that the sum of the powers of m uncorrelated Rayleigh envelopes is greater than b. (b) The generalised $Q(a, b)$ function is the probability that a sum-of-powers envelope of m Ricean signals with a sum-of-powers specular envelope $a = \sqrt{s_\mu}/\sigma$ is greater than b, i.e. $Q_m(a, b) = Q_m(\sqrt{s_\mu}/\sigma,\ \sqrt{s_0}/\sigma) = \text{Prob}\ (r_{Ri} > r_0)$*

which represents the main behavioural difference between the Q_m and is shown in Figure C.13(a). The example $Q_4(a, b)$ is given in Figure C.13(b).

For N odd, the pdf appears to need numeric integration for the cdf. For the Rice distribution, $N = 2$ and $s_\mu = k_{Rice} 2\sigma^2$, and the distribution equations here reduce to those from the Rice section above. Some cdfs are plotted in Figure C.14.



Figures C.14 *The cdf for the chi-squared distribution with N degrees of freedom (summation of N independent Gaussian variables). (a) For $N = 4$, corresponding to two uncorrelated diversity channels; (b) $N = 6$, three uncorrelated diversity channels. The abscissa is normalised by the total power, i.e. $N + s_m$, where s_μ is the sum of the squared specular components (the non-centrality parameter) and is rounded to an integer value for the plot labels.The plots can be interpreted as N/2 uncorrelated diversity channels, with each channel having a Rice factor of $k_{Rice} = s_\mu / N$. The signal combination corresponds to predetection maximum ratio*

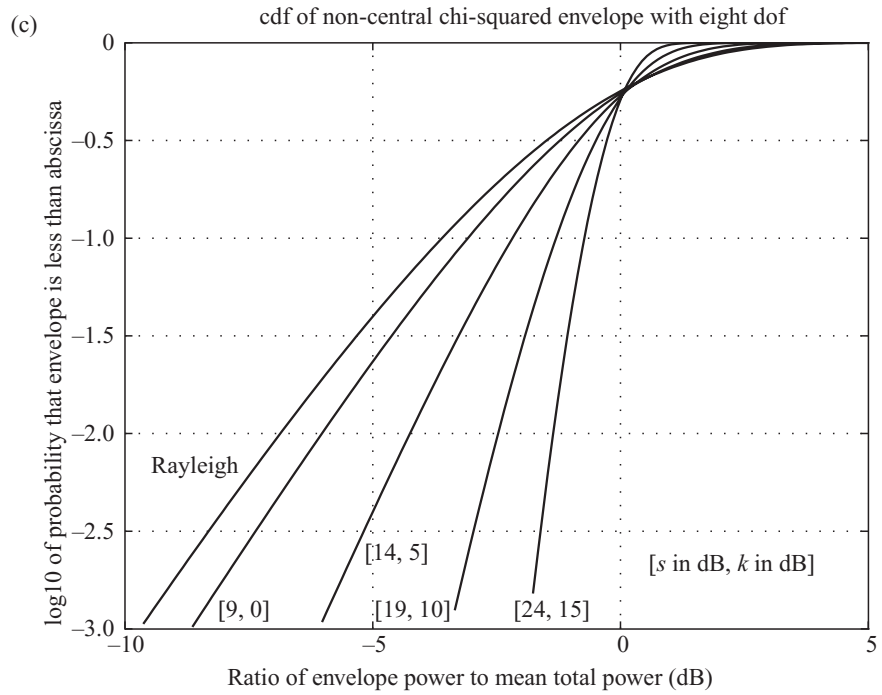


Figure C.14 *(c) As for (a) and (b), but with $N = 8$, four uncorrelated diversity channels*

For the case where the component Gaussians are all zero mean, the sum of the means squared is zero, i.e. $s_\mu = 0$ and the pdf reduces to the (central) chi-squared form

$$p_{s_X^{(N)}}(s) = \frac{1}{(2\sigma^2)^{N/2}\Gamma\left(\frac{N}{2}\right)} s^{(N/2-1)} \mathrm{e}^{-s/2\sigma^2}, \quad s \geq 0, \tag{C.137}$$

with the moments

$$E\{s_X^{(N)}\} = N\sigma^2, \tag{C.138}$$

$$E\{s_X^{(N)^2}\} = (2N + N^2)\sigma^4, \tag{C.139}$$

$$\sigma_{s_X^{(N)}}{}^2 = 2N\sigma^4, \tag{C.140}$$

respectively. The Rayleigh envelope (exponential in s) case is for $N = 2$. The cdf can be expressed as an incomplete gamma function (see section C.12). For N even, however, the cdf is

$$\text{Prob}\,(s \leq s_0) = P_{s_X^{(N)}}(s_0) = 1 - \mathrm{e}^{-s_0/2\sigma^2} \sum_{k=0}^{\frac{N}{2}-1} \frac{1}{k!} \left(\frac{s_0}{2\sigma^2}\right)^k. \tag{C.141}$$

The Rayleigh cdf cases are included in the Rice cdfs of Figure C.14.

C.20 Maximum ratio combination of Rayleigh envelopes

A benchmark for signal combining is maximum ratio, because the resulting SNR, γ_{R}, is the sum of the component SNRs,

$$\gamma_{\mathrm{R}} = \sum_{n=1}^{N} \gamma_n. \tag{C.141}$$

To obtain this result, the signals are weighted by their complex conjugate of the signal voltage (this aligns the component signal phases) and the inverse of the noise power, which emphasises the components with large SNRs.

For N branches with mean powers R_n, the pdf of the resultant SNR is

$$p_{\mathrm{MR}}(\gamma_{\mathrm{R}}) = \sum_{n=1}^{N} \frac{1}{\xi_n} \mathrm{e}^{-\gamma_{\mathrm{R}}/R_n}. \tag{C.142}$$

where

$$\xi_n = R_n \prod_{\substack{m=1 \\ m \neq n}}^{N} \left(1 - \frac{R_m}{R_n}\right). \tag{C.143}$$

If the branches all have the same mean powers, $R = R_1 = R_2 = \ldots = R_N$, then the pdf reduces to the chi-square distribution (Jakes, 1974), i.e.

$$p_{\mathrm{MR}}(\gamma_{\mathrm{R}}) = \frac{\gamma_{\mathrm{R}}^{N-1} \mathrm{e}^{-\gamma_{\mathrm{R}}/R}}{R^N (N-1)!} \tag{C.144}$$

and as noted above, this can be couched as the Nakagami-m distribution in voltage ratio, w, by the substitution $w = \sqrt{\gamma_R}$. The mean SNR of the combined signal is simply the sum of the contributing SNRs.

$$\begin{aligned} E\{\gamma_{\mathrm{R}}\} = \sum_{n=1}^{N} E\{\gamma_n\} &= \sum_{n=1}^{N} R_n \\ &= NR \quad \text{if } R = R_1 = R_2 \cdots = R_N. \end{aligned} \tag{C.145}$$

This simple result also is the best possible resultant SNR from linear signal combination, which is why it is a benchmark.

C.21 References

ABRAMOWITZ, M. and STEGUN, I.A.: 'Handbook of mathematical functions', Dover, 1965.

ANG, A. H.-S. and TANG, W. H.: *Probability Concepts in Engineering Planning and Design*, **1**, Wiley, 1975.

BECKMANN, P.: 'The probability distribution of a vector sum of n unit vectors with arbitrary phase distributions', *Acta Technica, CSAV*, 1959, (4), pp. 323–335.

BECKMANN, P.: *Probability in Communication Engineering*, Harcourt, Brace and World, Colorado, 1967.

BECKMANN, P. and SPIZZICHINO, A.: *Electromagnetic Waves*, No. 4, Pergamon Press, 1963. Later published as *The Scattering of Electromagnetic Waves from Rough Surfaces*, Artech House, 1987.

BIYARI, K.H. and LINDSEY, W.C.: 'Error performance of DPSK mobile communication systems over non-Rayleigh fading channels', *IEEE Transactions*, May 1995, **VT-44** (2), pp. 211–219.

COULSON, A.J., WILLIAMSON, A.G., and VAUGHAN, R.G.: 'Improved fading distribution for mobile radio', *IEE Proceedings-Communications*, June 1998, **145** (3), pp. 197–202.

GRIFFITHS, J. and McGEEHAN, J.: 'Interrelationship between some statistical distributions used in radio-wave propagation', *IEE Proceedings*, December 1982, **129** Part F, (6), pp. 411–417.

HOYT, R.S.: 'Probability functions for the modulus and angle of the normal complex variate', *Bell System Technical Journal*, 1947, **26**, pp. 318–359.

JAKES, W.C. (Ed.): *Mobile Microwave Communications*, AT&T, 1974.

MARCUM, J.I.: 'A statistical theory of target detection by pulsed radar', *IRE Transactions*, April 1960, **IT-6**, pp. 59–267.

MIDDLETON, D.: 'Statistical-physical models of electromagnetic interference', *IEEE Transactions*, August 1977, **EMC-19**, pp. 106–127.

MILLER, J. and THOMAS, J.: 'The detection of signals in impulsive noise modelled as a mixture process', *IEEE Transactions*, May 1976, **COM-24**, pp. 559–563.

NAKAGAMI, M.: 'The m-distribution – a general formula of intensity distribution of rapid fading', in W.C. Hoffman (ed.), *Statistical methods in radio wave propagation*, Pergamon, Oxford, 1960.

PAPOULIS, A.: *Probability, Random Variables, and Stochastic Processes*, 3rd edition, McGraw-Hill, New York, 1991.

PARL, S.: 'A new method for calculating the generalized Q function', *IEEE Transactions*, January 1980, **IT-26**, pp. 121–124.

PROAKIS, J.G.: *Digital Communications*, McGraw-Hill, New York, 1983.

RICE, S.O.: 'Mathematical analysis of random noise', *Bell System Technical Journal*, 1944, (3), and 1945, (1).

STACY, E.W.: 'A generalisation of the gamma function', *Annals of Mathematical Statistics*, 1962, **33**, pp. 1187–1192.

STACY, E.W. and MIHRAM, G.A.: 'Parameter estimation for a generalized gamma distribution', *Tecnometrics*, August 1965, **7** (3), pp. 349–358.

STEIN, S.: Part III of Schwartz, M., Bennett, W.R., and Stein, S., *Communications Systems and Techniques*, McGraw-Hill, 1966.

SUZUKI, H., 1977: 'A statistical model for urban radio propagation', *IEEE Transactions*, July 1977, **Com-25** (7), pp. 673–680.

Appendix D

Fresnel zones

In the context of Huygens' principle and physical optics, it makes sense to consider a volume over which the phase of contributing waves is fairly constant, centred around the line-of-sight where the phase is stationary. Consider the situation in Figure D.1 where the distance between the two antennas is d and the total distance to a scattering point is $R_1 + R_2$.

Moving on a line perpendicular to the line connecting the two antennas the phase difference increases, and we define the extent of the 'zone of influence' by the points where the excess path length equals $\lambda/2$,

$$R_1 + R_2 = d + \frac{\lambda}{2}. \tag{D.1}$$

The higher-order Fresnel zones are defined by using multiple values of $\lambda/2$. Making a transverse cut through the ellipse it is clear that the phase will be fairly constant inside the first Fresnel zone, but that contributions from the higher order zones will

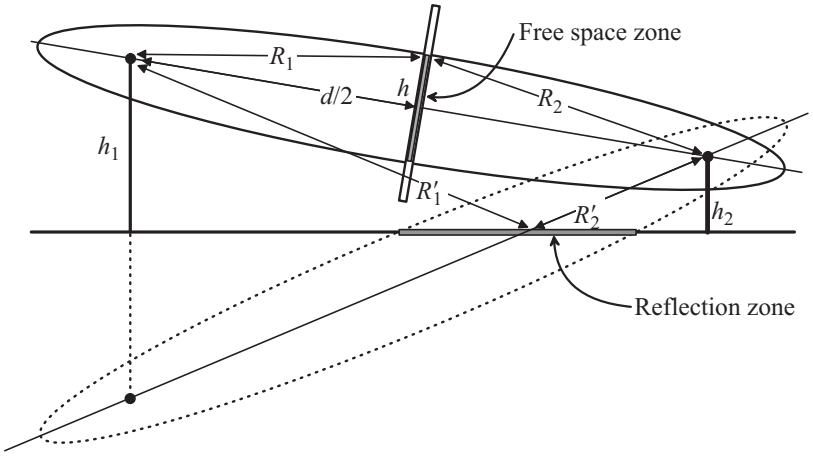


Figure D.1 Definition of the first Fresnel zone for the free space path and for the reflected path. The dotted ellipse is for the reflected path

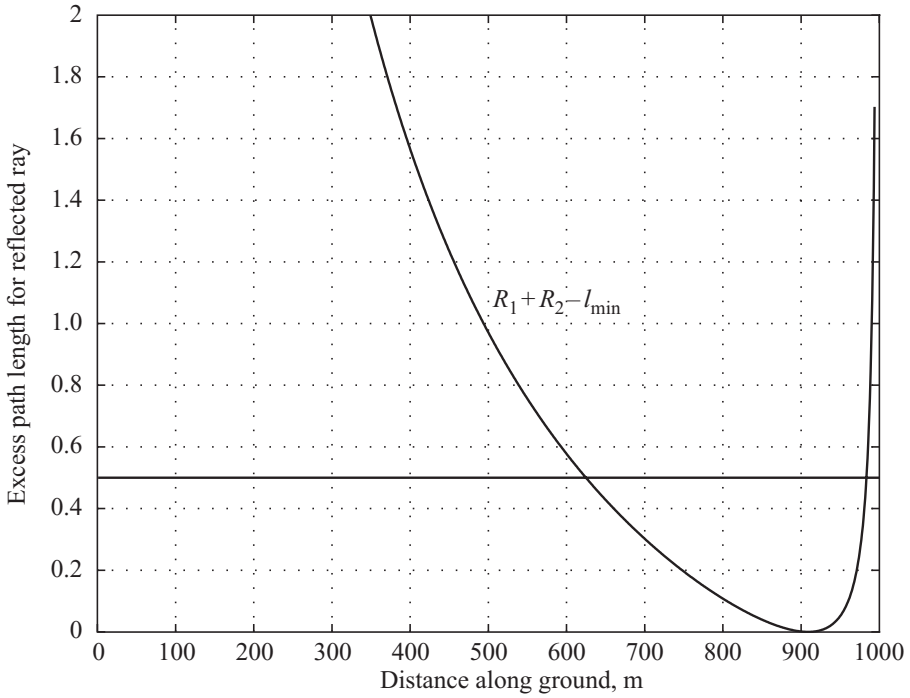


Figure D.2 *Excess path length in wavelengths along ground for $h_1 = 20\,m$, $h_2 = 2\,m$, $d = 1000\,m$, and $f = 1800\,MHz$. The cut with the line $y = 0.5$ defines the boundaries of the first Fresnel zone.*

tend to cancel due to the phase shifts. This notion gives rise to the first Fresnel zone corresponding to a zone of influence (see also Section 3.1 on reflection from a smooth surface, Section 3.2 on rough surfaces and the discussion in Section 3.3.2 on transition zones).

It is useful to consider the width of the Fresnel zone at the mid-point of the ellipse, $2h$, which for cases where the distance d is much larger than the wavelength is given by

$$h = \sqrt{\frac{d\lambda}{2}}. \tag{D.2}$$

For the example of $d = 1000\,\text{m}$ and $f = 1800\,\text{MHz}$, $h = 9\,\text{m}$.

The concept of a Fresnel zone may also be applied to the reflected wave from a smooth surface (Figure D.1). The ellipse defining the Fresnel zone has one antenna and the image of the other antenna as the two foci. Where the ellipsoid cuts the groundplane, a planar zone of influence may be defined, the reflection zone.

An example of the excess path length along the surface around the reflection point is shown in Figure D.2. Parameters are $h_1 = 20\,\text{m}$, $h_2 = 2\,\text{m}$, $f = 800\,\text{MHz}$, and $d = 1000\,\text{m}$. The reflection point is situated at $x = 909\,\text{m}$, close to the low antenna. If again $\lambda/2$ is chosen as the critical excess path length we can read the extent of the Fresnel zone off Figure D.2. The zone is highly asymmetric around the reflection point extending several hundred metres towards the high antenna. The total width of the zone is 360 m.

Appendix E

Group delay equivalence in the time and frequency domains

This appendix shows the equivalence between the time and frequency domain definitions for delay, and is associated with Chapter 5. The time domain expression for the mean excess delay is the energy-weighted delay. This is shown to be the same as the frequency domain energy-weighted group delay. The normalisation for both expressions is the total energy

$$E = \int_0^\infty |h(t)|^2 \, \mathrm{d}t \tag{E.1a}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(\omega)|^2 \, \mathrm{d}\omega \tag{E.1b}$$

where the equivalence is by Parseval's theorem:

$$\int x(t) y^*(t) \, \mathrm{d}t = \frac{1}{2\pi} \int X(\omega) Y^*(\omega) \, \mathrm{d}\omega. \tag{E.2}$$

The steps to establish equivalence are as follows. The time domain mean excess delay is

$$\bar{\tau} = \frac{1}{E} \int_0^\infty t|h(t)|^2 \, \mathrm{d}t. \tag{E.3}$$

This is re-written as

$$\bar{\tau} = \frac{1}{E} \int_0^\infty t h(t) h^*(t) \, \mathrm{d}t. \tag{E.4}$$

and putting

$$x(t) = th(t) \quad \Leftrightarrow \quad X(\omega) = \frac{1}{-\mathrm{j}} \frac{\mathrm{d}H(\omega)}{\mathrm{d}\omega} \tag{E.5}$$

and

$$y^*(t) = h^*(t), \tag{E.6}$$

into equation (E.2), there results

$$\bar{\tau} = \frac{1}{2\pi E} \int_{-\infty}^{\infty} \frac{1}{-\mathrm{j}} \frac{\mathrm{d}H(\omega)}{\mathrm{d}\omega} H^*(\omega)\, \mathrm{d}\omega. \tag{E.7}$$

Substituting the channel transfer function

$$H(\omega) = |H(\omega)|\mathrm{e}^{\mathrm{j}\phi(\omega)} \tag{E.8}$$

into equation (E.6) gives

$$\begin{aligned} \bar{\tau} &= \frac{1}{2\pi E} \int_{-\infty}^{\infty} \frac{1}{-\mathrm{j}} \frac{\mathrm{d}}{\mathrm{d}\omega} (|H(\omega)|\mathrm{e}^{\mathrm{j}\phi(\omega)}) |H(\omega)|\mathrm{e}^{-\mathrm{j}\phi(\omega)}\, \mathrm{d}\omega \\ &= \frac{1}{2\pi E} \int_{-\infty}^{\infty} \left(\left(\frac{1}{-\mathrm{j}} |H(\omega)| \frac{\mathrm{d}}{\mathrm{d}\omega} |H(\omega)| \right) - \left(|H(\omega)|^2 \frac{\mathrm{d}\phi(\omega)}{\mathrm{d}\omega} \right) \right) \mathrm{d}\omega. \end{aligned} \tag{E.9}$$

The first term is zero since the right-hand side is real, so

$$\bar{\tau} = \frac{1}{2\pi E} \int |H(\omega)|^2 \left(-\frac{\mathrm{d}}{\mathrm{d}\omega} \phi(\omega) \right) \mathrm{d}\omega \tag{E.10}$$

i.e.

$$\bar{\tau} = \frac{1}{2\pi E} \int |H(\omega)|^2\, \tau_g(\omega)\, \mathrm{d}\omega. \tag{E.11}$$

E.1 Reference

PAPOULIS, A.: *Signal Analysis*, McGraw-Hill, 1977.

Index

absorbing half-plane 132, 134
absorbing screen 123–4
absorption 533, 535
absorption density 533
ad hoc networks 13
adaptive antennas 631–42
additive mean 416–17
additive mixtures with non-Gaussianity 726–7
additive noise 662
adjacent channel interference 20–1
Akaike criterion 394
ambiguity function 213
amplitude dispersion 223
angle diversity 40, 578–80
angle of arrival 177, 268–71
  instantaneous 269–71
angular distribution 381–2
angular domain 173–8
angular power spectrum 255
angular spread 174–6, 178, 255, 630
angularly displaced beams 581–5
antenna beams 435–43, 577, 581, 583–5
antenna branches 45–6
antenna diversity 38, 40–1, 48
antenna effective cross section 683
antenna efficiency 438, 446–8
antenna efficiency factors 443
antenna elements 435
antenna impedance 460–1
antenna impedance mismatch 444
antenna matching 495, 510–13, 595
antenna parameters 435
antenna patterns 71–2, 436, 575–7, 593–4
  diversity 577–8
antenna principles 435
antenna radiation efficiency 447–8
antenna synthesis 508–9
antenna temperature 452–4, 456
antenna with transmission line 455–6, 458
antenna–environment interaction 537
antennas on a handset 532–8
  near a person 533–6
  random environment 535–8
antennas with circular polarisation 538–55
  patches 546–8
  scanning-mode helix 550–5
array antennas 629–31
  multi-port 630–1
  one-port 629–30
array directivity 588–9
  broadside 589
  endfire 589
array weights 41–3, 64, 632–46, 661–2
autocorrelation 75–8, 415–18, 560, 687, 691
  window size effect 416
autocorrelation function 203, 212–13, 283, 687
  envelope 406
autocovariance 415
autocovariance function 687
automatic gain control 18
auxiliary potential functions 470–1
average power density 448
averaged energy stored 502
averaged scenario models 300–2
averaging 80, 341–2, 661
axial ratio 541–3

Babinet's principle 475–6
backfire radiation 552
bandpass signal representation 32
bandwidth 188
  dynamic 228

bandwidth (*cont.*)
instantaneous 214
loaded 504
single resonance 503–4
unloaded 504
bandwidth improvement factors 511
base stations 315–17, 320, 622
polarisation diversity 622
baseband frequency 330
baseband transfer function 67
basic linear estimator 392
BCH coding scheme 30
beam misalignment 441–3
beam mode 639, 649
Beckmann distribution 714
Bello functions 247–8
biological effects 13–14
bit-error ratio 34–6, 51–5, 659
fading effect 54–5
irreducible 34–7
modulation 51–2
path loss 52
bit-error ratio bound 670
bluetooth systems 13
Booker's relation 475–8, 524
Box–Muller algorithm 360
Box–Muller technique 359–60
breakpoint 100
Brewster angle 93–7
bridged knife edge 154–5
broadband 39–40
buildings 163–5
burst errors 30
Butler matrix 631

capacity xxix, 29, 652–5, 658–9, 670–9
asymptotic 675–7
cdf 672–7
M transmit antennas and one receive antenna 653
(M,N) array 653–4
mean 653–5, 678–9
statistical model 672
unknown channel 671
variance 675–6
capacity outage 669–71
carrier frequency 650
Cartesian-spherical coordinate transformations 479
cascade circuits 458–9
causal impulse response 371
causality 371, 374
cavity model 520–2, 527
cavity-backed slot antenna 477
CCIR 10–12, 24
CDMA 22–3, 45, 47, 631–2
cells 14–17
hexagonal 15
pico 15
cellular layout 14–17, 21
cellular networks 631
central limit theorem 289, 303, 697
product giving lognormal distribution 715
channel coding 30
channel correlation functions 75–9
channel efficiency 31–3
channel function generation 359–70
channel gain 83, 254
channel sounding 419–28
full 420–3
partial 422–4
chi-square test 695
chi-squared distribution 254, 721, 728–32
cdf 731–2
pdf 236–7
chicken wire antenna 5–7
chirp 420–3
complex 420
linear FM 420, 423
circuits 457–9
cascade 458–9
circular array 569–70
circular coordinates 581
circular function beams 583–5
circular patch 546–8, 615–16, 618
circular polarisation 538–44
co-polar 541, 548
cross-polar 541, 548
left-hand 539, 543
reflections from a smooth surface 539
right-hand 539, 543
circular polarisation ratio 541–3
Clarke scenarios 262–3, 267, 273, 293, 297–9, 579, 601
simulation 327, 334, 338
Clarke's model 313–14, 578
clipped envelopes 407–14
'Clipper' antenna 6–7
clipping 408–14
hard 410–14
soft 409–10
'clock time' 38, 247
closely spaced elements 588, 605–6
clothesline antenna 3–5
cluster of scatterers 268

clutter correction factor 158
co-channel interference 15–20, 53
  multi-user scheme 53
co-polar gain 438
coded modulation 31
coding 30
coherence bandwidth 46, 203–6, 280
coherence bandwidth-delay spread product 205–8
coherent component 705
coherent field 113–18
combined signals
  signal statistics 564–71
combining algorithms 558–60
  equal gain combining 558–9
  maximum ratio combining 558–9
  selection 558–9
  switching 558
communications systems 9–14
compact antennas 501
complementary antenna elements 477
complementary error function 699
complementary normal function 699
complex correlation 344–5
complex envelope 448
complex height 436, 449–50
complex normal 83, 330, 347, 350, 359, 406, 661, 697
complex weights 42–3, 64, 632–46
complex zero-mean Gaussian process 697–8, 702, 704
compound distribution 719
conditioning theorem 254
confidence interval 290
confluent hypergeometric function 707
convolution 421, 688
coordinate system transformations 479–80
correlated branches 568–71
  three-branch 569–71
  two-branch 568–9
correlated noise 571–4
correlated signal generation 364–5
correlation 654–9, 687
correlation analysis 400–19
  envelope correlations 403–14
correlation bandwidth 271
correlation coefficient 84–7, 276–8, 283, 349, 415, 575, 577, 621, 686
  bivariate normally distributed samples 401
  confidence limits 403–4
  envelope 86, 350
  estimation 87, 400–3
  hard clipping 410–13
  independent samples 401–3
  phase 351–2
  power 85–6
  sloping monopoles 612–14
  soft clipping 409–10
correlation coefficient function 337
  loaded circuit envelope 602–3, 605
correlation distance 271–2
  directional antennas or scenarios 274–80
correlation distance–Doppler spread product 271
correlation frequency 280
correlation function 81–7, 212, 687–9
  angularly displaced beams 581–5
  envelope 83–4
  spatial 275
  surface height 115
correlation function spread 272–3
correlation length 115–16
correlation matrix 362–4, 568, 636
  estimation 393–4
  loaded circuit 595
  open circuit 595
correlation over a bandwidth 311
correlation spacing 271
correlation-ergodic processes 688
COST 13
  COST207 169–71
  COST231 model 154, 156–7, 165
covariance 686
covariance function 286, 687
  backward direction calculations 394
  forward direction calculations 394
covariance matrix 253–4, 348, 352–3, 661–2
  interference-plus-noise-signals 663
covariance spectrum 691
Cramér–Rao bound 692–4
cross-correlation coefficient 329, 560
cross-correlation coefficient function
  spatial envelope 604
cross-correlations 690–1
cross-polar discrimination 301–2, 537, 576–7
cumulative density function 193–4, 238, 570, 685, 696
  asymptotic expansion 674–5
  group delay 256–7
  maximum ratio combining 566
  selection combining 566
  student-T 256
  two-path model 197

cumulative distribution function 665–9
  Gaussian process 698
cumulative probability function 193, 685
curl 498–500
cyclostationarity 333, 335–8
cylindrical source coordinates 468–9, 479–80

data rate 649–50
decibel units 716, 718
deconvolution 424–8
  inverse filter 426–7
  subtractive 426
degrees of freedom 41–3, 46
del operator 498
delay profile
  continuous 369
  discrete 369–70
delay spectrum 81
delay spread 82, 168, 170, 183–4, 208, 222, 379–81
  finite pulse width effect 188–93
  instantaneous 184, 190–2, 256, 427–8
  local 209, 214–18
  path loss effect 338–9
  two-path model 192–3, 262
delay time 38, 211
  mean 183
delay time moments 183–4
demeaning of envelopes 414–19
  additive mean 416–17
  multiplicative mean 417–19
depolarisation 121
deterministic channel 688–9
dielectric constant 90
differential operators 498
differentiation of scalars and vectors 498
diffracted field 124, 126–7, 129, 132
diffraction 123–42
  transition zone 133–42
  UTD 125–42
diffraction coefficient 127–8, 135–6, 140
diffraction models
  two-dimensional 151–4
diffuse component 705
diffuse field 113–14, 118–23
  surface roughness effect 118–23
diffuse scattering 144–5
digital channel 26–8, 33–7
  fading 35–7
digital link outage probability 659
digital wireless link 48, 50
  example 57–9
dipole impedance 483
dipoles 461–5
  folded 462–3
  radiation resistance 480–8
  side-by-side 591–2
  spaced 592
  thin wire 465, 482, 484
direction of arrival 636–7
directional antennas 274–5
directional gain 436
directional scenarios 274, 279
directive gain 438
directivity 435–8, 449, 523, 528
  distributed 437–8, 441–3
  Gaussian beam 441
  linear array factor 588–9
  pencil beam formula 439
  uniform line source 474
dispersion 39, 209–10
  local 220
  two-path model 222–8
dispersion metrics 209
distortion 209, 226–8
distributed directivity 437–8, 441–3
  mobile to geostationary satellites 548–50
  sloping monopole 609–10
divergence (div) 498–9
divergence theorem 499
diversity 38–48, 655–9
  antenna 38, 40–1, 554
  macro-diversity 556
  micro-diversity 556
diversity action 48
diversity algorithm performance 557–8
diversity antennas 38, 40–1, 554, 575
  examples 600
  network model 594–5
  polarisation matrix 575
diversity array 590
diversity combined signals 354, 355
diversity conditions 573–7
diversity distributions 727
diversity gain 38, 40, 45, 556–8, 567–8
  equivalent number of ideal branches 604–6
  sloping monopoles 613, 615
diversity order 649
Dolph–Chebyshev taper 514
Doppler frequency 230, 252, 255, 330, 336
  excursions 258
  moments 233, 263–7, 346–7
  pdf 345
Doppler profile 81, 233, 267, 274

Doppler shift 67, 70, 73, 265
Doppler spectrum 81, 250–1, 265, 345–6
Doppler spread 82, 233, 263, 265–6, 357, 578
downlink with unknown channel 639, 642
DPSK 659
dual quantities for electric and magnetic sources 479
duplex techniques 23–4
duplexer action 618
dynamic bandwidth 228

Edgeworth expansions 674
effective area 449, 451
effective gain 452
effective height 64, 451, 495
effective permeability 496
effective permittivity 117
effective reflection coefficient 115
effective scatterers 65–6, 70, 308, 386–9
  distribution 68–9, 71–3, 75, 78–9
  power profiles 80–1
  resolvability 69–70
effective scattering distribution 253, 293–4
  2D 73–4
  3D 71–2
  Clarke scenario 298
  line source 294–7
  single source 294
  uniform scenario 300
eigenvalues 644–7
  excitation 646
electric dipole 507
  infinitesimal 507
electric energy 502
electric vector potential 521–2
electrically small antennas 506
electromagnetic signal channel 26–8
elliptic integral 350
elliptic polarisation 63–4, 436, 444, 539
elliptical arc length 324
elliptical coordinates 167
elliptical model 322–5
  directionality 322–3
elliptical relations 321–2
energy density antenna 585, 587–8
energy efficiency 424–5
energy per bit 32–3
ensemble averaged distribution 245–6
envelope 343
  clipped 407–14
  demeaning 414–19
  mean 701
  median 343
  Rayleigh distributed 408, 411
  single signal 349
envelope autocorrelation function 406
envelope correlation coefficients 86, 350, 605
envelope correlations 403–14
  clipped envelopes 407–14
envelope pdf 239–41
envelope spectrum 406
envelope variance 285, 287
envelope-phase relations 370–84
  frequency-selective channel 371–2
equal gain combining 558–9, 563, 568
equalisation 38, 44–5
equalisers 44–6
  decision feedback 44
  Viterbi 44
equivalent rectangular bandwidth 249
equivalent surface impedance 115
  horizontal polarisation 116–17
  vertical polarisation 117–18
ergodicity 688
Erlang 10–11
error function 699
ESPRIT 396, 398
ETSI 13
Euler's constant 481
expectation 686
exponential taper 514

fade depth 258–9
  1D model 259
  2D model 258
fade duration 355–7
fade length 104–5
fading 34–5
fading channels 38–40
  envelope-phase relations 370–84
  finite bandwidth effect 195–8
fading signal 376–7, 381–3
Fano criterion 510
Fano theory 510
far field 436–7, 466, 469–70
  integration 480–2
far-field approximation 468
far-field scatterer 452
Faraday rotation 538
fast fading 34–5
FDMA 22–3
feed impedance 522–6
ferrite core 494–7
few-path model 234–42

field component diversity 585
field correlation matrix 601
field quantities 497
field relations 497
field strength 683–4
fields for electric dipoles 471–3
  'infinitesimal', uniform dipole 471–2
  sinusoidal dipole current 472–3
  uniform distribution dipole source 473–4
fields of a current distribution 466–80
  combination of electric and magnetic sources 478–9
  electric dipoles 471–4
  magnetic sources 473–5
finite bandwidth effects 195–8, 248–50
finite impulse response 373
firecracker antenna 3–4
Fisher information matrix 694
Fisher's pdf 401
Fisher's score function 693–5
Fisher's transformation 402
flat fading 35
flat-fading channels 40–1
folded dipoles 462–3
folded *N*-pole 463–4
forward error correction 30–1
Foster's reactance theorem 502
Fourier angular resolution 251–2
  broadside to trajectory 251–2
  direction of trajectory 251–2
Fourier integral 514
Fourier model 68–9
Fourier resolution 309–10
Fourier transform 68, 76, 79–80, 212, 261, 275, 373, 420, 689
  pair 68
FPLMTS 9
frame antenna 4, 6
free space 532
free space path loss 50, 683
frequency coherence function 203, 208
frequency correlation 280
frequency correlation function 83
frequency covariance function 283
frequency diversity 46
frequency division duplexing 23–4
frequency re-use 14–16
frequency shift keying 51
frequency-fading channels 210
frequency-flat fading 21
frequency-selective channels 21–2, 42–6, 371–2
  channel characterisation 419
  equalisation 44–5
  static 209
  time diversity 43–4
frequency-selective fading 21, 35, 39
Fresnel coefficients 89–90, 93
Fresnel integral 128
Fresnel zones 100, 109–10, 133, 735–6
Friis transmission equation 443–4, 447, 449, 451–2, 672, 683
fringing fields 519
function transformation theorem 686

gain 435–8, 641, 648–9, 655–9
  co-polar 438
  directional 436–7
  directive 438
  maximum 645–8
gain enhancement 639–41
  correlated case 640
  uncorrelated case 640
gain reduction 588, 590
gain-to-temperature ratio 455, 457
gamma distribution 721
  generalised 723–6
gamma function 707
  incomplete 721, 725
Gans' law 206, 405
Gaussian beam 580–5
  circular coordinates 581
Gaussian beam pattern 439–41
Gaussian conditioning theorem 254–5
Gaussian continuous source model 392
Gaussian correlated samples 360–2
Gaussian density
  standardised 674
Gaussian distribution 210, 220, 237, 679–700
  multivariate 728
Gaussian independent samples 359–60
Gaussian process 697–8, 702, 704
  moments 698
Gaussian scenario 312
Gaussian spread parameter 439
general scattering 303–6
generation of channel functions 359–70
  correlated signals 364–5
  transfer functions 365–8
geometric directivity 588
geometric optics 123–4, 143
geometric theory of diffraction 123–5, 129
geostationary satellites 317–18, 540, 548–50
ghost cancellation 421–2

global positioning system 318, 540, 545, 548
global variance 209
Goubau antenna 464–5, 515–19
  Q 505
gradient (grad) 498
Gram matrix 642
grazing incidence 92–4, 99–102
  distance dependence 100–1
  frequency dependence 101–2
  receiver height dependence 101
Green's function 468
  far-field approximation 468
ground noise 453
groundplane 463–5
  finite 486–8, 490
groundplane effects 505–6
groundplane principles 460–1
group delay 186–8, 201–2, 208, 214, 252, 255–9, 737–8
  complex 219
  distribution 256
  excursions 256–9
  frequency domain 737
  generalised 221
  mean 186, 188, 199, 203, 208–9
  negative 198–203, 208
  relative mean 199–200
  time domain 737
  two-path model 187
GSM 631

half power beamwidth 276–9, 439–40
hand-off boundary 15–16
handset antennas 532–8
Hata's approximation 150–1, 153, 158, 160–2
helix antenna 533, 535, 542, 544
  multifilar 545
  quadrifilar 545
  scanning-mode 550–5
  spherical 505, 509
Helmholtz scattering formula 105–6
Hertz 3
Hilbert transform 211, 217, 276, 372
  channel characterisation 378–9
  phase ambiguity 383–4
HIPERLAN 13
Hoyt distribution 712, 714
Hoyt k-factor 714
Huygens' principle 124, 735
hypergeometric function 401

image model 98–9
impedance 461, 464
  monopole 488
impedance match efficiency 443
impedance matrix 590
  antenna 594–5
  load 594–5
impulse response 38–9, 70–1, 74–5
  correlation 81
  finite bandwidth effects 248–50
  many-path 147–50
  two-path model 182–3, 188–91
IMT-2000 9
in-phase and quadrature components 236–9
  bandwidth 406
incident field
  polarisation matrix 575
incident power distribution 575–7
incoherent component 705
incoherent field 113
incoherent scattering 119
indoor environments 171–3, 178
  angular domain 178
  time domain 171–3
indoor modelling 165–7
induced emf method 483
infinite impulse response 373
infinitesimal electric dipole 587
information channel 26–8
information matrix 694
information rate 30
infrastructure 1–2
input admittance 485
input delay-spread function 248
input impedance 465, 478
input reactance 484
inscribing sphere 526, 532
instantaneous bandwidth 214
instantaneous delay spread 256, 427–8
instantaneous frequency 214, 420
  complex 219
instantaneous scenarios 300, 328–9
integral equation formulation 158–62
interference 2, 12, 53, 56, 571–4
  intersymbol 36–7
interference rejection 631–2
interference-to-signal ratio 633–5
interferers 660–9
  fixed powers 665, 668
  variable powers 665–8
interleaving 30
intersymbol interference 36–7
Iridium satellite system 540

irreducible bit-error ratio 34–7
irregular terrain 163
isotropic surface 115–18
ITU 10, 24
ITU-R 24

Jacobian 686, 700–1
joint distribution
  between branch signals 557
  between Rayleigh envelopes 559
joint envelope 348–9
joint pdfs 347–8, 408, 728
joint phase pdf 348–9
jointly normal distribution 254

Kahunen–Loeve expansion 366–70
kernel function 213–14
kernels 216–17
keyholes (pinholes) xxix, 654–9
Kirchhoff's solution 105–17
  limitations 113–15
  phase-only effects 107–9
Klopfenstein taper 514
Kolmogorov–Smirnoff test 696
Kummers function 707

Laplace transform 372–3
Laplacian azimuth function 174–5
Laplacian distribution 260, 498, 500
Laplacian scenario 312–13
leaky waves 552
least mean squares method 633
level crossing 353–5
  diversity combined signals 354–5
  random FM 358
likelihood function 693–4
limiter 408
line source 294–6
  circular 296–7
linear systems model 246–7
load match parameters 444–6
loaded circuit 595
loaded circuit envelope correlation coefficient function 602–3
local delay spread 209, 214–18
local variances 214–15
log likelihood function 693
lognormal distribution 715–18
long-term fading 47
loss factor 90
loss tangent 90
low noise amplifier 459
low-earth-orbit satellite 318, 540, 545
low-profile antennas 531
low-profile dual layer patch 531–3

macro diversity 48
magnetic current 521
magnetic dipole 507
magnetic dipole moment 475
magnetic energy 502
magnetic loop 527
magnetic sources 473–5
MAN 13
man-made noise 454
many-path model 245–51
Marconi xxix, 3–4
Marcum Q function 709–11
  generalised 729–30
matching circuit 495, 510–13
  discrete component 512–13
matching network 447
Matlab 331–3
maximum gain 645–8
  known channel 645
  unknown channel 648
maximum gain combining 649
maximum likelihood estimation 45, 692–5
maximum phase 376, 380–3
maximum ratio combining 558, 563–4
  cdf 566
  diversity gain 567–8
  Rayleigh envelopes 733
Maxwell 3
Maxwell duals 474, 477
Maxwell's equations 499, 501
  differential 499
  integral 499
Maxwellians 3
mean capacity 678–9
mean delay 221
mean delay time 183, 209
mean direction 278
mean effective gain 443, 536–8, 577
mean estimate 692
  variance 692
mean excess delay 183
mean path loss 48–9
mean power 291
  spread 285–8, 292
  variance 283–4, 287
mean reflection 305
mean square error 659
  minimum 663
mean square excess delay 183

mean square time 183, 216
method of moments 480
Middleton's Class A impulsive noise distribution 727
Millington correction factor 157–8
MIMO systems 642, 650, 669–72, 674
  capacity xxix, 671–2
minimum descriptor length 394, 397
minimum norm algorithm 395
minimum phase 222–5, 372, 375, 380–1
minimum scattering antennas 593–4
  canonical 593
  mutual impedance 593
mixed path 50
mixture distribution 719–20
mobile channel terminology 25–33
mobile terminals 620–2
mobile to geostationary satellite links 548–51
modified Poisson process 365
modulation 30–1, 47, 51
modulation correction 377–8
  magnitude 377
  phase 378
monopoles 329, 461, 464–6
  finite groundplane 486–9
  folded 463
  infinite groundplane 484
  in-line 329
  input impedance 486, 488
  input reactance 489
  mutual impedance 592–3
  quarterwave 488–91
  radiation resistance 489
  top-loaded 515–16
monopoles on a groundplane 600–6
  four in-line monopoles 604–5
  sloping monopoles 607–15
moving average technique 364
moving receiver 65–7, 73–4, 328
  Brewster angle effect 94–5, 97
  signal dynamics 352–3
moving terminal 229–33, 246–7
MSK 36
multi-element arrays 642
multipath channel 247–8
multipath field components 63
multipath mechanisms 89
multipath mitigation 37–8, 47
multipath model testing 384–400
multipath propagation xxvii, xxviii, 33, 63
multipath reception 63
multiple access techniques 21–3, 47
multiple quarterwave sections 513–15
multiple scattering algorithm 159–60
multiplicative mean 417–19
multiscreen diffraction 153
multiscreen model 153
multi-user access technique 17–18
multi-user scheme 53–4
multi-user system requirements 14–24
multivariate Gaussian 728
MUSIC algorithm 395
mutual coupling 329, 493, 524, 590–1
  network models 591
mutual impedance 477, 590–4
  monopoles 592–3
mutual resistance 593–4

Nakagami 706
Nakagami distribution 720–2
Nakagami–Rice distribution 706
narrowband channels 35, 38–9
  modulation correction 377–8
narrowband Gaussian process 84–5
narrowband model 327–31
narrowband signals 400–19
  correlation coefficient estimation 400–3
  envelope correlations 403–14
near-far effect 18, 20
negative group delay 198–203
negatively correlated signals 560–1
network model 594–5
noise 662
noise factor 457
noise figure 455–9
  circuits 457–9
noise power 456–9
noise sources 454
non-Gaussianity 726–7
nonlinearities 28–9
non-line-of-sight xxvii, 54–6
  interference 56
non-minimum phase 223–4
normalised pattern 436–7
Norton surface wave 91–2
null steering 41–2
Nyquist's criteria 31
Nyquist's formula 453

ohmic loss
  wires 488–91
ohmic resistance 480, 492–4
Okumura-Hata case 149–51, 158
omnidirectional scenario 278, 404
open circuit 595

open circuit voltage 309
operating environments 10
optimum combining 41–3, 45–6, 632, 638, 641–2, 659–60
optimum diagonal load resistance 597–9
optimum weight vector 632
optimum weights 662–3
orthogonal frequency division modulation 13
OSI model 10
outages 38
  analogue 659
  capacity 669–71
  digital 659, 670
outdoor environments 320–5
  diffuse model 320–2
output delay spread function 248
output power 661–2

Page distribution 213
PAL television 421–2
PAN 13
parabolic method 163–4
parallel channels 669
parametric models 165–7
Parl's algorithm 710–11
Parseval's theorem 737
partitioning 387–90
patch antennas 6, 8, 519–32
  absorption 535
  circularly polarised 546–8
  low-profile dual layer 531–2
  modelling 474
  multiple stacked 529–31
  on a groundplane 615–18
  PIFA 519–26
  Q 505
  square 519, 526–9
path gain 48, 50
path loss 48, 50, 52, 54, 101–4, 338–9, 683–4
  circular polarisation 104
  free-space 684
  measurement-based 56–7
pattern models 439
  circular functions 440–1
patterns 435–6
  antenna 575–8
  circular patch 548, 616
  conical 549
  dipole antenna 473
  fan 554
  finite groundplane 487, 490
  Gaussian beam 439–41
  realistic 585–6
  satellite links 544–6
  scanning-mode helix 551–2, 555
  sloping monopole 607–9
pencil beam formula 439, 441
periodicity 273
personal antennas 8
personal communications 9, 12
personal terminals 314–15
perturbation theory 115–18
phase
  correlation coefficient 351–2
  single signal 349
phase derivatives 252–6
phase dispersion 187, 223
phase linearisation 66–7
phase mixing 66
phase roll 366–7
phased array 41–2
phi factors 121–1
physical optics 123–4, 735
PIFA 519–26
  cavity model 520–2
  feed impedance 522–4
  inscribing sphere 526
pinholes (keyholes) xxix, 654–9
plane wave illumination 299
Poincaré sphere 444
point scatterers 65
point source model 391
polarisation 63–4, 71, 89–90, 92, 97, 106, 585
  circular 94, 96, 98, 104, 444, 538–44
  cross-polar 104
  elliptical 63–4
  mixed 103
polarisation coupling 305
polarisation diversity 40, 577–8, 620–2
polarisation efficiency 443–4, 542
polarisation matrix 300–2, 310, 575
polarisation purity 542
pole-zero description 372–6
postdetection combining 563
Poulsen arc transmitter 3
power
  finite bandwidth effect 282–5
  finite distance effect 286–93
power azimuth spectrum 177
power consumption 2
power control 20–1
power correlation 84–5
power correlation coefficient 85–6, 583

power delay profile 80–3, 167–72, 203, 262, 324–5
two-path model 183, 190, 192
power delay spectrum 174
power density 683
power flux density 12–13
power profiles 80–1
power signal 283–5, 408–9
power spatial gradient 259–61
power spectral density 690
power spectrum 689–90
power transfer function 231–2
Poynting vector 448
predetection combining 563
prediction 385–6
prediction distance 396–7
prediction length 398–9
Principal Components Linear Prediction 396
principle of reciprocity 64, 435
probability 685
probability density function 238, 340, 685–6, 695
chi-squared 236–7
closed forms 240–2
envelope 239–41
in-phase and quadrature signal 238
student-T 256
probability functions 193–5
probability of a bit error 51
propagation channel 26–8
propagation channel models 181
propagation modelling 149
propagation time 247
proximity effect 493–4
pseudo beat distances 337
pseudo spectrum 395
PSTN 2

Q 501–6
loaded 504
measured 504
minimum 509
resonance 503–4
square patch 527–8
theoretical 507
unloaded 502, 509
Q-function 699
Marcum-*Q* 708–710
QPSK 51–3
quadratic phase channel 220
quadrature amplitude modulation 31
quadrature components 236–9
quality factor *see Q*
quality of service 10
quarterwave monopole 488–91
quarterwave section 513–14
multiple 513–14

radar cross section 452
radiation efficiency 446–7, 493–4
radiation from moving charges 466–7
radiation mechanisms 552–4
radiation resistance 447–9, 502
dipoles 480–9
monopole 489
sloping monopole 610–13
small loops 492–6
thin wire dipole 482
radio astronomy 12–13
radio carrier frequency 330
radio channel 26–8
radio data networks 13
Radio Regulations 25
radio spectrum 24–5
radiosity 142, 144–5
rake 38, 43–4, 46
random frequency modulation 35, 252, 256, 352
level crossing rate 358
spectrum 357–8
random phasor plus Rayleigh 715
range 650
raw data channel 26–8
ray launching 144
ray model 303–6
ray tracing 142–4
Rayleigh channels 675–9
correlated 675–6
uncorrelated 675–9
Rayleigh distribution 239, 254, 700–4
Rayleigh envelope 282, 704–5
higher moments 703
maximum ratio combination 733
Rayleigh fading 17, 19, 21, 34, 48, 54
Rayleigh roughness factor 112, 114, 117
reactance 502
received power 595–6, 598–600, 663
received signal 340–1, 663
received signal strength indicator 21–2, 283, 405, 407
receiving antenna
equivalent circuit 450–1
receiving arrays 659–69
optimally combined 659
statistical model 660–3

reciprocity 33, 64, 435
rectangular patch 527
rectangular source coordinates 468–9, 479–80
Reed–Solomon coding scheme 30
reflection coefficients 89, 91, 93–4, 96, 99, 103, 112–13, 304
  co-polar 94, 96, 98, 305
  cross-polar 94, 96, 98, 305
  specular 118
reflection from a rough surface 105–23
  Kirchhoff's solution 105–15
  perturbation theory 115–18
reflection from a smooth planar surface 89–105, 306–8
  grazing incidence 92–3, 99–102
  normal incidence 95–7
  Norton surface wave 91–2
  polarisation 539
regularity condition 694
relative permittivity 90
resistive loads 595–6
resolvability of scatterers 69–70
resonance 445, 461, 503–4, 510–11
  single 503–4
retarded charge 467
retarded current 467
return loss 446
Rice 706
Rice distribution 704–7
Rice envelope 706–9
  moment formula 706–7
Rice K-factor 317, 706–7, 712–13
Rice phase 712–13
Rician channels 672, 674
Rihaczek bandwidth 199–200, 219–20, 224–8
  two-path model 226
Rihaczek time 217
RLAN 9, 13
rod antenna 5–6
rough dielectric medium 115–18, 121
rough surface scattering
  polarisation aspects 121–3
rough surface statistics 115
roughness factor 109, 112
RSSI *see* received signal strength indicator
running mean 687
rural and hilly terrain 157–64
  building effects 163–4
  open undulating terrain 158
rural environments 167–9, 173–5
  angular domain 173–5
  time domain 167–9
Rutherford xxix, 3

sample average 686
sampled power 288–93
satellite antennas 455
  mobile 455
satellite links 539–40, 544–6
  mobile 539–40
  space loss 544–6
satellites 317–19
  geostationary 317–18, 540, 548–50
  low-earth-orbit 318, 540, 545
  non-geostationary 318–19
scanning diversity 561
scanning-mode helix 550–5
  radiation mechanism 552–4
scatter distribution models 293–300
scattered field
  mean 111
  scalar version 105
scattering cross section 120, 168, 452
scattering distribution spectrum 246
scattering function 75
scattering medium 68
  continuous 68
scattering parameters 590
scattering wave numbers 120
scenario 181
  Clarke 262–3, 267, 273, 293, 297–9, 579, 601
  directional 274, 279
  instantaneous 300, 328–9
  omnidirectional 278, 404
  static two-source 182
  uniform 265, 267, 297, 299–300, 333–5, 342–3, 346, 582, 585
  vehicular 609–10
scenario models 181, 245
  static two-source 182
selection combining 558–9, 561–3, 566–7
  cdf 566
  diversity gain 567–8
selection diversity 558
  cdf 558
self-complementary antennas 477–8
  helix 477
  planar spiral 477
self-impedance 477
shadow fading 47–8
shadow zone 124–5
Shannon capacity 669

Shannon capacity theorem 29–30, 32, 34
Shannon channel efficiency 33
shifting property 299
short-term channel behaviour
  many-path model 245
  two-path model 181
short-term fading 34, 327–39
  prediction 384–400
signal combining 556–64
  algorithms 558–60
signal fading 34
signal in a channel 221–2
signal prediction 385–6
  additive noise 391
  noise-free case 387–90
  results 396–9
signal processing 37–8
signal time 211
signal-to-noise-ratio (SNR) 34, 51, 53–6, 633–4
signal-to-noise-plus-interference ratio (SNIR) 557
simulation
  reference parameters 340–58
  short-term fading 327–39
single source 294–5
singular value decomposition 642–5
  expansion 643
SINR 632–9, 659, 662–9
  cumulative distribution function 665–9
  maximum 662–3
skin depth 489, 491
sky noise 453–4
slope diffraction 133–5, 137, 139
sloping monopoles 607–15
  correlation coefficients 612–14
  distributed directivity 609–10
  diversity gain 613–15
  radiation resistance 610–13
slot antenna 476–7
slot feeding 475–6
slow fading 47–8
slow frequency hopping 46–7
small antennas 505–15
  bandwidth enhancement 509–15
  groundplane effects 505–6
  see also Goubau antenna
  see also patch antennas
small loops 492–6
  circular 492
  closely spaced 493–4
  ferrite loaded 494–7
Smith chart 445, 512, 529
snapshot of discrete scatterers 245–6
solenoidal field 469
Sommerfeld's attenuation factor 91
sounding 419–28
sounding pulses 248
source coding 30
source elements 460
source estimation 394
space diversity 40, 577–9
space/frequency correlation 280–2
space-loss factor 318–19, 544–6
space-time processing 38
spatial envelope cross-correlation coefficient function 604
spatially correlated signals 634–7
spatially uncorrelated signals 632–5
specific absorption rate 533
spectral efficiency 14, 16, 31
  parallel channels 650–4
spectral occupation efficiency 423–5
spectral requirements 11
spectral usage 2, 24–5
spectrogram 210
spectrum 689, 691
spectrum allocation 25
specular component 705
specular scattering 109, 111–13
spherical helix antenna 505, 509
spherical source coordinates 468–9, 479–80
square patch antenna 519, 526–9
  array 528–9
standard deviation 686
standard normal function 699
static multipath 34
static two-source scenario 182
statistical average 686
statistical independence 685
statistics 74
  few-path model 234–42
  for mobile communications 685–96
Stoke's theorem 500
street corner 145–6
strict-sense stationarity 689
student-T distribution 255–6, 260–1, 269, 271
subspace methods 391–400
suburban environments 158–9
super-resolution methods 391–400
  deconvolution 424–8
surface height distribution 111–12, 115
  correlation function 115
  Gaussian 111–12
surface roughness 109–11

surface-wave antenna 550
susceptance 502
Suzuki distribution 304, 719
sweep rate 420, 422
switched combining 558, 562–3, 571–3
switched diversity 571
switched diversity antenna 617–20
switched parasitic elements 617–20
synchronisation 202–3
system temperature 456–7, 459

tail-area function 699
tapered section 514–15
  Dolph–Chebyshev 514
  exponential 514
  Klopfenstein 514
tapered transmission lines 514–15
TDMA 22–3
TE mode 508, 546
temperature distribution model 455–6
thermal noise 662
thin wire dipole
  input reactance 484, 487
  radiation resistance 482
thin wire monopole
  input impedance 486
three-branch circular array 599
three-dimensional effects 164–5
three-path model 234–5
time diversity 43–4
time division duplexing 23–4, 642
time domain modelling 167–73
time-dispersed signals 45
time-frequency distribution 211–14
time-frequency theory 209–28
timing errors 35–6
time series correlation 310–11
timing jitter 202–3
timing synchronisation 257
TM mode 508
Toeplitz correlation structure 363
traffic asymmetry 10–11
traffic density 10–12
transfer function 74–87, 661
  2D 229–32
  autocorrelation 75–6
  correlation 81–3
  derivatives 252–63
  finite bandwidth effect 232–3
  integrating over finite bandwidth 282–5
  integrating over finite distance 286–93
  local approximation 218–21
  phase 185–6
  spatial 381–2
  spatial periodicity 267
  spectral functions 75
  two-path model 184–5, 188–9
transfer function generation
  frequency domain 366–8
  time (space) domain 365–6
transfer function model 63, 65–6, 68–71, 74
  time domain representation 70
transition function 128
transition zone boundary 132–3
transition zone diffraction 133–42
  N screens 140–2
  one screen 136–7
  two screens 138–40
transmission line 455–6, 458
  tapered 514–15
transmission line transformers 513
transmit diversity 46–7
transmit-receive arrays 669–71
transmit-receive diversity 47, 645–8
  known channel 646
  unknown channel 648
transmitting antenna
  equivalent circuit 450–1
transmitting pattern 64
Turin model 170–1, 249
two signals 343–6
  joint envelope 348–9
  joint pdf 347–8
  joint phase pdf 348–9
two-element arrays 598–9, 642
two-path model 181–2, 267
  dynamic 261–3

UMTS 9
uncertainty principle 205–6, 210, 272, 278
uncorrelated branches
  common mean powers 565–6
  different mean powers 564–5
uncorrelated noise 691
uncorrelated scatterers 300, 308–12
uncorrelated scattering 76
uncorrelated signals 300, 632–5
uniform line source 474
uniform scenario 265, 267, 297, 299–300, 342–3, 346, 582, 585
  discretisation 333–5
  simulation 327–8, 331, 339
uniform slot 523

uniform theory of diffraction 123, 125–42, 153–4
  transition zone 133–42
  wedge 125–32
unresolvable physical scatterers 74
uplink 637–8, 641
urban environments 149–57, 164, 170–1, 174–7
  angular domain 174–7
  diffraction to street level 155
  propagation above rooftops 149–57
  time domain 170–1
user discrimination
  uplink case 637–8

variance 701
variance of mean estimate 692
vector magnetic potential 470–1
vector notation 391–2
vehicle antennas 3–8
vehicular communications 9, 313–14
  power flux density 12
vehicular scenario 609–10
voltage reflection coefficient 444–5
voltage standing wave ratio 445–6, 504, 529

Walfisch–Bertoni model 151–3
Walfisch–Ikegami model 154, 156–7
WAN 13
water filling 651
wave equation 500
wave structure approach 483–6
wedge 125–32
weights 41–3, 64, 632–46, 661–2
whip antenna 6
wideband channels 21–2, 35, 39–40, 203
  channel sounding 419–28
wide-sense stationarity 76–80, 689
  frequency domain 76, 79
  space domain 77, 79
Wiener–Khinchin theorem 690
Wigner distribution 213, 216
wire antennas
  ohmic loss 488–91
wire dipoles 465, 482, 484
WLAN 13, 165
WSSUS 77, 79–80

zero description 376–7
zero flipping 380–1
zero-forcing algorithm 44

Channels, Propagation and Antennas for Mobile Communication

Written by two leading experts in the field, this exceptional book introduces the reader to the principles, theory and applications of physical layer wireless/mobile communications. In the area of wireless, the antennas, propagation and the radio channel used are inter-related; this book offers an explanation of that relationship, which is fundamental to the development of systems with high spectral efficiency.

Channels, Propagation and Antennas for Mobile Communications emphasises the basic principles needed to establish an understanding of this technology. However, the tools required – such as the mathematics and statistics – are treated in the manner of a practical handbook, avoiding detailed derivations. The reader will develop a clear insight into the physics and effects of multipath and the use of multipath channels for communications. Propagation modelling, simulation and measurement, scattering, antenna principles, array antennas, adaptive antennas and smart antennas are also covered.

This book will be a valuable reference for senior undergraduates and postgraduates in electrical engineering and communications. Practising design engineers, systems designers and engineering managers will also gain a sound understanding of the field.

Rodney Vaughan is senior research engineer at Industrial Research Limited, New Zealand. He has degrees in electrical engineering from the University of Canterbury, New Zealand, and gained his Ph.D. from Aalborg University, Denmark. Since then he has worked for Telecom NZ and Industrial Research Limited. For the latter he established research and development groups in signal processing and multipath communications. Applications include designs in antennas and signal processing, which have been used widely. In addition to his work on many industrial projects, his research includes topics in multipath, signal processing and multiport antennas, and he has published over 100 papers and conference contributions.

Jørgen Bach Andersen is the director of the Center for PersonKommunikation and Professor of Wireless Communications at Aalborg University in Denmark. He gained his degree and Ph.D. from the Technical University of Denmark in 1961 and 1971 respectively. His research has focused on radio propagation, antennas and signal processing, as well as biomedical studies related to electromagnetic waves. He has been active in European cooperation with the COST framework for wireless communications and has served on several COST management committees. He has published over 200 papers and conference contributions, and contributed to nine books. A former vice president of URSE, and former chairman of URSI Commissions B and K, he is a Life Fellow of IEEE.



The Institution of Engineering and Technology
www.theiet.org
0 85296 084 0
978-0-85296-084-4