

Thin-Film Optical Filters



Preparing a plant for the manufacture of narrowband filters. (Courtesy of Walter Nurnberg FIEP FRPS, the editors of *Engineering*, and Sir Howard Grubb, Parsons & Co Ltd.)

Thin-Film Optical Filters

THIRD EDITION

H A Macleod

Thin Film Center Inc.

Tucson, Arizona

and

Professor Emeritus of Optical Sciences

University of Arizona

IOP

Institute of Physics Publishing

Bristol and Philadelphia

© H A Macleod 1986, 2001

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher. Multiple copying is permitted in accordance with the terms of licences issued by the Copyright Licensing Agency under the terms of its agreement with the Committee of Vice-Chancellors and Principals.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

ISBN 0 7503 0688 2

Library of Congress Cataloguing-in-Publication Data are available

Consultant Editor: **Professor W T Welford**, Imperial College, London

Production Editor: Simon Laurenson

Production Control: Sarah Plenty

Cover Design: Victoria Le Billon

Marketing Executive: Colin Fenton

Published by Institute of Physics Publishing, wholly owned by The Institute of Physics, London

Institute of Physics Publishing, Dirac House, Temple Back, Bristol BS1 6BE, UK

US Office: Institute of Physics Publishing, The Public Ledger Building, Suite 1035, 150 South Independence Mall West, Philadelphia, PA 19106, USA

Typeset in TeX using the IOP Bookmaker Macros

Printed in the UK by J W Arrowsmith Ltd, Bristol

*To
my Mother and Father
Agnes Donaldson Macleod
John Macleod*

Contents

Foreword to the third edition	xiii
Foreword to the second edition	xv
Apologia to the first edition	xix
Symbols and abbreviations	xxiii
1 Introduction	1
1.1 Early history	1
1.2 Thin-film filters	5
References	9
2 Basic theory	12
2.1 Maxwell's equations and plane electromagnetic waves	12
2.1.1 The Poynting vector	17
2.2 The simple boundary	18
2.2.1 Normal incidence	20
2.2.2 Oblique incidence	23
2.2.3 The optical admittance for oblique incidence	27
2.2.4 Normal incidence in absorbing media	29
2.2.5 Oblique incidence in absorbing media	34
2.3 The reflectance of a thin film	37
2.4 The reflectance of an assembly of thin films	40
2.5 Reflectance, transmittance and absorptance	43
2.6 Units	46
2.7 Summary of important results	46
2.8 Potential transmittance	50
2.9 Quarter- and half-wave optical thicknesses	52
2.10 A theorem on the transmittance of a thin-film assembly	53
2.11 Admittance loci	55
2.12 Electric field and losses in the admittance diagram	60
2.13 The vector method	66
2.14 Incoherent reflection at two or more surfaces	67
2.15 Other techniques	72

2.15.1	The Herpin index	72
2.15.2	Alternative method of calculation	73
2.15.3	Smith's method of multilayer design	75
2.15.4	The Smith chart	77
2.15.5	Reflection circle diagrams	80
	References	85
3	Antireflection coatings	86
3.1	Antireflection coatings on high-index substrates	87
3.1.1	The single-layer antireflection coating	87
3.1.2	Double-layer antireflection coatings	92
3.1.3	Multilayer coatings	102
3.2	Antireflection coatings on low-index substrates	108
3.2.1	The single-layer antireflection coating	110
3.2.2	Two-layer antireflection coatings	111
3.2.3	Multilayer antireflection coatings	118
3.3	Equivalent layers	135
3.4	Antireflection coatings for two zeros	139
3.5	Antireflection coatings for the visible and the infrared	144
3.6	Inhomogeneous layers	152
3.7	Further information	156
	References	156
4	Neutral mirrors and beam splitters	158
4.1	High-reflectance mirror coatings	158
4.1.1	Metallic layers	158
4.1.2	Protection of metal films	160
4.1.3	Overall system performance, boosted reflectance	164
4.1.4	Reflecting coatings for the ultraviolet	167
4.2	Neutral beam splitters	169
4.2.1	Beam splitters using metallic layers	169
4.2.2	Beam splitters using dielectric layers	172
4.3	Neutral-density filters	176
	References	177
5	Multilayer high-reflectance coatings	179
5.1	The Fabry–Perot interferometer	179
5.2	Multilayer dielectric coatings	185
5.2.1	All-dielectric multilayers with extended high-reflectance zones	193
5.2.2	Coating uniformity requirements	200
5.3	Losses	204
	References	208

6 Edge filters	210
6.1 Thin-film absorption filters	210
6.2 Interference edge filters	211
6.2.1 The quarter-wave stack	211
6.2.2 Symmetrical multilayers and the Herpin index	213
6.2.3 Performance calculations	223
References	255
7 Band-pass filters	257
7.1 Broadband-pass filters	257
7.2 Narrowband filters	260
7.2.1 The metal–dielectric Fabry–Perot filter	260
7.2.2 The all-dielectric Fabry–Perot filter	266
7.2.3 The solid etalon filter	280
7.2.4 The effect of varying the angle of incidence	283
7.2.5 Sideband blocking	293
7.3 Multiple cavity filters	293
7.3.1 Thelen’s method of analysis	300
7.4 Higher performance in multiple cavity filters	306
7.4.1 Effect of tilting	315
7.4.2 Losses in multiple cavity filters	316
7.4.3 Case I: high-index cavities	317
7.4.4 Case II: low-index cavities	318
7.4.5 Further information	318
7.5 Phase dispersion filter	319
7.6 Multiple cavity metal–dielectric filters	325
7.6.1 The induced-transmission filter	328
7.6.2 Examples of filter designs	334
7.7 Measured filter performance	342
References	345
8 Tilted coatings	348
8.1 Introduction	348
8.2 Modified admittances and the tilted admittance diagram	349
8.3 Polarisers	362
8.3.1 The Brewster angle polarising beam splitter	362
8.3.2 Plate polariser	366
8.3.3 Cube polarisers	367
8.4 Nonpolarising coatings	368
8.4.1 Edge filters at intermediate angle of incidence	368
8.4.2 Reflecting coatings at very high angles of incidence	374
8.4.3 Edge filters at very high angles of incidence	376
8.5 Antireflection coatings	377
8.5.1 p-polarisation only	378
8.5.2 s-polarisation only	379

8.5.3	s- and p-polarisation together	381
8.6	Retarders	382
8.6.1	Achromatic quarter- and half-wave retardation plates	382
8.6.2	Multilayer phase retarders	385
8.7	Optical tunnel filters	389
References		391
9	Production methods and thin-film materials	393
9.1	The production of thin films	394
9.1.1	Thermal evaporation	395
9.1.2	Energetic processes	405
9.1.3	Other processes	413
9.1.4	Baking	417
9.2	Measurement of the optical properties	418
9.3	Measurement of the mechanical properties	436
9.4	Toxicity	445
9.5	Summary of some properties of common materials	446
References		456
10	Factors affecting layer and coating properties	462
10.1	Microstructure and thin-film behaviour	462
10.2	Sensitivity to contamination	478
References		485
11	Layer uniformity and thickness monitoring	488
11.1	Uniformity	488
11.1.1	Flat plate	490
11.1.2	Spherical surface	490
11.1.3	Rotating substrates	490
11.2	Substrate preparation	497
11.3	Thickness monitoring	499
11.3.1	Optical monitoring techniques	500
11.3.2	The quartz-crystal monitor	509
11.4	Tolerances	511
References		520
12	Specification of filters and environmental effects	523
12.1	Optical properties	523
12.1.1	Performance specification	523
12.1.2	Manufacturing Specification	526
12.1.3	Test Specification	527
12.2	Physical properties	530
12.2.1	Abrasion Resistance	530
12.2.2	Adhesion	533
12.2.3	Environmental Resistance	533
References		535

13 System considerations: applications of filters and coatings	536
13.1 Potential energy grasp of interference filters	540
13.2 Narrowband filters in astronomy	545
13.3 Atmospheric temperature sounding	550
13.4 Order-sorting filters for grating spectrometers	559
13.5 Glare suppression filters and coatings	570
13.6 Some coatings involving metal layers	575
13.6.1 Electrode films for Schottky-barrier photodiodes	575
13.6.2 Spectrally selective coatings for photothermal solar energy conversion	579
13.6.3 Heat reflecting metal–dielectric coatings	583
References	585
14 Other topics	588
14.1 Rugate filters	588
14.2 Ultrafast coatings	599
14.3 Automatic methods	610
References	619
15 Characteristics of thin-film dielectric materials	621
References	628
Index	631

Foreword to the third edition

The foreword to the second edition of this book identified increasing computer power and availability as especially significant influences in optical coating design. This has continued to the point where any description I might give of current computing speed and capacity would be completely out of date by the time this work is in print. Software for coating design (and for other tasks) is now so advanced that commercial packages have almost completely replaced individually written programs. I have often heard it suggested that this removes all need for skill or even knowledge from the act of coating design. I firmly believe that the need for skill and understanding is actually increased by the availability of such powerful tools. The designer who knows very well what he or she is doing is always able to achieve better results than the individual who does not. Coating design still contains compromises. Some aspects of performance are impossible to attain. The results offered by an automatic process that is attempting to reach impossible goals are usually substantially poorer than those when the goals are realistic. The aim of the book, therefore, is still to improve understanding.

During the years since publication of the second edition, the energetic processes, and particularly ion-assisted deposition, have been widely adopted. There are several consequences. The improved stability of optical constants of the materials has enabled the reliable production of coatings of continuously increasing complexity. We even see coatings produced now purely for their aesthetic appeal. Then the enormous improvement in environmental stability has opened up new applications, especially in communications. Unprecedented temperature stability of optical coatings can now be achieved. Specially designed coatings have simplified the construction of ultrafast lasers. Banknotes of many countries inhibit counterfeiting by carrying patches exhibiting the typical iridescence of optical coatings. Coatings to inhibit the effects of glare are now integral parts of visual display units.

I mentioned in my previous foreword the difficulty I experienced in bringing the earlier edition up to date. This time the task has been even more difficult. The volume of literature has expanded to the extent that it is almost impossible to keep up with all of it. The pressure on workers to publish has in many cases reached almost intolerable levels. I regret I do not remember exactly who introduced the idea of the half-life of a publication after which it sinks into obscurity but it is

clear that the half-life has become quite short. Comprehensively to review this vast volume of material that has appeared and continues to appear would have changed completely the style of the book. The continuing demand for the now out-of-print second edition of the book suggests that it is used much more as a learning tool than a research reference and so my aim has been to try to keep it so. There have been few fundamental changes that affect our basic understanding of optical coatings and so this third edition reflects that.

I appreciate very much the help of various organizations and individuals who provided material. Many are named in the foreword to the second edition and in the apologia to the first. Additional names include Shincron Company Ltd, Ion-Tech Inc, Applied Vision Ltd, Professor Frank Placido of the University of Paisley and Roger Hunneman of the Department of Cybernetics, University of Reading.

Again I am grateful for all the helpful comments and suggestions from all my friends and colleagues. The enormous list of names is beyond what can be reproduced here but I must mention my debt to my old friend Professor Lee Cheng-Chung who took the trouble to work completely through the book and provided me with what has to be the most detailed list of misprints and mistakes and Professor Shigetaro Ogura who was instrumental in the translation of the second edition into Japanese. The people at Adam Hilger must be the most patient people on earth. I think finally it was my shame at so trying the endurance of Kathryn Cantley who simply responded with encouragement and understanding that drove me to complete the work.

My eternal and grateful thanks to my wife. She did not write the book but she made sure that I did.

Angus Macleod
Tucson 1999

Foreword to the second edition

A great deal has happened in the subject of optical coatings since the first edition of this book. This is especially true of facilities for thin-film calculations. In 1969 my thin-film computing was performed on an IBM 1130 computer that had a random access memory of 10 kbytes. Time had to be booked in advance, sometimes days in advance. Calculations remote from this computer were performed either by slide rule, log tables or electromechanical calculator. Nowadays my students scarcely know what a slide rule is, my pocket calculator accommodates programs that can calculate the properties of thin-film multilayers and I have on my desk a microcomputer with a random access memory of 0.5 Mbytes, which I can use as and when I like. The earlier parts of this revision were written on a mechanical typewriter. The final parts were completed on my own word processor. These advances in data processing and computing are without precedent and, of course, have had a profound and irreversible effect on many aspects of everyday life as well as on the whole field of science and technology.

There have been major developments, too, in the deposition of thin-film coatings, and although these lack the spectacular, almost explosive, character of computing programs, nevertheless important and significant advances have been made. Electron-beam sources have become the norm rather than the exception, with performance and reliability beyond anything available in 1969. Pumping systems are enormously improved, and the box-coater is now standard rather than unusual. Microprocessors control the entire operation of the pumping system and, frequently, even the deposition process. We have come to understand that many of our problems are inherent in the properties of our thin films rather than in the complexity of our designs. Microstructure and its influence on material properties is especially important. Ultimate coating performance is determined by the losses and instabilities of our films rather than the accuracy and precision of our monitoring systems.

My own circumstances have changed too. I wrote the first edition in industry. I finish the second as a university professor in a different country.

All this change has presented me with difficult problems in the revision of this book. I want to bring it up to date but do not want to lose what was useful in the first edition. I believe that in spite of the great advances in computers, there is still an important place for the appreciation of the fundamentals of thin-film coating design. Powerful synthesis and refinement techniques are available and are enormously useful, but an understanding of thin-film coating performance and the important design parameters is still an essential ingredient of success. The computer frees us from much of the previous drudgery and puts in our hands more powerful tools for improving our understanding. The availability of programmable calculators and of microcomputers implies easy handling of more complex expressions and formulae in design and performance calculations. The book, therefore, contains many more of these than did the first edition. I hope they are found useful. I have included a great deal of detail on the admittance diagram and admittance loci. I use them in my teaching and research and have taken this opportunity to write them up. SI units, rather than Gaussian, have been adopted, and I think chapter 2 is much the better for the change. There is more on coatings for oblique incidence including the admittance diagram beyond the critical angle, which explains and predicts many of the resonant effects that are observed in connection with surface plasmons, effects used by Greenland and Billington (Chapter 8, reference 12) in the late 1940s and early 1950s for monitoring thin-film deposition.

Inevitably, the first edition contained a number of mistakes and misprints and I apologise for them. Many were picked up by friends and colleagues who kindly pointed them out to me. Perhaps the worse mistake was in figure 9.4 on uniformity. The results were quoted as for a flat plate but, in fact, referred to a spherical work holder. These errors have been corrected in this edition and I hope that I have avoided making too many fresh ones. I am immensely grateful to all the people who helped in this correction process. I hope they will forgive me for not including the huge list of their names here. My thanks are also due to J J Apfel, G DeBell, E Pelletier and W T Welford who read and commented on various parts of the manuscript.

To the list in the foreword of the first edition of organisations kindly providing material should be added the names Leybold-Heraeus GmbH, and Optical Coating Laboratory Inc. Airco-Temescal is now known as Temescal, a Division of the BOC Group Inc., and the British Scientific Instrument Research Association as Sira Institute.

My publisher is still the same Adam Hilger, but now part of the Institute of Physics. I owe a very great debt to Neville Goodman who was responsible for the first edition and who also persuaded and encouraged me into the second. He retired while it was still in preparation, and the task of extracting the final manuscript from me became Jim Revill's. Ian Kingston and Brian McMahon did a tremendous job on the manuscript at a distance of 3000 miles. Their patience

with me in the delays I have caused them has been amazing.

My wife and family have once again been a great source of support and encouragement.

Angus Macleod

Newcastle-Upon-Tyne

and

Tucson

1985

Apologia to the first edition

When I first became involved with the manufacture of thin-film optical filters, I was particularly fortunate to be closely associated with Oliver Heavens, who gave me invaluable help and guidance. Although I had not at that time met him, Dr L Holland also helped me through his book, *The Vacuum Deposition of Thin Films*. Lacking, however, was a book devoted to the design and production of multilayer thin-film optical filters, a lack which I have since felt especially when introducing others to the field. Like many others in similar situations I produced from time to time notes on the subject purely for my own use. Then in 1967, I met Neville Goodman of Adam Hilger, who had apparently long been hoping for a book on optical filters in general. I was certainly not competent to write a book on this wide subject, but, in the course of conversation, the possibility of a book solely on thin-film optical filters arose. Neville Goodman's enthusiasm was infectious, and with his considerable encouragement, I dug out my notes and began writing. This, some two years and much labour later, is the result. I have tried to make it the book that I would like to have had myself when I first started in the field, and I hope it may help to satisfy also the needs of others. It is not in any way intended to compete with the existing works on optical thin films, but rather to supplement them, by dealing with one aspect of the subject which seems to be only lightly covered elsewhere.

It will be immediately obvious to even the most casual of readers that a very large proportion of the book is a review of the work of others. I have tried to acknowledge this fully throughout the text. Many of the results have been recast to fit in with the unified approach which I have attempted to adopt throughout the book. Some of the work is, I fondly imagine, completely my own, but at least a proportion of it may, unknown to me, have been anticipated elsewhere. To any authors concerned I humbly apologise, my only excuse being that I also thought of it. I promise, as far as I can, to correct the situation if ever there is a second edition. I can, however, say with complete confidence that any shortcomings of the book are entirely my own work.

Even the mere writing of the book would have been impossible without the willing help, so freely given, of a large number of friends and colleagues. Neville Goodman started the whole thing off and has always been ready with just the right sort of encouragement. David Tomlinson, also of Adam Hilger, edited the work

and adjusted it where necessary so that all sounded just as I had meant it to, but had not quite managed to achieve. The drawings were the work of Mrs Jacobi. At Grubb Parsons, Jim Mills performed all the calculations, using an IBM 1130 (he appears in the frontispiece for which I am also grateful), Fred Ritchie kindly gave me permission to quote many of his results and helped considerably by reading the manuscript, and Helen Davis transformed my almost illegible first manuscript into one which could be read without considerable strain. Stimulating discussion with John Little and other colleagues over the years has also been invaluable. Desmond Smith of Reading University kindly gave me much material especially connected with the section on atmospheric temperature sounding which he was good enough to read and correct. John Seeley and Alan Thetford, both of Reading University, helped me by amplifying and explaining their methods of design. Jim Ring, of Imperial College, read and commented on the section on astronomical applications and Dr J Meaburn kindly provided the photographs for it. Dr A F Turner gave me much information on the early history of multiple half-wave filters. It is impossible to mention by name all those others who have helped but they include: M J Shadbolt, S W Warren, A J N Hope, H Bucher and all the authors who led the way and whose work I have used and quoted.

Journals, publishers and organisations which provided and gave permission for the reproduction of material were:

Journal of the Optical Society of America (The Optical Society of America)
Applied Optics (The Optical Society of America)
Optica Acta (Taylor and Francis Limited)
Proceedings of the Physical Society (The Institute of Physics and the Physical Society)
IEEE Transactions on Aerospace (The Institute of Electrical and Electronics Engineers, Inc.)
Zeitschrift für Physik (Springer Verlag)
Bell System Technical Journal (The American Telephone and Telegraph Co.)
Philips Engineering Technical Journal (Philips Research Laboratories)
Methuen & Co. Ltd
OCLI Optical Coatings Limited
Standard Telephones and Cables Limited
Balzers Aktiengesellschaft für Hochvacuumtechnik und dünne Schichten
Edwards High Vacuum Limited
Airco Temescal (A Division of Air Reduction Company Inc.)
Hawker Siddeley Dynamics Limited
System Computers Limited
Ferranti Limited
British Scientific Instrument Research Association
And lastly, but far from least, the management of Sir Howard Grubb, Parsons & Co. Ltd, particularly Mr G M Sisson and MR G E Manville, for much material, for facilities and for permission to write this book.

To all these and to all the others, who are too numerous to name and who I hope will excuse me for not attempting to name them, I am truly grateful.

I should add that my wife and children have been particularly patient with me during the long writing process which has taken up so much of the time that would normally have been theirs. Indeed my children eventually began to worry if ever I appeared to be slacking and, by their comments, prodded me into redoubled efforts.

H A Macleod
Newcastle-Upon-Tyne
May 1969

Symbols and abbreviations

The following table gives those more important symbols used in at least several places in the text. We have tried as far as possible to create a consistent set of symbols but there are several well known and accepted symbols that are universally used in the field for certain quantities and changing them would probably lead to even greater confusion than would retaining them. This has meant that in some cases the same symbol is used in different places for different quantities. The table should make it clear. Less important symbols defined and used only in very short sections have been omitted.

A	Absorptance—the ratio of the energy absorbed in the structure to the energy incident on it.
\mathcal{A}	A quantity used in the calculation of the absorptance of dielectric assemblies. It is equivalent to $(1 - \psi)$.
B	One of the elements of the characteristic matrix of a thin-film assembly. It can be identified as a normalised electric field amplitude.
C	One of the elements of the characteristic matrix of a thin-film assembly. It can be identified as a normalised magnetic field amplitude.
d_q	The physical thickness of the q th layer in a thin-film assembly.
E	The electric vector in the electromagnetic field.
E	The amplitude of the tangential component of electric field, that is the field parallel to a boundary.
E	The equivalent admittance. See also η_E .
\mathcal{E}	The electric amplitude.
F	A function used in the theory of the Fabry–Perot interferometer.
\mathcal{F}	Finesse—the ratio of the separation of adjacent fringes to the fringe halfwidth in the Fabry–Perot interferometer.

<i>g</i>	$g = \lambda_0/\lambda = \nu/\nu_0$ sometimes called the relative wavelength of the relative wavenumber or the wavelength ratio. λ_0 and ν_0 are usually chosen to be the wavelength or wavenumber, respectively, at which the optical thicknesses of the more important layers in the assembly are quarter-waves. The phase thickness, δ , of quarter-wave layers is given by $\delta = (\pi/2)g$.
<i>H</i>	The magnetic amplitude.
<i>H</i>	The magnetic vector in the electromagnetic field.
<i>H</i>	The amplitude of the tangential component of magnetic field, that is the field parallel to a boundary.
<i>H</i>	Represents a quarter-wave of high index.
<i>I</i>	The intensity of the wave. A measure of the energy per unit area per unit time carried by the wave.
<i>k</i>	The extinction coefficient. The complex refractive index is given by $N = n - ik$. A finite value of <i>k</i> for a medium denotes the presence of absorption. See also the absorption coefficient α .
<i>L</i>	Represents a quarter-wave of low index.
<i>M</i>	Represents a quarter-wave of intermediate index.
<i>M_a</i>	A symbol denoting the elements of the characteristic matrix of layer <i>a</i> .
<i>N</i>	The complex refractive index. $N = n - ik$.
<i>n</i>	The real part of the refractive index.
<i>n*</i>	The effective index, that is the index of an equivalent layer that shifts in wavelength by the same amount as a narrowband filter when tilted with respect to the incident light.
<i>p</i>	Packing density of a film.
<i>p</i>	Indicates the plane of polarisation in which the electric vector is parallel to the plane of incidence. Equivalent to TM.
<i>R</i>	The reflectance. The ratio at a boundary of the reflected intensity to the incident intensity. At oblique incidence the components normal to the boundary are used.
<i>4</i>	Indicates the plane of polarisation in which the electric vector is normal to the plane of incidence. (From the German <i>senkrecht</i>). Equivalent to TE.
<i>T</i>	The transmittance. The ratio at a boundary of the transmitted intensity to the incident intensity. At oblique incidence the components normal to the boundary are used.
TE	Transverse electric. The plane of polarisation in which the electric vector is normal to the plane of incidence. Equivalent to s-polarisation.
TM	Transverse magnetic. The plane of polarisation in which the magnetic vector is normal to the plane of incidence. Equivalent to p-polarisation.

x, y, z	The three axes defining the orientation of a thin-film assembly. z is normally taken normal to the interfaces and with positive direction in the sense of the propagation of the incident wave, x and y in the plane of the interfaces with x also in the plane of incidence. x, y and z form a right-handed set.
$X + iZ$	The optimum exit admittance for a metal layer in order to achieve the maximum potential transmittance.
\mathcal{Y}	The admittance of free space.
y	The admittance of a medium. In SI units y is measured in siemens. $y = N\mathcal{Y}$ and so is numerically equal to the refractive index if measured in free space units.
Y	The admittance of a surface or multilayer. It is given by C/B .
y_0	The admittance of the incident medium.
y_m (y_{sub} or y_s)	The admittance of the substrate upon which the film system is deposited.
α	The absorption coefficient. The inverse of the distance along the direction of propagation in which the intensity of a wave falls to 1/e times its original value. $\alpha = 4\pi k/\lambda$ where k is the extinction coefficient.
α	A symbol used to represent $2\pi nd/\lambda$.
α, β, γ	The three direction cosines.
$(\alpha - i\beta)$	Symbols used to represent the admittance of a metal. Similar to $n - ik$.
β	A symbol used to represent $2\pi kd/\lambda$.
γ	The equivalent phase thickness of a symmetrical assembly.
Δ_q	(η_p/η_s) where η_p and η_s are modified admittances. This is a quantity used in the design of polarisation-free coatings.
ε	Indicates a small error or a departure from a reference value of a number.
ε	The permittivity of a medium.
η	The tilted optical admittance.
η_m	The tilted admittance of the substrate. See y_m .
η_E	The equivalent admittance of a symmetrical assembly. See also E .
θ	The angle of incidence in a medium.
θ_0	The angle of incidence in the incident medium.
λ	The wavelength of the light, usually the wavelength in free space.
λ_0	The reference wavelength. See g .
ν_0	The reference wavenumber. $\nu_0 = 1/\lambda_0$. See g .
ρ	The amplitude reflection coefficient.
ρ	The electric charge density.
γ	The amplitude transmission coefficient.
ϕ	The phase shift on reflection.

ψ Potential transmittance. $\psi = T/(1 - R)$. ψ Used in some limited calculations to represent $2\delta_p/\delta_q$.

Chapter 1

Introduction

This book is intended to form an introduction to thin-film optical filters for both the manufacturer and the user. It does not pretend to present a detailed account of the entire field of thin-film optics, but it is hoped that it will form a supplement to those works already available in the field and which only briefly touch on the principles of filters. For the sake of a degree of completeness, it has been thought desirable to repeat again some of the information that will be found elsewhere in textbooks, referring the reader to more complete sources for greater detail. The topics covered are a mixture of design, manufacture, performance and application, including enough of the basic mathematics of optical thin films for the reader to carry out thin-film calculations. The aim has been to present, as far as possible, a unified treatment, and there are some alternative methods of analysis which are not discussed.

When the book was first written there were just a few books available that covered aspects of the field. Now the situation has changed somewhat and there is an array of relevant books. Some of these are listed in the bibliography at the end of this chapter. However, the half-life of a work these days is so short that knowledge can actually disappear. It is well worthwhile taking the time to go back to some of the earlier books. Heavens [1], Holland [2], Anders [3], Knittl [4] are just some of those that will repay study, and they are listed in the bibliography along with some more recent volumes.

In a work of this size, it is not possible to cover the entire field of thin-film optical devices in the detail that some of them may deserve. The selection of topics is due, at least in part, to the author's own preferences and knowledge. Optical filters have been interpreted fairly broadly to include such items as antireflection and high-reflectance coatings.

1.1 Early history

The earliest of what might be called modern thin-film optics was the discovery by Robert Boyle and Robert Hooke, independently, of the phenomenon now known

as ‘Newton’s rings’. The explanation of this is nowadays thought to be a very simple matter, being due to interference in a single thin film of varying thickness. However, at that time, the theory of the nature of light was not sufficiently far advanced, and the explanation of this, and a number of similar observations made in the same period by Sir Isaac Newton on thin films, eluded scientists for almost a further 150 years. Then, on 12 November 1801, in a Bakerian Lecture to the Royal Society, Thomas Young enunciated the principle of the interference of light and produced the first satisfactory explanation of the effect. As Henry Crew [5] has put it, ‘This simple but tremendously important fact that two rays of light incident upon a single point can be added together to produce darkness at that point is, as I see it, the one outstanding discovery which the world owes to Thomas Young.’

Young’s theory was far from achieving universal acceptance. Indeed Young became the victim of a bitter personal attack, against which he had the greatest difficulty defending himself. Recognition came slowly and depended much on the work of Augustin Jean Fresnel [6] who, quite independently, also arrived at a wave theory of light. Fresnel’s discovery, in 1816, that two beams of light which are polarised at right angles could never interfere, established the transverse nature of light waves. Then Fresnel combined Young’s interference principle and Huygens’s ideas of light propagation into an elegant theory of diffraction. It was Fresnel who put the wave theory of light on such a firm foundation that it has never been shaken. For the thin-film worker, Fresnel’s laws, governing the amplitude and phase of light reflected and transmitted at a single boundary, are of major importance. Knittl [7] has reminded us that it was Fresnel who first summed an infinite series of rays to determine the transmittance of a thick sheet of glass and that it was Simeon Denis Poisson, in correspondence with Fresnel, who included interference effects in the summation to arrive at the important results that a half-wave thick film does not change the reflectance of a surface, and that a quarter-wave thick film of index $(n_0 n_1)^{1/2}$ will reduce to zero the reflectance of a surface between two media of indices n_1 and n_0 . Fresnel died in 1827, at the early age of 39.

In 1873, the great work of James Clerk Maxwell, *A Treatise on Electricity and Magnetism* [8], was published, and in his system of equations we have all the basic theory for the analysis of thin-film optical problems.

Meanwhile, in 1817, Joseph Fraunhofer had made what were probably the first ever antireflection coatings. It is worth quoting his observations at some length because they show the considerable insight that he had, even at that early date, into the physical causes of the effects that were produced. The following is a translation of part of the paper as it appears in the collected works [9].

Before I quote the experiments which I have made on this I will give the method which I have made use of to tell in a short time whether the glass will withstand the influence of the atmosphere. If one grinds and then polishes, as finely as possible, one surface of glass which has become etched through long exposure to the atmosphere, then wets one part of the surface, for example half, with concentrated sulphuric or nitric acid

and lets it work on the surface for 24 hours, one finds after cleaning away the acid that that part of the surface on which the acid was, reflects much less light than the other half, that is it shines less although it is not in the least etched and still transmits as much light as the other half, so that one can detect no difference on looking through. The difference in the amount of reflected light will be most easily detected if one lets the light strike approximately vertically. It is the greater the more the glass is liable to tarnish and become etched. If the polish on the glass is not very good this difference will be less noticeable. On glass which is not liable to tarnish, the sulphuric and nitric acid does not work. Through this treatment with sulphuric or nitric acid some types of glasses get on their surfaces beautiful vivid colours which alter like soap bubbles if one lets the light strike at different angles.

Then, in an appendix to the paper added in 1819:

Colours on reflection always occur with all transparent media if they are very thin. If for example, one spreads polished glass thinly with alcohol and lets it gradually evaporate, towards the end of the evaporation, colours appear as with tarnished glass. If one spreads a solution of gum-lac in a comparatively large quantity of alcohol very thinly over polished warmed metal the alcohol will very quickly evaporate, and the gum-lac remains behind as a transparent hard varnish which shows colours if it is thinly enough laid on. Since the colours, in glasses which have been coloured through tarnishing, alter themselves if the inclination of the incident light becomes greater or smaller, there is no doubt that these colours are quite of the same nature as those of soap bubbles, and those which occur through the contact of two polished flat glass surfaces, or generally as thin transparent films of material. Thus there must be on the surface of tarnished glass which shows colours, a thin layer of glass which is different in refractive power from the underlying. Such a situation must occur if a component is partly removed from the surface of the glass or if a component of the glass combines at the surface with a related material into a new transparent product.

It seems that Fraunhofer did not follow up this particular line into the development of an antireflection coating for glass, perhaps because optical components were not, at that time, sufficiently complicated for the need for antireflection coatings to be obvious. Possibly the important point that, not only was the reflectance less, but the transmittance also greater had escaped him.

In 1886, Lord Rayleigh reported to the Royal Society an experimental verification of Fresnel's reflection law at near-normal incidence [10]. In order to attain a sufficiently satisfactory agreement between measurement and prediction, he had found it necessary to use freshly polished glass because the reflectance

of older material, even without any visible signs of tarnish, was too low. One possible explanation which he suggested was the formation, on the surface, of a thin layer of different refractive index from the underlying material. He was apparently unaware of the earlier work of Fraunhofer.

Then, in 1891, Dennis Taylor published the first edition of his famous book *On the Adjustment and Testing of Telescopic Objectives* and mentioned [11, 12] that ‘as regards the tarnish which we have above alluded to as being noticeable upon the flint lens of an ordinary objective after a few years of use, we are very glad to be able to reassure the owner of such a flint that this film of tarnish, generally looked upon with suspicion, is really a very good friend to the observer, inasmuch as it increases the transparency of his objective’.

In fact, Taylor went on to develop a method of artificially producing the tarnish by chemical etching [13]. This work was followed up by Kollmorgen, who developed the chemical process still further for different types of glasses [14].

At the same time, in the nineteenth century, a great deal of progress was being made in the field of interferometry. The most significant development, from the thin-film point of view, was the Fabry–Perot interferometer [15] described in 1899, which has become one of the basic structures for thin-film filters.

Developments became much more rapid in the 1930s, and indeed it is in this period that we can recognise the beginnings of the modern thin-film optical coating. In 1932, Rouard [16] observed that a very thin metallic film reduced the internal reflectance of a glass plate, although the external reflectance was increased. In 1934, Bauer [17], in the course of fundamental investigations of the optical properties of halides, produced reflection-reducing coatings, and Pfund [18] evaporated zinc sulphide layers to make low-loss beam splitters for Michelson interferometers, noting, incidentally, that titanium dioxide could be a better material. In 1936, John Strong [19] produced antireflection coatings by evaporation of fluorite to give inhomogeneous films which reduced the reflectance of glass to visible light by as much as 89%, a most impressive figure. Then, in 1939, Geffcken [20] constructed the first thin-film metal–dielectric interference filters. A fascinating account of Geffcken’s work is given by Thelen [21] who describes Geffcken’s search for improved antireflection coatings and his creation of the famous quarter–half–quarter design.

The most important factor in this sudden expansion of thin-film optical coatings was the manufacturing process. Although sputtering was discovered around the middle of the nineteenth century, and vacuum evaporation around the beginning of the twentieth, they were not considered as useful manufacturing processes. The main difficulty was the lack of really suitable pumps, and it was not until the early 1930s that the work of C R Burch on diffusion pump oils made it possible for this process to be used satisfactorily. Since then, tremendous strides have been made, particularly in the last few years. Filters with greater than 100 layers are not uncommon and uses have been found for them in almost every branch of science and technology.

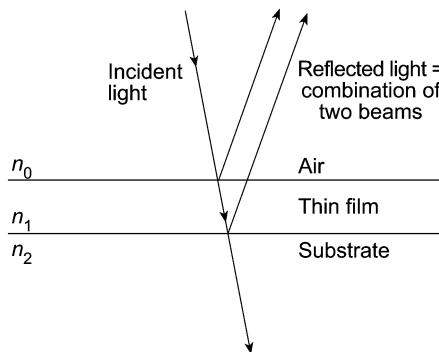


Figure 1.1. A single thin film.

1.2 Thin-film filters

To understand in a qualitative way the performance of thin-film optical devices, it is necessary to accept several simple statements. The first is that the amplitude reflectance of light at any boundary between two media is given by $(1-\rho)/(1+\rho)$, where ρ is the ratio of the optical admittances at the boundary, which, in the optical region, is also the ratio of the refractive indices. The reflectance (the ratio of irradiances or intensities) is the square of this quantity. The second is that there is a phase shift of 180° when the reflectance takes place in a medium of lower refractive index than the adjoining medium, and zero if the medium has a higher index than the one adjoining it. The third is that if light is split into two components by reflection at the top and bottom surfaces of a thin film, then the beams will recombine in such a way that the resultant amplitude will be the difference of the amplitudes of the two components if the relative phase shift is 180° , or the sum of the amplitudes if the relative phase shift is either zero or a multiple of 360° . In the former case, we say that the beams interfere destructively and in the latter constructively. Other cases where the phase shift is different will be intermediate between these two possibilities.

The antireflection coating depends for its operation on the more or less complete cancellation of the light reflected at the upper and lower of the two surfaces of the thin film. Let the index of the substrate be n_{sub} , that of the film n_1 , and that of the incident medium, which will in almost all cases be air, n_0 . For complete cancellation of the two beams of light, the amplitudes of the light reflected at the upper and lower boundaries of the film should be equal, which implies that the ratios of the refractive indices at each boundary should be equal, i.e. $n_0/n_1 = n_1/n_{\text{sub}}$, or $n_1 = (n_0 n_{\text{sub}})^{1/2}$. This shows that the index of the thin film should be intermediate between the indices of air, which may be taken as unity, and of the substrate, which may be taken as at least 1.52. At both the upper and lower boundaries of the antireflection film, the reflection takes place in

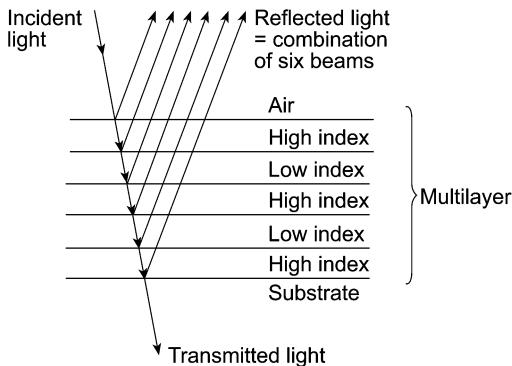


Figure 1.2. A multilayer.

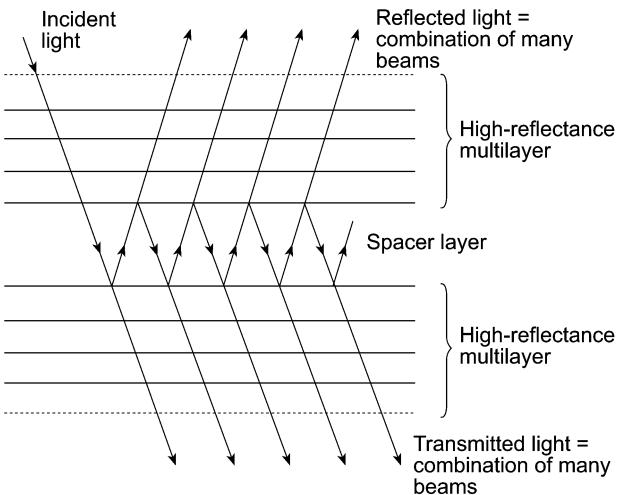


Figure 1.3. A Fabry-Perot filter showing multiple reflections in the spacer layer.

a medium of lower refractive index than the adjoining medium. Thus, to ensure that the relative phase shift is 180° so that the beams cancel, the optical thickness of the film should be made one quarter wavelength.

A simple antireflection coating should, therefore, consist of a single film of refractive index equal to the square root of that of the substrate, and of optical thickness one quarter of a wavelength. As will be explained in the chapter on antireflection coatings, there are other improved coatings covering wider wavelength ranges involving greater numbers of layers.

Another basic type of thin-film structure is a stack of alternate high- and low-index films, all one quarter wavelength thick (see figure 1.2). Light reflected within the high-index layers will not suffer any phase shift on reflection, while that

reflected within the low-index layers will suffer a change of 180° . It is fairly easy to see that the various components of the incident light produced by reflection at successive boundaries throughout the assembly will reappear at the front surface all in phase so that they will recombine constructively. This implies that the effective reflectance of the assembly can be made very high indeed, as high as may be desired, merely by increasing the number of layers. This is the basic form of the high-reflectance coating. When such a coating is constructed, it is found that the reflectance remains high over only a limited range of wavelengths, depending on the ratio of high and low refractive indices. Outside this zone, the reflectance changes abruptly to a low value. Because of this behaviour, the quarter-wave stack is used as a basic building block for many types of thin-film filters. It can be used as a longwave-pass filter, a shortwave-pass filter, a bandstop filter, a straightforward high-reflectance coating, for example in laser mirrors, and as a reflector in a thin-film Fabry–Perot interferometer (figure 1.3), which is another basic filter type described in some detail in chapters 5 and 7. Here, it is sufficient to say that it consists of a spacer or cavity layer which is usually half a wavelength thick, bounded by two high-reflectance coatings. Multiple-beam interference in the spacer or cavity layer causes the transmission of the filter to be extremely high over a narrow band of wavelengths around that for which the spacer is a multiple of one half wavelength thick. It is possible, as with lumped electric circuits, to couple two or more Fabry–Perot filters in series to give a more rectangular pass band.

In the great majority of cases the thin films are completely transparent, so that no energy is absorbed. The filter characteristic in reflection is the complement of that in transmission. This fact is used in the construction of such devices as dichroic beam splitters for colour primary separation in, for example, colour television cameras.

This brief description has neglected the effect of multiple reflections in most of the layers and, for an accurate evaluation of the performance of a filter, these extra reflections must be taken into account. This involves extremely complex calculations and an alternative, and more effective, approach has been found in the development of entirely new forms of solution of Maxwell's equations in stratified media. This is, in fact, the principal method used in chapter 2 where basic theory is considered. The solution appears as a very elegant product of 2×2 matrices, each matrix representing a single film. Unfortunately, in spite of the apparent simplicity of the matrices, calculation by hand of the properties of a given multilayer, particularly if there are absorbing layers present and a wide spectral region is involved, is an extremely tedious and time-consuming task. The preferred method of calculation is to use a computer. This makes calculation so rapid and straightforward that it makes little sense to use anything else. Even pocket calculators, especially the programmable kind, can be used to great effect. However, in spite of the enormous power of the modern computer it is still true that skill and experience play a major part in successful coating design. The computer brings little in the way of understanding. Understanding

is the emphasis in the bulk of this book. There are many techniques that date back to times when computers were expensive, cumbersome and scarce, and alternatives, usually approximate, were required. These would not be used for calculation today but they bring an insight that straightforward calculation cannot deliver, even if it is very fast. Thus we include many such techniques and it is convenient to introduce them often in a historical context. The matrix method itself brings many advantages. For example, it has made possible the development of exceedingly powerful design techniques based on the algebraic manipulation of the matrices themselves. These are also included. Graphical techniques are of considerable usefulness in visualisation of the properties of coatings. There are many such techniques but in this book we pay particular attention to one such method known as the admittance diagram. This is one that the author has found of considerable assistance over the years. It is an accurate technique in the sense that it contains no approximations other than those involved perhaps in sketching it, but it is used normally as an aid to understanding rather than as a calculation tool.

In the design of a thin-film multilayer, we are required to find an arrangement of layers which will give a performance specified in advance, and this is much more difficult than straightforward calculation of the properties of a given multilayer. There is no analytical solution to the general problem. The normal method of design is to arrive at a possible structure for a filter, using techniques which will be described, and which consist of a mixture of analysis, experience and the use of well-known building blocks. The evaluation is then completed by calculating the performance on a computer. Depending on the results of the computations, adjustments to the proposed design may be made, then recomputed, until a satisfactory solution is found. This adjustment process can itself be undertaken by a computer and is often known by the term ‘refinement’. A related term is synthesis, which implies an element of construction as well as adjustment. The ultimate in synthesis would be the complete construction of a design with no starting information beyond the performance specification, but, at the present state of the art it is normal to provide some starting information, such as materials to be used, total thickness of coating and, perhaps, a very rough starting design.

The successful application of refinement techniques depends largely on a starting solution that has a performance close to that required. Under these conditions it has been made to work exceedingly well. The operation of a refinement process involves the adjustment of the parameters of the system to minimise a merit coefficient (in some less common versions a measure of merit may be maximised) representing the gap between the performance achieved by the design at any stage and the desired performance. The main difference between the various techniques is in the details of the rules used to control and adjust the design. A major problem is the enormous number of parameters that can potentially be involved. Refinement is usually kept within bounds by limiting the search to small changes in an almost acceptable starting design. In synthesis with no starting design, the possibilities are virtually infinite, and so the rules

governing the search procedure have to be very carefully organised. The most effective techniques incorporate two elements, an effective refinement technique that operates until it reaches a limit and a procedure for complicating the design that is then applied. These two elements alternate as the design is gradually constructed. Automatic design synthesis is undoubtedly increasing in importance in step with developments in computers, but it is still true that in the hands of a skilled practitioner the achievements of both refinement and synthesis are much more impressive than when no skill is involved. Someone who knows well what he or she is doing will succeed much better than someone who does not. This branch of the subject is much more a matter of computing technique rather than fundamental to the understanding of thin-film filters, and so it is largely outside the scope of this book. The book by Liddell [22] and the more recent text by Furman and Tikhonravov [23] give good accounts of various methods. The real limitation to what is, at the present time, possible in optical thin-film filters and coatings is the capability of the manufacturing process to produce layers of precisely the correct optical constants and thickness, rather than any deficiency in design techniques.

The common techniques for the construction of thin-film optical coatings can be classified as physical vapour deposition. They are vacuum processes where a solid film condenses from the vapour phase. The most straightforward and the traditional method is known as thermal evaporation and this is still much used. Because of defects of solidity possessed by thermally evaporated films there has, in recent years, been a shift, now accelerating, towards what are described as the energetic processes. Here, mechanical momentum is transferred to the growing film, either by deliberate bombardment or by an increase in the momentum of the arriving film material, and this added momentum drives the outermost material deeper into the film, increasing its solidity. These processes are described briefly in the later chapters of the book but much more information will be found in the books listed in the bibliography at the end of this chapter.

References

- [1] Heavens O S 1955 *Optical Properties of Thin Solid Films* (London: Butterworths)
- [2] Holland L 1956 *Vacuum Deposition of Thin Films* (London: Chapman and Hall)
- [3] Anders H 1965 *Dünne Schichten für die Optik* (Stuttgart: Wissenschaftliche Verlagsgesellschaft) [English translation 1967 *Thin Films in Optics* (Focal Press)]
- [4] Knittl Z 1976 *Optics of Thin Films* (London: Wiley)
- [5] Crew H 1930 Thomas Young's place in the history of the wave theory of light *J. Opt. Soc. Am.* **20** 3–10
- [6] Senarmont H d, Verdet E and Fresnel L (ed) 1866–1870 *Oeuvres Complètes d'Augustin Fresnel* (Paris: Imprimerie Impériale)
- [7] Knittl Z 1978 Fresnel historique et actuel *Opt. Acta* **25** 167–73
- [8] Maxwell J C 1873 *A Treatise on Electricity and Magnetism* (Oxford: Clarendon)
- [9] Fraunhofer J v 1817 Versuche über die Ursachen des Anlaufens und Mattwerdens

des Glases und die Mittel, denselben zuvorzukommen *Joseph von Fraunhofer's Gesammelte Schriften* (München)

- [10] Rayleigh L 1886 On the intensity of light reflected from certain surfaces at nearly perpendicular incidence *Proc. R. Soc.* **41** 275–94
- [11] Taylor H D 1891 *On the Adjustment and Testing of Telescopic Objectives* (T Cooke)
- [12] Taylor H D 1983 *The Adjustment and Testing of Telescopic Objectives* 5th edn (Bristol: Adam Hilger)
- [13] Taylor H D 1904 *Lenses* UK Patent 29561
- [14] Kollmorgen F 1916 Light transmission through telescopes *Trans. Am. Illumin. Eng. Soc.* **11** 220–8
- [15] Fabry C and Perot A 1899 Théorie et applications d'une nouvelle méthode de spectroscopie interférentielle *Ann. Chim. Phys.* **16** 115–44
- [16] Rouard P 1932 Sur le pouvoir réflecteur des métaux en lames très minces *Contes Rendus de l'Academie de Science* **195** 869–72
- [17] Bauer G 1934 Absolutwerte der optischen Absorptionskonstanten von Alkalihalogenidkristallen im Gebiet ihrer ultravioletten Eigenfrequenzen *Ann. Phys.* **19** 434–64
- [18] Pfund A H 1934 Highly reflecting films of zinc sulphide *J. Opt. Soc. Am.* **24** 99–102
- [19] Strong J 1936 On a method of decreasing the reflection from non-metallic substances *J. Opt. Soc. Am.* **26** 73–4
- [20] Geffcken W 1939 *Interferenzlichtfilter* Germany Patent 716153
- [21] Thelen A 1997 The pioneering contributions of W Geffcken *Thin Films on Glass* ed H Bach and D Krause (Berlin: Springer) pp 227–39
- [22] Liddell H M 1981 *Computer-Aided Techniques for the Design of Multilayer Filters* (Bristol: Adam Hilger)
- [23] Furman S A and Tikhonravov A V 1992 *Basics of Optics of Multilayer Systems* 1st edn (Gif-sur-Yvette: Editions Frontières)

Bibliography

A complete bibliography of primary references would stretch to an enormous length. This list is, therefore, primarily one of secondary references wherever possible. Primary references are given usually only where secondary references are difficult to obtain or do not exist.

- Anders H 1965 *Dünne Schichten für die Optik* (Stuttgart: Wissenschaftliche Verlagsgesellschaft) [English translation 1967 *Thin Films in Optics* (Focal Press)]
- Bach H and Krause D (ed) 1997 *Thin Films on Glass* (Berlin: Springer)
- Flory F R (ed) 1995 Thin films for optical systems 1 *Optical Engineering* ed B J Thomson (New York: Marcel Dekker) p 49
- Frey H and Kienel G (ed) 1987 *Dünnenschicht Technologie* (Düsseldorf: VDI-Verlag)
- Furman S A and Tikhonravov A V 1992 *Basics of Optics of Multilayer Systems* 1st edn (Gif-sur-Yvette: Editions Frontières)
- Hartnagel H L, Dawar A L, Jain A K and Jagadish C 1995 *Semiconducting Transparent Thin Films* (Bristol: Institute of Physics)
- Heavens O S 1955 *Optical Properties of Thin Solid Films* (London: Butterworths)
- Holland L 1956 *Vacuum Deposition of Thin Films* (London: Chapman and Hall)

- Hummel R E and Guenther K H (ed) 1995 Thin films for optical coatings *Handbook of Optical Properties* 1st edn (Boca Raton, FL: Chemical Rubber Company)
- Jacobson M R (ed) 1989 *Deposition of Optical Coatings 1 (SPIE Milestone Series)* ed B J Thompson (Bellingham: SPIE) MS 6
- Jacobson M R (ed) 1990 *Design of Optical Coatings (SPIE Milestone Series)* ed B J Thompson (Bellingham: SPIE) MS 26
- Jacobson M R (ed) 1992 *Characterization of Optical Coatings 1 (SPIE Milestone Series)* ed B J Thompson (Bellingham: SPIE) MS 63
- Knittl Z 1976 *Optics of Thin Films* (London: Wiley)
- Liddell H M 1981 *Computer-Aided Techniques for the Design of Multilayer Filters* (Bristol: Adam Hilger)
- Lissberger P H 1970 Optical applications of dielectric thin films *Rep. Prog. Phys.* **33** 197–268
- Pulker H K 1984 *Coatings on Glass* (Amsterdam: Elsevier)
- Rancourt J D 1987 *Optical Thin films: Users' Handbook* (New York: Macmillan)
- Vasicek A 1960 *Optics of Thin Films* (Amsterdam: North-Holland)
- Willey R R 1996 *Practical Design and Production of Optical Thin Films* (New York: Marcel Dekker)

Chapter 2

Basic theory

This next part of the book is a long and rather tedious account of some basic theory which is necessary in order to make calculations of the properties of multilayer thin-film coatings. It is perhaps worth reading just once, or when some deeper insight into thin-film calculations is required. In order to make it easier for those who have read it to find the basic results, or for those who do not wish to read it at all to proceed with the remainder of the book, the principal results are summarised, beginning on page 46.

2.1 Maxwell's equations and plane electromagnetic waves

For those readers who are still with us we begin our attack on thin-film problems by solving Maxwell's equations together with the appropriate material equations. In isotropic media these are:

$$\text{curl } \mathbf{H} = \mathbf{j} + \partial \mathbf{D} / \partial t \quad (2.1)$$

$$\text{curl } \mathbf{E} = -\partial \mathbf{B} / \partial t \quad (2.2)$$

$$\text{div } \mathbf{D} = \rho \quad (2.3)$$

$$\text{div } \mathbf{B} = 0 \quad (2.4)$$

$$\mathbf{j} = \sigma \mathbf{E} \quad (2.5)$$

$$\mathbf{D} = \epsilon \mathbf{E} \quad (2.6)$$

$$\mathbf{B} = \mu \mathbf{H}. \quad (2.7)$$

In anisotropic media, equations (2.1) to (2.7) become much more complicated with σ , ϵ and μ being tensor rather than scalar quantities. Anisotropic media are covered by Yeh [1] and Hodgkinson and Wu [2].

The International System of Units (SI) is used as far as possible throughout this book. Table 2.1 shows the definitions of the quantities in the equations together with the appropriate SI units.

Table 2.1.

Symbol	Physical quantity	SI unit	Symbol for SI unit
E	Electric field strength	volts per metre	V m^{-1}
D	Electric displacement	coulombs per square metre	C m^{-2}
H	Magnetic field strength	amperes per metre	A m^{-1}
j	Electric current density	amperes per square metre	A m^{-2}
B	Magnetic flux density or magnetic induction	tesla	T
ρ	Electric charge density	coulombs per cubic metre	C m^{-3}
σ	Electric conductivity	siemens per metre	S m^{-1}
μ	Permeability	henries per metre	H m^{-1}
ε	Permittivity	farads per metre	F m^{-1}

Table 2.2.

Symbol	Physical quality	Value
c	Speed of light in a vacuum	$2.997925 \times 10^8 \text{ m s}^{-1}$
μ_0	Permeability of a vacuum	$4\pi \times 10^{-7} \text{ H m}^{-1}$
ε_0	Permittivity of a vacuum ($= \mu_0^{-1} c^{-2}$)	$8.8541853 \times 10^{-12} \text{ F m}^{-1}$

To the equations we can add

$$\varepsilon = \varepsilon_r \varepsilon_0 \quad (2.8)$$

$$\mu = \mu_r \mu_0 \quad (2.9)$$

$$\varepsilon_0 = 1/(\mu_0 c^2) \quad (2.10)$$

where ε_0 and μ_0 are the permittivity and permeability of free space, respectively. ε_r and μ_r are the relative permittivity and permeability, and c is a constant that can be identified as the velocity of light in free space. ε_0 , μ_0 and c are important constants, the values of which are given in table 2.2.

The following analysis is brief and incomplete. For a full, rigorous treatment of the electromagnetic field equations the reader is referred to Born and Wolf [3].

$$\operatorname{div} \mathbf{D} = 0$$

and, solving for \mathbf{E}

$$\nabla^2 \mathbf{E} = \varepsilon \mu \frac{\partial^2 \mathbf{E}}{\partial t^2} + \mu \sigma \frac{\partial \mathbf{E}}{\partial t}. \quad (2.11)$$

A similar expression holds for \mathbf{H} .

First of all we look for a solution of equation (2.11) in the form of a plane-polarised plane harmonic wave, and we choose the complex form of this wave, the physical meaning being associated with the real part of the expression.

$$\mathbf{E} = \mathcal{E} \exp[i\omega(t - x/v)] \quad (2.12)$$

represents such a wave propagating along the x axis with velocity v . \mathcal{E} is the vector amplitude and ω the angular frequency of this wave. The advantage of the complex form of the wave is that phase changes can be dealt with very readily by including them in a complex amplitude. If we include a relative phase, φ , in (2.12) then it becomes

$$\mathbf{E} = \mathcal{E} \exp[i\{\omega(t - x/v) + \varphi\}] = \mathcal{E} \exp(i\varphi) \exp[i\omega(t - x/v)] \quad (2.13)$$

where $\mathcal{E} \exp(i\varphi)$ is the complex vector amplitude. The complex scalar amplitude is given by $\mathcal{E} \exp(i\varphi)$ where $\mathcal{E} = |\mathcal{E}|$. Equation (2.13), which has phase φ relative to expression (2.12), is simply expression (2.12) with the amplitude replaced by the complex amplitude.

For equation (2.12) to be a solution of equation (2.11) it is necessary that

$$\omega^2/v^2 = \omega^2 \varepsilon \mu - i\omega \mu \sigma. \quad (2.14)$$

In a vacuum we have $\sigma = 0$ and $v = c$, so that from equation (2.14)

$$c^2 = 1/\varepsilon_0 \mu_0 \quad (2.15)$$

which is identical to equation (2.10). Multiplying equation (2.14) by equation (2.15) and dividing through by ω^2 , we obtain

$$\frac{c^2}{v^2} = \frac{\varepsilon \mu}{\varepsilon_0 \mu_0} - i \frac{\mu \sigma}{\omega \varepsilon_0 \mu_0},$$

where c/v is clearly a dimensionless parameter of the medium, which we denote by N :

$$N^2 = \varepsilon_r \mu_r - i \frac{\mu_r \sigma}{\omega \varepsilon_0}. \quad (2.16)$$

This implies that N is of the form

$$N = c/v = n - ik. \quad (2.17)$$

There are two possible values of N from (2.16), but for physical reasons we choose that which gives a positive value of n . N is known as the complex refractive index, n as the real part of the refractive index (or often simply as the refractive index because N is real in an ideal dielectric material) and k is known as the extinction coefficient.

From equation (2.16)

$$n^2 - k^2 = \epsilon_r \mu_r \quad (2.18)$$

$$2nk = \frac{\mu_r \sigma}{\omega \epsilon_0}. \quad (2.19)$$

Equation (2.12) can now be written

$$\mathbf{E} = \mathcal{E} \exp[i\omega t - (2\pi N/\lambda)x], \quad (2.20)$$

where we have introduced the wavelength in free space, $\lambda (= 2\pi c/\omega)$.

Substituting $n - ik$ for N in equation (2.20) gives

$$\mathbf{E} = \mathcal{E} \exp[-(2\pi k/\lambda)x] \exp[i\omega t - (2\pi n/\lambda)x] \quad (2.21)$$

and the significance of k emerges as being a measure of absorption in the medium. The distance $\lambda/(2\pi k)$ is that in which the amplitude of the wave falls to $1/e$ of its original value. The way in which the power carried by the wave falls off will be considered shortly.

The change in phase produced by a traversal of distance x in the medium is the same as that produced by a distance nx in a vacuum. Because of this, nx is known as the optical distance, as distinct from the physical or geometrical distance. Generally, in thin-film optics one is more interested in optical distances and optical thicknesses than in geometrical ones.

Equation (2.16) represents a plane-polarised plane wave propagating along the x axis. For a similar wave propagating in a direction given by direction coefficient (α, β, γ) the expression becomes

$$\mathbf{E} = \mathcal{E} \exp[i\omega t - (2\pi N/\lambda)(\alpha x + \beta y + \gamma z)]. \quad (2.22)$$

This is the simplest type of wave in an absorbing medium. In an assembly of absorbing thin films, we shall see that we are occasionally forced to adopt a slightly more complicated expression for the wave.

There are some important relationships for this type of wave which can be derived from Maxwell's equations. Let the direction of propagation of the wave be given by unit vector \hat{s} where

$$\hat{s} = \alpha \mathbf{i} + \beta \mathbf{j} + \gamma \mathbf{k}$$

and where \mathbf{i} , \mathbf{j} and \mathbf{k} are unit vectors along the x , y and z axes, respectively. From equation (2.22) we have

$$\partial \mathbf{E} / \partial t = i\omega \mathbf{E}$$

and from equations (2.1), (2.5) and (2.6)

$$\begin{aligned} \text{curl } \mathbf{H} &= \sigma \mathbf{E} + \epsilon \partial \mathbf{E} / \partial t \\ &= (\sigma + i\omega \epsilon) \mathbf{E} \\ &= i \frac{\omega N^2}{c^2 \mu} \mathbf{E}. \end{aligned}$$

Now

$$\text{curl} = \left(\frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k} \right) \times$$

where \times denotes the vector product. But

$$\begin{aligned} \frac{\partial}{\partial x} &= -i \frac{2\pi N}{\lambda} \alpha = -i \frac{\omega N}{c} \alpha, \\ \frac{\partial}{\partial y} &= -i \frac{\omega N}{c} \beta, \quad \frac{\partial}{\partial z} = -i \frac{\omega N}{c} \gamma \end{aligned}$$

so that

$$\text{curl } \mathbf{H} = -i \frac{\omega N}{c} (\hat{s} \times \mathbf{H}).$$

Then

$$-i \frac{\omega N}{c} (\hat{s} \times \mathbf{H}) = i \frac{\omega N^2}{c^2 \mu} \mathbf{E},$$

i.e.

$$(\hat{s} \times \mathbf{H}) = -\frac{N}{c\mu} \mathbf{E} \quad (2.23)$$

and similarly

$$\frac{N}{c\mu} (\hat{s} \times \mathbf{E}) = \mathbf{H}. \quad (2.24)$$

For this type of wave, therefore, \mathbf{E} , \mathbf{H} and \hat{s} are mutually perpendicular and form a right-handed set. The quantity $N/c\mu$ has the dimensions of an admittance and is known as the characteristic optical admittance of the medium, written y . In free space it can be readily shown that the optical admittance is given by

$$y = (\epsilon_0 / \mu_0)^{1/2} = 2.6544 \times 10^{-3} \text{ S.} \quad (2.25)$$

Now

$$\mu = \mu_r \mu_0 \quad (2.26)$$

and at optical frequencies μ_r is unity so that we can write

$$y = N \mathcal{Y} \quad (2.27)$$

and

$$\mathbf{H} = y (\hat{s} \times \mathbf{E}) = N \mathcal{Y} (\hat{s} \times \mathbf{E}). \quad (2.28)$$

2.1.1 The Poynting vector

An important feature of electromagnetic radiation is that it is a form of energy transport, and it is the energy associated with the wave which is normally observed. The instantaneous rate of flow of energy across unit area is given by the Poynting vector

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}. \quad (2.29)$$

The direction of the vector is the direction of energy flow.

This expression is nonlinear (\mathbf{E} is multiplied by \mathbf{H}) and so we cannot use directly the complex form of the wave, which is not valid for nonlinear operations. Either the real or the imaginary part of the wave expression should be inserted. The instantaneous value of the Poynting vector oscillates at twice the frequency of the wave and it is its mean value which is significant. This is defined as the irradiance or, in the older systems of units, intensity. In the SI system of units it is measured in watts per square metre. An unfortunate feature of the SI system, for our purposes, is that the symbol for irradiance is E . Use of this symbol would make it very difficult for us to distinguish between irradiance and electric field. Since both are extremely important in almost everything we do we must be able to differentiate between them, and so we adopt a nonstandard symbol, I , for irradiance. For a harmonic wave we find that we can derive a very attractive and simple expression for the irradiance using the complex form of the wave. This is

$$\mathbf{I} = \frac{1}{2}\text{Re}(\mathbf{E} \times \mathbf{H}^*), \quad (2.30)$$

where $*$ denotes complex conjugate. It should be emphasised that the complex form must be used in equation (2.30). The irradiance \mathbf{I} is written in (2.30) as a vector quantity, when it has the same direction as the flow of energy of the wave. The more usual scalar irradiance I is simply the magnitude of \mathbf{I} . Since \mathbf{E} and \mathbf{H} are perpendicular, equation (2.30) can be written

$$I = \frac{1}{2}\text{Re}(E H^*), \quad (2.31)$$

where E and H are the scalar magnitudes.

It is important to note that the electric and magnetic vectors in equation (2.30) should be the total resultant fields due to all the waves which are involved. This is implicit in the derivation of the Poynting vector expression. We will return to this point when calculating reflectance and transmittance.

For a single, homogeneous, harmonic wave of the form (2.22):

$$\mathbf{H} = y(\hat{s} \times \mathbf{E})$$

so that

$$\begin{aligned} \mathbf{I} &= \text{Re}\left(\frac{1}{2}y\mathbf{E}\mathbf{E}^*\hat{s}\right) \\ &= \frac{1}{2}n\gamma\mathbf{E}\mathbf{E}^*\hat{s}. \end{aligned} \quad (2.32)$$

Now, from equation (2.22), the magnitude of E is given by

$$\begin{aligned} \mathbf{E} &= \mathcal{E} \exp[i\{\omega t - (2\pi[n - ik]/\lambda)(\alpha x + \beta y + \gamma z)\}] \\ &= \mathcal{E} \exp[-(2\pi k/\lambda)(\alpha x + \beta y + \gamma z)] \exp[i\{\omega t - (2\pi n/\lambda)(\alpha x + \beta y + \gamma z)\}] \end{aligned}$$

implying

$$\mathbf{E}\mathbf{E}^* = \mathcal{E}\mathcal{E}^* \exp[-(4\pi k/\lambda)(\alpha x + \beta y + \gamma z)]$$

and

$$I = \frac{1}{2}n\mathcal{Y}|\mathcal{E}|^2 \exp[-(4\pi k/\lambda)(\alpha x + \beta y + \gamma z)].$$

The expression $(\alpha x + \beta y + \gamma z)$ is simply the distance along the direction of propagation, and thus the irradiance drops to $1/e$ of its initial value in a distance given by $\lambda/4\pi k$. The inverse of this distance is defined as the absorption coefficient α , that is

$$\alpha = 4\pi k/\lambda. \quad (2.33)$$

The absorption coefficient α should not be confused with the direction cosine.

However,

$$|\mathcal{E}| \exp[-(2\pi k/\lambda)(\alpha x + \beta y + \gamma z)]$$

is really the amplitude of the wave at the point (x, y, z) so that a much simpler way of writing the expression for irradiance is

$$I = \frac{1}{2}n\mathcal{Y}(\text{amplitude})^2 \quad (2.34)$$

or

$$I \propto n \times (\text{amplitude})^2. \quad (2.35)$$

This expression is a better form than the more usual

$$I \propto (\text{amplitude})^2. \quad (2.36)$$

The expression will frequently be used for comparing irradiances, in calculating reflectance or transmittance, for example, and if the media in which the two waves are propagating are of different index; errors will occur unless n is included as above.

2.2 The simple boundary

Thin-film filters usually consist of a number of boundaries between various homogeneous media and it is the effect which these boundaries will have on an incident wave which we will wish to calculate. A single boundary is the simplest

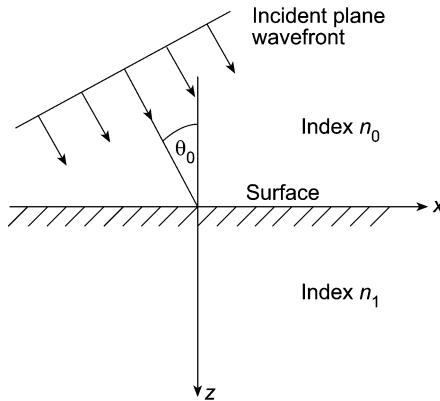


Figure 2.1. Plane wavefront incident on a single surface.

case. First of all we consider absorption-free media, i.e. $k = 0$. The arrangement is sketched in figure 2.1. At a boundary, the tangential components of \mathbf{E} and \mathbf{H} , that is, the components along the boundary, are continuous across it. In this case, the boundary is defined by $z = 0$, and the tangential components must be continuous for all values of x , y and t .

Let us retain our plane-polarised plane harmonic form for the incident wave; we can be safe in assuming that this wave will be split into a reflected wave and a transmitted wave at the boundary, and our objective is the calculation of the parameters of these waves. Without specifying their exact form for the moment, we can, however, be certain that they will consist of an amplitude term and a phase factor. The amplitude terms will not be functions of x , y or r , any variations due to these being included in the phase factors.

Let the direction cosines of the \hat{s} vectors of the transmitted and reflected waves be $(\alpha_t, \beta_t, \gamma_t)$ and $(\alpha_r, \beta_r, \gamma_r)$ respectively. We can then write the phase factors in the form:

$$\begin{aligned} \text{Incident wave} & \quad \exp\{i[\omega_i t - (2\pi n_0/\lambda_i)(x \sin \vartheta_0 + z \cos \vartheta_0)]\} \\ \text{Reflected wave} & \quad \exp\{i[\omega_r t - (2\pi n_0/\lambda_r)(\alpha_r x + \beta_r y + \gamma_r z)]\} \\ \text{Transmitted wave} & \quad \exp\{i[\omega_t t - (2\pi n_1/\lambda_t)(\alpha_t x + \beta_t y + \gamma_t z)]\}. \end{aligned}$$

The relative phases of these waves are included in the complex amplitudes. For waves with these phase factors to satisfy the boundary conditions for all x , y , t at $z = 0$ implies that the coefficients of these variables must be separately identically equal:

$$\omega \equiv \omega_r \equiv \omega_t$$

that is, there is no change of frequency in reflection or refraction and hence no change in free space wavelength either. This implies that the free space

wavelengths are equal:

$$\lambda \equiv \lambda_r \equiv \lambda_t.$$

Next

$$0 \equiv n_0 \beta_r \equiv n_1 \beta_t$$

that is, the directions of the reflected and transmitted or refracted beams are confined to the plane of incidence. This, in turn, means that the direction cosines of the reflected and transmitted waves are of the form

$$\alpha = \sin \vartheta \quad \gamma = \cos \vartheta. \quad (2.37)$$

Also

$$n_0 \sin \vartheta_0 \equiv n_0 \alpha_r \equiv n_1 \alpha_t$$

so that if the angles of reflection and refraction are ϑ_r and ϑ_t , respectively, then

$$\vartheta_0 = \vartheta_r \quad (2.38)$$

that is, the angle of reflection equals the angle of incidence, and

$$n_0 \sin \vartheta_0 = n_1 \sin \vartheta_t.$$

The result appears more symmetrical if we replace ϑ_t by ϑ_1 , giving

$$n_0 \sin \vartheta_0 = n_1 \sin \vartheta_1 \quad (2.39)$$

which is the familiar relationship known as Snell's law. γ_r and γ_t are then given either by equation (2.37) or by

$$\alpha_r^2 + \gamma_r^2 = 1 \quad \text{and} \quad \alpha_t^2 + \gamma_t^2 = 1. \quad (2.40)$$

Note that for the reflected beam we must choose the negative root of (2.40) so that the beam will propagate in the correct direction.

2.2.1 Normal incidence

Let us limit our initial discussion to normal incidence and let the incident wave be a plane-polarised plane harmonic wave. The coordinate axes are shown in figure 2.2. The xy plane is the plane of the boundary. The incident wave we can take as propagating along the z axis with the positive direction of the E vector along the x axis. Then the positive direction of the H vector will be the y axis. It is clear that the only waves which satisfy the boundary conditions are plane polarised in the same plane as the incident wave.

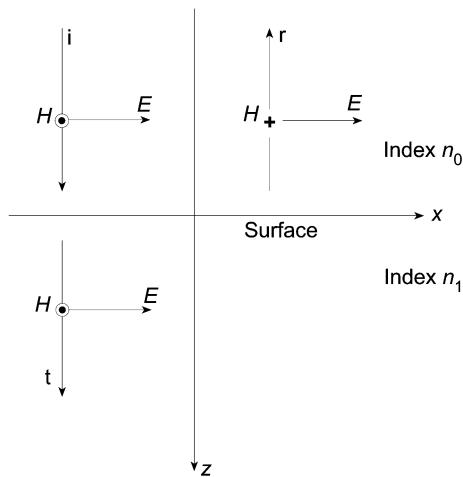


Figure 2.2. Convention defining positive directions of the electric and magnetic vectors for reflection and transmission at an interface at normal incidence.

A quoted phase difference between two waves travelling in the same direction is immediately meaningful. A phase difference between two waves travelling in opposite directions is absolutely meaningless, unless a reference plane at which the phase difference is measured is first defined. This is simply because the phase difference between oppositely propagating waves of the same frequency has a term $(\pm 4\pi ns/\lambda)$ in it where s is a distance measured along the direction of propagation. Before proceeding further, therefore, we need to define the reference point for measurements of relative phase between the oppositely propagating beams.

Then there is another problem. The waves have electric and magnetic fields that with the direction of propagation form right-handed sets. Since the direction of propagation is reversed in the reflected beam, the orientation of electric and magnetic fields cannot remain the same as that in the incident beam, otherwise we would no longer have a right-handed set. We need to decide on how we are going to handle this. Since the electric field is the one that is most important from the point of view of interaction with matter, we will define our directions with respect to it.

The matter of phase references and electric field directions are what we call conventions because we do have complete freedom of choice, and any self-consistent arrangement is possible. We must simply ensure that once we have made our choice we adhere to it. A good rule, however, is never to make things difficult when we can make them easy, and so we will normally choose the rule that is most convenient and least complicated. We define the positive direction of \mathbf{E} along the x axis for all the beams that are involved. Because of this choice, the positive direction of the magnetic vector will be along the y axis for the

incident and transmitted waves, but along the negative direction of the y axis for the reflected wave.

We are now in a position to apply the boundary conditions. Since we have already made sure that the phase factors are satisfactory, we have only to consider the amplitudes, and we will be including any phase changes in these.

(a) Electric vector continuous across the boundary:

$$\mathcal{E}_i + \mathcal{E}_r = \mathcal{E}_t. \quad (2.41)$$

(b) Magnetic vector continuous across the boundary:

$$\mathcal{H}_i - \mathcal{H}_r = \mathcal{H}_t$$

where we must use a minus sign because of our convention for positive directions. The relationship between magnetic and electric field through the characteristic admittance gives

$$y_0 \mathcal{E}_i - y_0 \mathcal{E}_r = y_1 \mathcal{E}_t. \quad (2.42)$$

This can also be derived using the vector relationship (2.28) and (2.41). We can eliminate \mathcal{E}_t to give

$$y_1(\mathcal{E}_i + \mathcal{E}_r) = y_0(\mathcal{E}_i - \mathcal{E}_r),$$

i.e.

$$\frac{\mathcal{E}_r}{\mathcal{E}_i} = \frac{y_0 - y_1}{y_0 + y_1} = \frac{n_0 - n_1}{n_0 + n_1} \quad (2.43)$$

the second part of the relationship being correct only because at optical frequencies we can write

$$y = n\mathcal{Y}.$$

Similarly, eliminating \mathcal{E}_r ,

$$\frac{\mathcal{E}_t}{\mathcal{E}_i} = \frac{2y_0}{y_0 + y_1} = \frac{2n_0}{n_0 + n_1}. \quad (2.44)$$

These quantities are called the amplitude reflection and transmission coefficients and are denoted by ρ and τ respectively. Thus

$$\rho = \frac{y_0 - y_1}{y_0 + y_1} = \frac{n_0 - n_1}{n_0 + n_1} \quad (2.45)$$

$$\tau = \frac{2y_0}{y_0 + y_1} = \frac{2n_0}{n_0 + n_1}. \quad (2.46)$$

In this particular case, all y real, these two quantities are real. τ is always a positive real number, indicating that according to our phase convention there is no phase shift between the incident and transmitted beams at the interface. The

behaviour of ρ indicates that there will be no phase shift between the incident and reflected beams at the interface provided $n_0 > n_1$, but that if $n_0 < n_1$ there will be a phase change of π because the value of ρ becomes negative.

We now examine the energy balance at the boundary. Since the boundary is of zero thickness, it can neither supply energy to nor extract energy from the various waves. The Poynting vector will therefore be continuous across the boundary, so that we can write:

$$\begin{aligned}\text{net irradiance} &= \operatorname{Re} \left[\frac{1}{2} (\mathcal{E}_i + \mathcal{E}_r)(y_0 \mathcal{E}_i - y_0 \mathcal{E}_r)^* \right] \\ &= \operatorname{Re} \left[\frac{1}{2} \mathcal{E}_i (y_1 \mathcal{E}_t)^* \right]\end{aligned}$$

[using $\operatorname{Re}(\frac{1}{2} \mathbf{E} \times \mathbf{H}^*)$ and equations (2.41) and (2.42)]. Now

$$\mathcal{E}_r = \rho \mathcal{E}_i \quad \text{and} \quad \mathcal{E}_t = \tau \mathcal{E}_i,$$

i.e.

$$\text{net irradiance} = \frac{1}{2} y_0 \mathcal{E}_i \mathcal{E}_i^* (1 - \rho^2) = \frac{1}{2} y_0 \mathcal{E}_i \mathcal{E}_i^* (y_1/y_0) \tau^2. \quad (2.47)$$

Now, $(1/2)y_0 \mathcal{E}_i \mathcal{E}_i^*$ is the irradiance of the incident beam I_i . We can identify $\rho^2(1/2)y_0 \mathcal{E}_i \mathcal{E}_i^* = \rho^2 I_i$ as the irradiance of the reflected beam I_r and $(y_1/y_0) \times \tau^2(1/2)y_0 \mathcal{E}_i \mathcal{E}_i^* = (y_1/y_0)\tau^2 I_i$ as the irradiance of the transmitted beam I_t . We define the reflectance R as the ratio of the reflected and incident irradiances and the transmittance T as the ratio of the transmitted and incident irradiances. Then

$$\begin{aligned}T &= \frac{I_t}{I_i} = \frac{y_1 \tau^2}{y_0} = \frac{4 y_0 y_1}{(y_0 + y_1)^2} = \frac{4 n_0 n_1}{(n_0 + n_1)^2} \\ R &= \frac{I_r}{I_i} = \rho^2 = \left(\frac{y_0 - y_1}{y_0 + y_1} \right)^2 = \left(\frac{n_0 - n_1}{n_0 + n_1} \right)^2.\end{aligned} \quad (2.48)$$

From equation (2.47) we have, using equations (2.48),

$$(1 - R) = T. \quad (2.49)$$

Equations (2.47), (2.48) and (2.49) are therefore consistent with our ideas of splitting the irradiances into incident, reflected and transmitted irradiances which can be treated as separate waves, the energy flow into the second medium being simply the difference of the incident and reflected irradiances. Remember that all this, so far, assumes that there is no absorption. We shall shortly see that the situation changes slightly when absorption is present.

2.2.2 Oblique incidence

Now let us consider oblique incidence, still retaining our absorption-free media. For any general direction of the vector amplitude of the incident wave we quickly

find that the application of the boundary conditions leads us into complicated and difficult expressions for the vector amplitudes of the reflected and transmitted waves. Fortunately there are two orientations of the incident wave which lead to reasonably straightforward calculations: the vector electrical amplitudes aligned in the plane of incidence (i.e. the xy plane of figure 2.1) and the vector electrical amplitudes aligned normal to the plane of incidence (i.e. parallel to the y axis in figure 2.1). In each of these cases, the orientations of the transmitted and reflected vector amplitudes are the same as for the incident wave. Any incident wave of arbitrary polarisation can therefore be split into two components having these simple orientations. The transmitted and reflected components can be calculated for each orientation separately and then combined to yield the resultant. Since, therefore, it is necessary to consider two orientations only, they have been given special names. A wave with the electric vector in the plane of incidence is known as p-polarised or, sometimes, as TM (for transverse magnetic) and a wave with the electric vector normal to the plane of incidence as s-polarised or, sometimes, TE (for transverse electric). The p and s are derived from the German parallel and senkrecht (perpendicular). Before we can actually proceed to the calculation of the reflected and transmitted amplitudes, we must choose the various reference directions of the vectors from which any phase differences will be calculated. We have, once again, complete freedom of choice, but once we have established the convention we must adhere to it, just as in the normal incidence case. The conventions which we will use in this book are illustrated in figure 2.3. They have been chosen to be compatible with those for normal incidence already established. In some works, an opposite convention for the p-polarised reflected beam has been adopted, but this leads to an incompatibility with results derived for normal incidence, and we prefer to avoid this situation. (Note that for reasons connected with consistency of reference directions for elliptically polarised light, the normal convention in ellipsometric calculations is opposite to that of figure 2.3 for reflected p-polarised light. When ellipsometric parameters are compared with the results of the expressions we shall use, it will usually be necessary to introduce a shift of 180° in the p-polarised reflected results.)

We can now apply the boundary conditions. Since we have already ensured that the phase factors will be correct, we need only consider the vector amplitudes.

2.2.2.1 *p*-polarised light

(a) Electric component parallel to the boundary, continuous across it:

$$\mathcal{E}_i \cos \vartheta_0 + \mathcal{E}_r \cos \vartheta_0 = \mathcal{E}_t \cos \vartheta_1. \quad (2.50)$$

(b) Magnetic component parallel to the boundary and continuous across it. Here we need to calculate the magnetic vector amplitudes, and we can do this either by using equation (2.28) to operate on equation (2.50) directly, or, since the magnetic vectors are already parallel to the boundary we can use figure 2.3 and then convert,

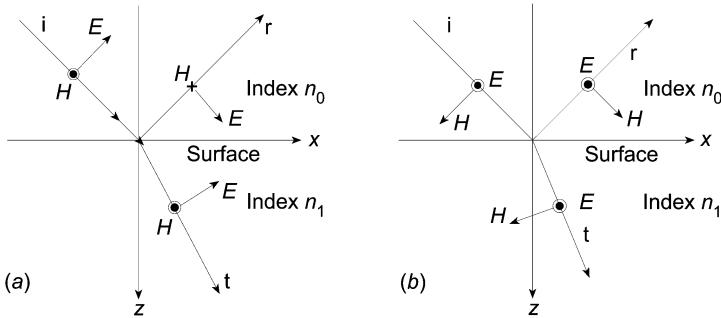


Figure 2.3. (a) Convention defining the positive directions of the electric and magnetic vectors for p-polarised light (TM waves). (b) Convention defining the positive directions of the electric and magnetic vectors for s-polarised light (TE waves).

since $\mathcal{H} = y\mathcal{E}$:

$$y_0\mathcal{E}_i - y_0\mathcal{E}_r = y_1\mathcal{E}_t. \quad (2.51)$$

At first sight it seems logical just to eliminate first \mathcal{E}_t and then \mathcal{E}_r from these two equations to obtain $\mathcal{E}_r/\mathcal{E}_i$ and $\mathcal{E}_t/\mathcal{E}_i$

$$\begin{aligned} \frac{\mathcal{E}_r}{\mathcal{E}_i} &= \frac{y_0 \cos \vartheta_1 - y_1 \cos \vartheta_0}{y_0 \cos \vartheta_1 + y_1 \cos \vartheta_0} \\ \frac{\mathcal{E}_t}{\mathcal{E}_i} &= \frac{2y_0 \cos \vartheta_0}{y_0 \cos \vartheta_1 + y_1 \cos \vartheta_0} \end{aligned} \quad (2.52)$$

and then simply to set

$$R = \left(\frac{\mathcal{E}_r}{\mathcal{E}_i} \right)^2 \quad \text{and} \quad T = \frac{y_1}{y_0} \left(\frac{\mathcal{E}_t}{\mathcal{E}_i} \right)^2$$

but when we calculate the expressions which result, we find that $R + T \neq 1$. In fact, there is no mistake in the calculations. We have computed the irradiances measured along the direction of propagation of the waves and the transmitted wave is inclined at an angle which differs from that of the incident wave. This leaves us with the problem that to adopt these definitions will involve the rejection of the ($R + T = 1$) rule.

We could correct this situation by modifying the definition of T to include this angular dependence, but an alternative, preferable and generally adopted approach is to use the components of the energy flows which are normal to the boundary. The \mathbf{E} and \mathbf{H} vectors that are involved in these calculations are then parallel to the boundary. Since these are those that enter directly into the boundary it seems appropriate to concentrate on them when we are dealing with the amplitudes of the waves.

The thin-film approach to all this, then, is to use the components of \mathbf{E} and \mathbf{H} parallel to the boundary, what are called the tangential components, in the expressions ρ and τ that involve amplitudes. Note that the normal approach in other areas of optics is to use the full components of \mathbf{E} and \mathbf{H} in amplitude expressions but to use the components of irradiance in reflectance and transmittance. The amplitude coefficients are then known as the Fresnel coefficients. The thin-film coefficients are not the Fresnel coefficients except at normal incidence, although the only coefficient that actually has a different value is the amplitude transmission coefficient for p-polarisation.

The tangential components of \mathbf{E} and \mathbf{H} , that is, the components parallel to the boundary, have already been calculated for use in equations (2.50) and (2.51). However, it is convenient to introduce special symbols for them: E and H .

Then we can write

$$E_i = \mathcal{E}_i \cos \vartheta_0 \quad H_i = \mathcal{H}_i = y_0 \mathcal{E}_i = \frac{y_0}{\cos \vartheta_0} E_i \quad (2.53)$$

$$E_r = \mathcal{E}_r \cos \vartheta_0 \quad H_r = \frac{y_0}{\cos \vartheta_0} E_r \quad (2.54)$$

$$E_t = \mathcal{E}_t \cos \vartheta_1 \quad H_t = \frac{y_1}{\cos \vartheta_1} E_t. \quad (2.55)$$

The orientations of these vectors are exactly the same as for normally incident light.

Equations (2.50) and (2.51) can then be written as follows.

(a) Electric field parallel to the boundary:

$$E_i + E_r = E_t$$

(b) Magnetic field parallel to the boundary:

$$\frac{y_0}{\cos \vartheta_0} H_i - \frac{y_0}{\cos \vartheta_0} H_r = \frac{y_1}{\cos \vartheta_1} H_t$$

giving us, by a process exactly similar to that we have already used for normal incidence,

$$\rho_p = \frac{E_r}{E_i} = \left(\frac{y_0}{\cos \vartheta_0} - \frac{y_1}{\cos \vartheta_1} \right) / \left(\frac{y_0}{\cos \vartheta_0} + \frac{y_1}{\cos \vartheta_1} \right) \quad (2.56)$$

$$\tau_p = \frac{E_t}{E_i} = \left(\frac{2y_0}{\cos \vartheta_0} \right) / \left(\frac{y_0}{\cos \vartheta_0} + \frac{y_1}{\cos \vartheta_1} \right) \quad (2.57)$$

$$R_p = \left[\left(\frac{y_0}{\cos \vartheta_0} - \frac{y_1}{\cos \vartheta_1} \right) / \left(\frac{y_0}{\cos \vartheta_0} + \frac{y_1}{\cos \vartheta_1} \right) \right]^2 \quad (2.58)$$

$$T_p = \left(\frac{4y_0 y_1}{\cos \vartheta_0 \cos \vartheta_1} \right) / \left(\frac{y_0}{\cos \vartheta_0} + \frac{y_1}{\cos \vartheta_1} \right)^2, \quad (2.59)$$

where $y_0 = n_0 \mathcal{Y}$ and $y_1 = n_1 \mathcal{Y}$ and the ($R + T = 1$) rule is retained. The suffix p has been used in the above expressions to denote p-polarisation.

It should be noted that the expression for τ_p is now different from that in equation (2.52), the form of the Fresnel amplitude transmission coefficient. Fortunately, the reflection coefficients in equations (2.52) and (2.58) are identical, and since much more use is made of reflection coefficients confusion is rare.

2.2.2.2 s-polarised light

In the case of s-polarisation the amplitudes of the components of the waves parallel to the boundary are

$$\begin{aligned} E_i &= \mathcal{E}_i & H_i &= \mathcal{H}_i \cos \vartheta_0 = y_0 \cos \vartheta_0 E_i \\ E_r &= \mathcal{E}_r & H_r &= \mathcal{H}_r \cos \vartheta_0 = y_0 \cos \vartheta_0 E_r \\ E_t &= \mathcal{E}_t & H_t &= y_1 \cos \vartheta_1 E_t \end{aligned}$$

and here we have again an orientation of the tangential components exactly as for normally incident light, and so a similar analysis leads to

$$\rho_s = \frac{E_r}{E_i} = (y_0 \cos \vartheta_0 - y_1 \cos \vartheta_1) / (y_0 \cos \vartheta_0 + y_1 \cos \vartheta_1) \quad (2.60)$$

$$\tau_s = \frac{E_t}{E_i} = (2y_0 \cos \vartheta_0) / (y_0 \cos \vartheta_0 + y_1 \cos \vartheta_1) \quad (2.61)$$

$$R_s = [(y_0 \cos \vartheta_0 - y_1 \cos \vartheta_1) / (y_0 \cos \vartheta_0 + y_1 \cos \vartheta_1)]^2 \quad (2.62)$$

$$T_s = (4y_0 \cos \vartheta_0 y_1 \cos \vartheta_1) / (y_0 \cos \vartheta_0 + y_1 \cos \vartheta_1)^2 \quad (2.63)$$

where once again $y_0 = n_0 \mathcal{Y}$ and $y_1 = n_1 \mathcal{Y}$ and the ($R + T = 1$) rule is retained. The suffix s has been used in the above expressions to denote s-polarisation.

2.2.3 The optical admittance for oblique incidence

The expressions which we have derived so far have been in their traditional form (except for the use of the tangential components rather than the full vector amplitudes) and they involve the characteristic admittances of the various media, or their refractive indices together with the admittance of free space, \mathcal{Y} . However, the notation is becoming increasingly cumbersome and will appear even more so when we consider the behaviour of thin films.

Equation (2.28) gives $\mathbf{H} = y(\hat{s} \times \mathbf{E})$ where $y = N\mathcal{Y}$ is the optical admittance. We have found it convenient to deal with E and H , the components of \mathbf{E} and \mathbf{H} parallel to the boundary, and so we introduce a tilted optical admittance η which connects E and H as

$$\eta = \frac{H}{E}. \quad (2.64)$$

At normal incidence $\eta = y = n\mathcal{Y}$ while at oblique incidence

$$\eta_p = \frac{y}{\cos \vartheta} = \frac{n\mathcal{Y}}{\cos \vartheta} \quad (2.65)$$

$$\eta_s = y \cos \vartheta = n\mathcal{Y} \cos \vartheta \quad (2.66)$$

where the ϑ and the y in (2.65) and (2.66) are those appropriate to the particular medium. In particular, Snell's law, equation (2.39), must be used to calculate ϑ . Then, in all cases, we can write

$$\rho = \left(\frac{\eta_0 - \eta_1}{\eta_0 + \eta_1} \right) \quad \tau = \left(\frac{2\eta_0}{\eta_0 + \eta_1} \right) \quad (2.67)$$

$$R = \left(\frac{\eta_0 - \eta_1}{\eta_0 + \eta_1} \right)^2 \quad T = \frac{4\eta_0\eta_1}{(\eta_0 + \eta_1)^2}. \quad (2.68)$$

These expressions can be used to compute the variation of reflectance of simple boundaries between extended media. Examples are shown in figure 2.4 of the variation of reflectance with angle of incidence. In this case, there is no absorption in the material and it can be seen that the reflectance for p-polarised light (TM) falls to zero at a definite angle. This particular angle is known as the Brewster angle and is of some importance. There are many applications where the windows of a cell must have close to zero reflection loss. When it can be arranged that the light will be linearly polarised a plate tilted at the Brewster angle will be a good solution. The light that is reflected at the Brewster angle is also linearly polarised with electric vector normal to the plane of incidence. This affords a way of identifying the absolute direction of polarisers and analysers—very difficult in any other way.

The expression for the Brewster angle can be derived as follows. For the p-reflectance to be zero, from equation (2.58)

$$\frac{y_0}{\cos \vartheta_0} = \frac{n_0 \mathcal{Y}}{\cos \vartheta_0} = \frac{y_1}{\cos \vartheta_1} = \frac{n_1 \mathcal{Y}}{\cos \vartheta_1}.$$

Snell's law gives another relationship between ϑ_0 and ϑ_1 :

$$n_0 \sin \vartheta_0 = n_1 \sin \vartheta_1.$$

Eliminating ϑ_1 from these two equations gives an expression for ϑ_0

$$\tan \vartheta_0 = n_1/n_0. \quad (2.69)$$

Note that this derivation depends on the relationship $y = n\mathcal{Y}$ valid at optical frequencies.

Nomograms which connect the angle of incidence ϑ referred to an incident medium of unity refractive index, the refractive index of a dielectric film and the optical admittance of the film at ϑ are reproduced in figure 2.5.

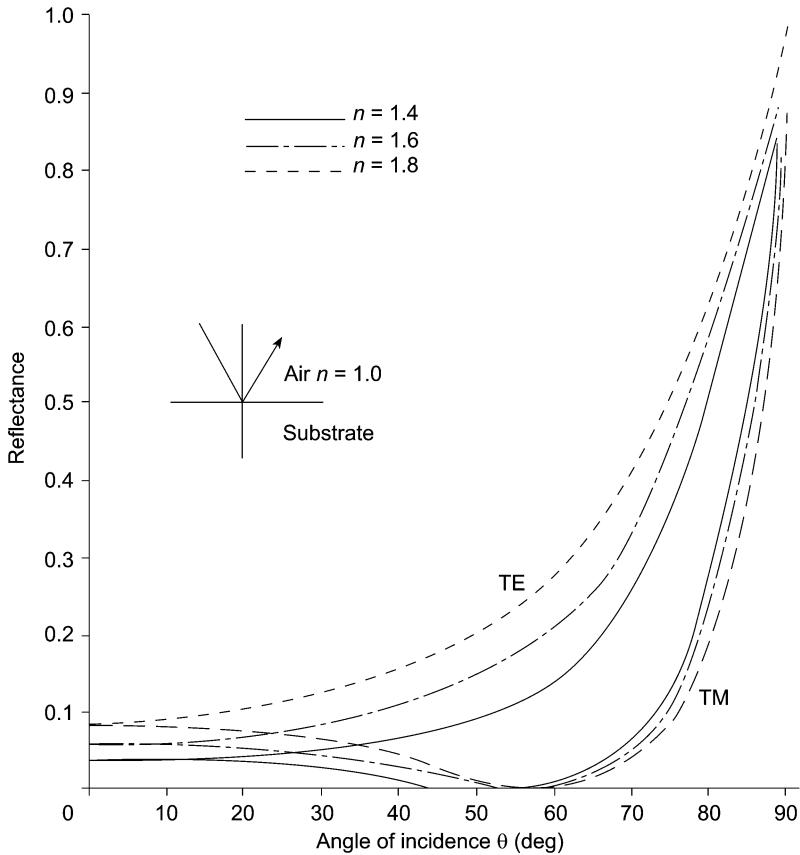


Figure 2.4. Variation of reflectance with angle of incidence for various values of refractive index. p-Reflectance (TM) to be zero, from equation (2.37).

2.2.4 Normal incidence in absorbing media

We must now examine the modifications necessary in our results in the presence of absorption. First we consider the case of normal incidence and write

$$\begin{aligned} N_0 &= n_0 - ik_0 \\ N_1 &= n_1 - ik_1 \\ y_0 &= N_0 \mathcal{Y} = (n_0 - ik_0) \mathcal{Y} \\ y_1 &= N_1 \mathcal{Y} = (n_1 - ik_1) \mathcal{Y}. \end{aligned}$$

The analysis follows that for absorption-free media. The boundary conditions are, as before:

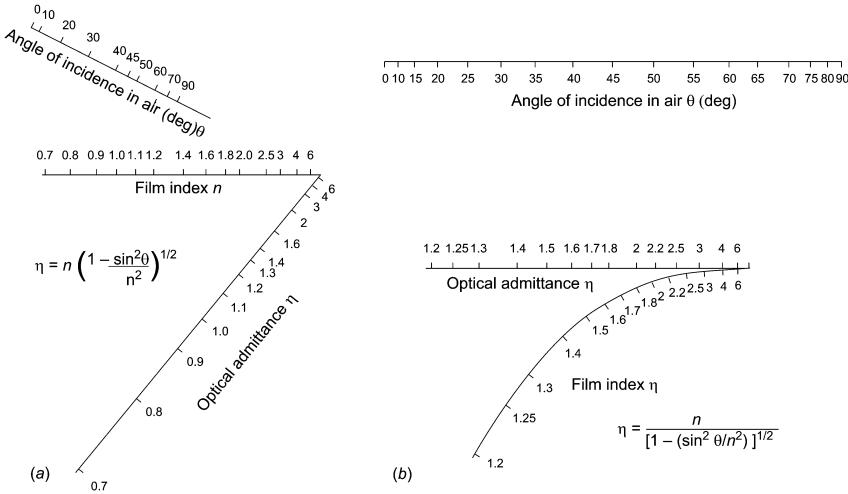


Figure 2.5. (a) Nomogram giving variation of optical admittance with angle of incidence for s-polarised light (TE waves). (b) Nomogram giving variation of optical admittance with angle of incidence for p-polarised light (TM waves).

(a) Electric vector continuous across the boundary:

$$\mathcal{E}_i + \mathcal{E}_r = \mathcal{E}_t.$$

(b) Magnetic vector continuous across the boundary:

$$y_0 \mathcal{E}_i - y_0 \mathcal{E}_r = y_1 \mathcal{E}_t$$

and eliminating first \mathcal{E}_t and then \mathcal{E}_r we obtain the expressions for the amplitude coefficients

$$\begin{aligned} \rho &= \frac{\mathcal{E}_r}{\mathcal{E}_i} = \frac{y_0 - y_1}{y_0 + y_1} = \frac{(n_0 - ik_0)\mathcal{Y} - (n_1 - ik_1)\mathcal{Y}}{(n_0 - ik_0)\mathcal{Y} + (n_1 - ik_1)\mathcal{Y}} \\ &= \frac{(n_0 - n_1) - i(k_0 - k_1)}{(n_0 + n_1) - i(k_0 + k_1)} \end{aligned} \quad (2.70)$$

$$\begin{aligned} \tau &= \frac{\mathcal{E}_t}{\mathcal{E}_i} = \frac{2y_0}{y_0 - y_1} = \frac{2(n_0 - ik_0)\mathcal{Y}}{(n_0 - ik_0)\mathcal{Y} + (n_1 - ik_1)\mathcal{Y}} \\ &= \frac{2(n_0 - ik_0)}{(n_0 + n_1) - i(k_0 + k_1)}. \end{aligned} \quad (2.71)$$

Our troubles begin when we try to extend this to reflectance and transmittance. We remain at normal incidence. Following the method for the absorption-free case, we compute the Poynting vector at the boundary in each medium and equate the

two values obtained. In the incident medium the resultant electric and magnetic fields are

$$\mathcal{E}_i + \mathcal{E}_r = \mathcal{E}_i(1 + \rho)$$

and

$$\mathcal{H}_i - \mathcal{H}_r = y_0(1 - \rho)\mathcal{E}_i,$$

respectively, where we have used the notation for tangential components, and in the second medium the fields are $\tau\mathcal{E}_i$ and $y_1\tau\mathcal{E}_i$ respectively. Then the net irradiance on either side of the boundary is

$$\begin{aligned} \text{Medium 0: } I &= \operatorname{Re} \left\{ \frac{1}{2} [\mathcal{E}_i(1 + \rho)][y_0^*(1 - \rho^*)\mathcal{E}_i^*] \right\} \\ \text{Medium 1: } I &= \operatorname{Re} \left\{ \frac{1}{2} [\tau\mathcal{E}_i][y_1^*\tau^*\mathcal{E}_i^*] \right\}. \end{aligned}$$

We then equate these two values which gives, at the boundary,

$$\begin{aligned} \operatorname{Re} \left[\frac{1}{2} y_0^* \mathcal{E}_i \mathcal{E}_i^* (1 + \rho - \rho^* - \rho\rho^*) \right] &= \frac{1}{2} \operatorname{Re}(y_1) \tau \tau^* \mathcal{E}_i \mathcal{E}_i^* \\ \frac{1}{2} \operatorname{Re}(y_0^*) \mathcal{E}_i \mathcal{E}_i^* - \frac{1}{2} \operatorname{Re}(y_0^*) \rho \rho^* \mathcal{E}_i \mathcal{E}_i^* + \frac{1}{2} \operatorname{Re}[y_0^*(\rho - \rho^*)] \mathcal{E}_i \mathcal{E}_i^* \\ &= \frac{1}{2} \operatorname{Re}(y_1) \tau \tau^* \mathcal{E}_i \mathcal{E}_i^*. \end{aligned} \quad (2.72)$$

We can replace the different parts of the expression (2.72) with their normal interpretations to give

$$I_i - RI_i + \frac{1}{2} \operatorname{Re}[y_0^*(\rho - \rho^*)] \mathcal{E}_i \mathcal{E}_i^* = TI_i, \quad (2.73)$$

where $(\rho - \rho^*)$ is imaginary. This implies that if y_0 is real the third term in (2.73) is zero. The other terms then make up the incident, the reflected and the transmitted irradiances, and these balance. If y_0 is complex then its imaginary part will combine with the imaginary $(\rho - \rho^*)$ to produce a real result that will imply that $T + R \neq 1$. The irradiances involved in the analysis are those actually at the boundary, which is of zero thickness, and it is impossible that it should either remove or donate energy to the waves. Our assumption that the irradiances can be divided into separate incident, reflected and transmitted irradiances is therefore incorrect. The source of the difficulty is a coupling between the incident and reflected fields which occurs only in an absorbing medium and which must be taken into account when computing energy transport. The expressions for the amplitude coefficients are perfectly correct. The explanation has been given by a number of people. The account by Berning [4] is probably the most accessible.

The extra term is of the order of (k^2/n^2) . For any reasonable experiment to be carried out the incident medium must be sufficiently free of absorption for the necessary comparative measurements to be performed with acceptably small errors. Although we will certainly be dealing with absorbing media in thin-film assemblies, our incident media will never be heavily absorbing and it will not be a serious lack of generality if we assume that our incident media are absorption-free. Since our expressions for the amplitude coefficients are valid, then any calculations of amplitudes in absorbing media will be correct. We simply have to ensure that calculations of reflectances are carried out in a transparent medium. With this restriction, then, we have

$$R = \left(\frac{y_0 - y_1}{y_0 + y_1} \right) \left(\frac{y_0 - y_1}{y_0 + y_1} \right)^* \quad (2.74)$$

$$T = \frac{4y_0 \operatorname{Re}(y_1)}{(y_0 + y_1)(y_0 + y_1)^*} \quad (2.75)$$

where y_0 is real.

We have avoided the problem connected with the definition of reflectance in a medium with complex y_0 simply by not defining it unless the incident medium is sufficiently free of either gain or absorption. Without a definition of reflectance, however, we have trouble with the meaning of antireflection and there are cases such as the rear surface of an absorbing substrate where an antireflection coating would be relevant. We do need to deal with this problem and although we have not yet discussed antireflection coatings it is most convenient to include the discussion here where we already have the basis for the theory. The discussion was originally given by Macleod [5].

The usual purpose of an antireflection coating is the reduction of reflectance, but frequently the objective of the reflectance reduction is the corresponding increase in transmittance. Although an absorbing or amplifying medium will rarely present us with a problem in terms of a reflectance measurement, we must occasionally treat a slab of such material on both sides to increase overall transmittance. In this context, therefore, we define an antireflection coating as one that increases transmittance and in the ideal case maximises it. But to accomplish this we need to define what we mean by transmittance.

We have no problem with the measurement of irradiance at the emergent side of our system, even if the emergent medium is absorbing. The incident irradiance is more difficult. This we can define as the irradiance if the transmitting structure were removed and replaced by an infinite extent of incident medium material. Then the transmittance will simply be the ratio of these two values, i.e.

$$I_{\text{inc}} = \frac{1}{2} \operatorname{Re}(y_0) \mathcal{E}_i \mathcal{E}_i^*,$$

and then

$$T = \frac{\frac{1}{2} \operatorname{Re}(y_1) \mathcal{E}_t \mathcal{E}_t^*}{\frac{1}{2} \operatorname{Re}(y_0) \mathcal{E}_i \mathcal{E}_i^*}.$$

This is completely consistent with (2.73), that is, with a slight manipulation,

$$T = 1 - \rho\rho^* + \frac{\operatorname{Re}[y_0^*(\rho - \rho^*)]}{\operatorname{Re}(y_0)}. \quad (2.76)$$

An alternative form uses

$$\mathcal{E}_t = \frac{2y_0}{(y_0 + y_1)} \mathcal{E}_i$$

so that

$$T = \frac{4y_0y_0^*\operatorname{Re}(y_1)}{\operatorname{Re}(y_0) \cdot [(y_0 + y_1)(y_0 + y_1)^*]}. \quad (2.77)$$

Now let the surface be coated with a dielectric system so that it presents the admittance Y . Then, since, in the absence of absorption, the net irradiance entering the thin-film system must also be the emergent irradiance,

$$T = \frac{4y_0y_0^*\operatorname{Re}(Y)}{\operatorname{Re}(y_0) \cdot [(y_0 + Y)(y_0 + Y)^*]}. \quad (2.78)$$

Let $Y = \alpha + i\beta$ then

$$T = \frac{4\alpha(n_0^2 + k_0^2)}{n_0[(n_0 + \alpha)^2 + (k_0 - \beta)^2]}$$

and T can readily be shown to be a maximum when

$$Y = \alpha + i\beta = n_0 + ik_0 = (n_0 - ik_0)^*. \quad (2.79)$$

The matching admittance should therefore be the *complex conjugate* of the incident admittance. For this perfect matching the transmittance becomes

$$T = \left(1 + \frac{k_0^2}{n_0^2}\right)$$

and this is greater than unity. This is not a mistake but rather a consequence of the definition of transmittance. Irradiance falls by a factor of roughly $4\pi k_0$ in a distance of one wavelength, rather larger than any normal value of k_0^2/n_0^2 , so that the effect is quite small. It originates in a curious pattern in the otherwise exponentially falling irradiance. It is caused by the presence of the interface and is a cyclic fluctuation in the rate of irradiance reduction. Note that the transmittance is unity if the coating is designed to match $n_0 - ik_0$ rather than its complex conjugate.

A dielectric coating that transforms an admittance y_1 to an admittance of y_0^* will also, when reversed, exactly transform an admittance of y_0 to y_1^* . This is dealt with in more detail later when induced transmission filters are discussed. Thus, the optimum coating to give highest transmittance will be the same in both

directions. This implies that an absorbing substrate in identical dielectric incident and emergent media should have exactly similar antireflection coatings on both front and rear surfaces.

Although also a little premature, it is convenient to mention here that the calculation of the properties of a coated slice of material involves multiple beams that are combined either coherently or incoherently. The coherent case is simply the usual interference calculation and we will return to that when we deal with induced transmission filters. We will see then that as the absorbing film becomes thicker, the matching rules for an induced transmission filter tend to (2.79). The incoherent case is at first sight less obvious. An estimate of the reflected beam is necessary for a multiple beam calculation. Such calculations imply that the absorption is not sufficiently high to eliminate completely a beam that suffers two traversals of the system. This implies, in turn, a negligible absorption in the space of one wavelength, in other words $4\pi k_0$ is very small. The upper limit on the size of the effect under discussion is k_0^2/n_0^2 and this will be still less significant. For an incoherent calculation to be appropriate there must be a jumbling of phase that washes out its effect. We can suppose for this discussion that the jumbling comes from a variation in the position of the reflecting surface over the aperture. The variation of the extra term in (2.79) is locked for its phase to the reflecting surface and so at any exactly plane surface that may be chosen as a reference, an average of the extra term is appropriate and this will be zero because ρ will have a phase that varies throughout the four quadrants. For multiple beam calculations, therefore, the reflectance can be taken simply as $\rho\rho^*$. Where k_0^2/n_0^2 is significant, the absorption will be very high and certainly enough for the influence of the multiple beams to be automatically negligible.

2.2.5 Oblique incidence in absorbing media

Remembering what we said in the previous section, we limit this to a transparent incident medium and an absorbing second, or emergent, medium. Our first aim must be to ensure that the phase factors are consistent. Taking advantage of some of the earlier results, we can write the phase factors as:

$$\begin{aligned} \text{incident: } & \exp\{i[\omega t - (2\pi n_0/\lambda)(x \sin \vartheta_0 + z \cos \vartheta_0)]\} \\ \text{reflected: } & \exp\{i[\omega t - (2\pi n_0/\lambda)(x \sin \vartheta_0 - z \cos \vartheta_0)]\} \\ \text{transmitted: } & \exp\{i[\omega t - (2\pi\{n_1 - ik_1\}/\lambda)(\alpha x + \gamma z)]\}, \end{aligned}$$

where α and γ in the transmitted phase factors are the only unknowns. The phase factors must be identically equal for all x and t with $z = 0$. This implies

$$\alpha = \frac{n_0 \sin \vartheta_0}{(n_1 - ik_1)}$$

and, since $\alpha^2 + \gamma^2 = 1$

$$\gamma = (1 - \alpha^2)^{1/2}.$$

There are two solutions to this equation and we must decide which is to be adopted. We note that it is strictly $(n_1 - ik_1)\alpha$ and $(n_1 - ik_1)\gamma$ that are required:

$$\begin{aligned}(n_1 - ik_1)\gamma &= [(n_1 - ik_1)^2 - n_0^2 \sin^2 \vartheta_0]^{1/2} \\ &= [n_1^2 - k_1^2 - n_0^2 \sin^2 \vartheta_0 - i2n_1k_1]^{1/2}.\end{aligned}$$

The quantity within the square root is in either the third or fourth quadrant and so the square roots are in the second quadrant (of the form $-a + ib$) and in the fourth quadrant (of the form $a - ib$). If we consider what happens when these values are substituted into the phase factors, we see that the fourth quadrant solution must be correct because this leads to an exponential fall-off with z amplitude, together with a change in phase of the correct sense. The second quadrant solution would lead to an increase with z and a change in phase of the incorrect sense, which would imply a wave travelling in the opposite direction. The fourth quadrant solution is also consistent with the solution for the absorption-free case. The transmitted phase factor is therefore of the form

$$\begin{aligned}\exp\{i[\omega t - (2\pi n_0 \sin \vartheta_0 x/\lambda) - (2\pi/\lambda)(a - ib)z]\} \\ = \exp(-2\pi bz/\lambda) \exp\{i[\omega t - (2\pi n_0 \sin \vartheta_0 x/\lambda) - (2\pi az/\lambda)]\}\end{aligned}$$

where

$$(a - ib) = [n_1^2 - k_1^2 - n_0^2 \sin^2 \vartheta_0 - i2n_1k_1]^{1/2}.$$

A wave which possesses such a phase factor is known as inhomogeneous. The exponential fall-off in amplitude is along the z axis, while the propagation direction in terms of phase is determined by the direction cosines, which can be extracted from

$$(2\pi n_0 \sin \vartheta_0 x/\lambda) + (2\pi az/\lambda).$$

The existence of such waves is another good reason for our choosing to consider the components of the fields parallel to the boundary and the flow of energy normal to the boundary.

We should note at this stage that provided we include the possibility of complex angles, the formulation of the absorption-free case applies equally well to absorbing media and we can write

$$\begin{aligned}(n_1 - ik_1) \sin \vartheta_1 &= n_0 \sin \vartheta_0 \\ \alpha &= \sin \vartheta_1 \\ \gamma &= \cos \vartheta_1 \\ (a - ib) &= (n_1 - ik_1) \cos \vartheta_1.\end{aligned}$$

The calculation of amplitudes follows the same pattern as before. However, we have not previously examined the implications of an inhomogeneous wave. Our

main concern is the calculation of the tilted admittance connected with such a wave. Since the x , y and t variations of the wave are contained in the phase factor, we can write

$$\begin{aligned}\operatorname{curl} &\equiv \left(\frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial x} \mathbf{j} + \frac{\partial}{\partial x} \mathbf{k} \right) \times \\ &\equiv \left(-i \frac{2\pi N}{\lambda} \alpha \mathbf{i} - i \frac{2\pi N}{\lambda} \gamma \mathbf{k} \right) \times\end{aligned}$$

and

$$\frac{\partial}{\partial t} \equiv i\omega,$$

where the \mathbf{k} is a unit vector in the z direction and should not be confused with the extinction coefficient k .

For p-waves the \mathbf{H} vector is parallel to the boundary in the y direction and so $\mathbf{H} = H_y \mathbf{j}$. The component of \mathbf{E} parallel to the boundary will then be in the x direction, $E_x \mathbf{i}$. We follow the analysis leading up to equation (2.23) and as before

$$\begin{aligned}\operatorname{curl} \mathbf{H} &= \sigma \mathbf{E} + \varepsilon \frac{\partial \mathbf{E}}{\partial t} \\ &= (\sigma + i\omega\varepsilon) \mathbf{E} \\ &= \frac{i\omega N^2}{c^2 \mu} \mathbf{E}.\end{aligned}$$

Now the tangential component of $\operatorname{curl} \mathbf{H}$ is in the x direction so that

$$-i \frac{2\pi N}{\lambda} \gamma (\mathbf{k} \times \mathbf{j}) H_y = i \frac{\omega N^2}{c^2 \mu} E_x i.$$

But

$$-(\mathbf{k} \times \mathbf{j}) = \mathbf{i}$$

so that

$$\begin{aligned}\eta_p &= \frac{H_y}{E_x} = \frac{\omega N \lambda}{2\pi c^2 \mu \gamma} = \frac{N}{c \mu \gamma} \\ &= \frac{N \gamma}{\gamma} = \frac{y}{\gamma}.\end{aligned}$$

For the s-waves we use

$$\operatorname{curl} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = -\mu \frac{\partial \mathbf{H}}{\partial t}.$$

E is now along the y axis and a similar analysis to that for p-waves yields

$$\eta_s = \frac{H_x}{E_y} = N\mathcal{Y}\gamma = y\gamma.$$

Now γ can be identified as $\cos \vartheta$, provided that ϑ is permitted to be complex, and so

$$\begin{aligned}\eta_p &= y/\cos \vartheta \\ (2.80) \end{aligned}$$

$$\eta_s = y \cos \vartheta.$$

Thus the amplitude and irradiance coefficients become as before

$$\rho = \frac{\eta_0 - \eta_1}{\eta_0 + \eta_1} \quad (2.81)$$

$$\tau = \frac{2\eta_0}{\eta_0 + \eta_1} \quad (2.82)$$

$$R = \left(\frac{\eta_0 - \eta_1}{\eta_0 + \eta_1} \right) \left(\frac{\eta_0 - \eta_1}{\eta_0 + \eta_1} \right)^* \quad (2.83)$$

$$T = \frac{4\eta_0 \operatorname{Re}(\eta_1)}{(\eta_0 + \eta_1)(\eta_0 + \eta_1)^*}. \quad (2.84)$$

These expressions are valid for absorption-free media as well.

2.3 The reflectance of a thin film

A simple extension of the above analysis occurs in the case of a thin, plane, parallel film of material covering the surface of a substrate. The presence of two (or more) interfaces means that a number of beams will be produced by successive reflections and the properties of the film will be determined by the summation of these beams. We say that the film is thin when interference effects can be detected in the reflected or transmitted light, that is, when the path difference between the beams is less than the coherence length of the light, and thick when the path difference is greater than the coherence length. The same film can appear thin or thick depending entirely on the illumination conditions. The thick case can be shown to be identical with the thin case integrated over a sufficiently wide wavelength range or a sufficiently large range of angles of incidence. Normally, we will find that the films on the substrates can be treated as thin while the substrates supporting the films can be considered thick. Thick films and substrates will be considered towards the end of this chapter. Here we concentrate on the thin case.

The arrangement is illustrated in figure 2.6. At this stage it is convenient to introduce a new notation. We denote waves in the direction of incidence by the

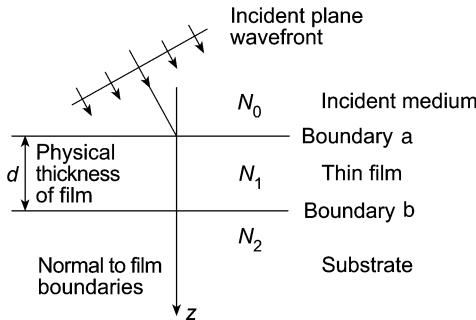


Figure 2.6. Plane wave incident on a thin film.

symbol + (that is, positive-going) and waves in the opposite direction by – (that is, negative-going).

The interface between the film and the substrate, denoted by the symbol b, can be treated in exactly the same way as the simple boundary already discussed. We consider the tangential components of the fields. There is no negative-going wave in the substrate and the waves in the film can be summed into one resultant positive-going wave and one resultant negative-going wave. At this interface, then, the tangential components of \mathbf{E} and \mathbf{H} are

$$\begin{aligned} \mathbf{E}_b &= E_{1b}^+ + E_{1b}^- \\ \mathbf{H}_b &= \eta_1 E_{1b}^+ - \eta_1 E_{1b}^-, \end{aligned}$$

where we are neglecting the common phase factors and where E_b and H_b represent the resultants. Hence

$$E_{1b}^+ = \frac{1}{2}(H_b/\eta_1 + E_b) \quad (2.85)$$

$$E_{1b}^- = \frac{1}{2}(-H_b/\eta_1 + E_b) \quad (2.86)$$

$$H_{1b}^+ = \eta_1 E_{1b}^+ = \frac{1}{2}(H_b + \eta_1 E_b) \quad (2.87)$$

$$H_{1b}^- = -\eta_1 E_{1b}^- = \frac{1}{2}(H_b - \eta_1 E_b). \quad (2.88)$$

The fields at the other interface a at the same instant and at a point with identical x and y coordinates can be determined by altering the phase factors of the waves to allow for a shift in the z coordinate from 0 to $-d$. The phase factor of the positive-going wave will be multiplied by $\exp(i\delta)$ where

$$\delta = 2\pi N_1 d \cos \vartheta_1 / \lambda$$

and ϑ_1 may be complex, while the negative-going phase factor will be multiplied by $\exp(-i\delta)$. We imply that this is a valid procedure when we say that the film

is thin. The values of \mathbf{E} and \mathbf{H} at the interface are now, using equations (2.85) to (2.88),

$$\begin{aligned} E_{1a}^+ &= E_{1b}^+ e^{i\delta} = \frac{1}{2}(H_b/\eta_1 + E_b)e^{i\delta} \\ E_{1a}^- &= E_{1b}^- e^{-i\delta} = \frac{1}{2}(-H_b/\eta_1 + E_b)e^{-i\delta} \\ H_{1a}^+ &= H_{1b}^+ e^{i\delta} = \frac{1}{2}(H_b + \eta_1 E_b)e^{i\delta} \\ H_{1a}^- &= H_{1b}^- e^{-i\delta} = \frac{1}{2}(H_b - \eta_1 E_b)e^{-i\delta} \end{aligned}$$

so that

$$\begin{aligned} E_a &= E_{1a}^+ + E_{1a}^- \\ &= E_b \left(\frac{e^{i\delta} + e^{-i\delta}}{2} \right) + H_b \left(\frac{e^{i\delta} - e^{-i\delta}}{2\eta_1} \right) \\ &= E_b \cos \delta + H_b \frac{i \sin \delta}{\eta_1} \\ H_a &= H_{1a}^+ + H_{1a}^- \\ &= E_b \eta_1 \left(\frac{e^{i\delta} - e^{-i\delta}}{2} \right) + H_b \left(\frac{e^{i\delta} + e^{-i\delta}}{2} \right) \\ &= E_b i \eta_1 \sin \delta + H_b \cos \delta. \end{aligned}$$

This can be written in matrix notation as

$$\begin{bmatrix} E_a \\ H_a \end{bmatrix} = \begin{bmatrix} \cos \delta & (i \sin \delta)/\eta_1 \\ i \eta_1 \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} E_b \\ H_b \end{bmatrix}. \quad (2.89)$$

Since the tangential components of \mathbf{E} and \mathbf{H} are continuous across a boundary, and since there is only a positive-going wave in the substrate, this relationship connects the tangential components of \mathbf{E} and \mathbf{H} at the incident interface with the tangential components of \mathbf{E} and \mathbf{H} which are transmitted through the final interface. The 2×2 matrix on the right-hand side of equation (2.89) is known as the characteristic matrix of the thin film.

We define the input optical admittance of the assembly by analogy with equation (2.64) as

$$Y = H_a/E_a \quad (2.90)$$

when the problem becomes merely that of finding the reflectance of a simple interface between an incident medium of admittance η_0 and a medium of admittance Y , i.e.

$$\rho = \frac{\eta_0 - Y}{\eta_0 + Y}$$

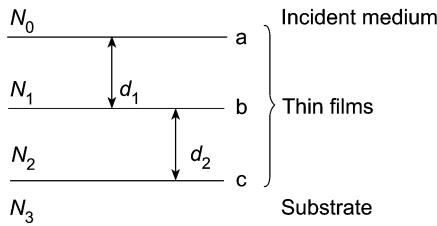


Figure 2.7. Notation for two films on a surface.

$$(2.91) \quad R = \left(\frac{\eta_0 - Y}{\eta_0 + Y} \right) \left(\frac{\eta_0 - Y}{\eta_0 + Y} \right)^*. \quad (2.91)$$

We can normalise equation (2.89) by dividing through by E_b to give

$$\begin{bmatrix} E_a/E_b \\ H_a/E_b \end{bmatrix} = \begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta & (i \sin \delta)/\eta_1 \\ i\eta_1 \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} 1 \\ \eta_2 \end{bmatrix} \quad (2.92)$$

and B and C , the normalised electric and magnetic fields at the front interface, are the quantities from which we will be extracting the properties of the thin-film system. Clearly, from (2.90) and (2.92), we can write

$$Y = \frac{H_a}{E_a} = \frac{C}{B} = \frac{\eta_2 \cos \delta + i\eta_1 \sin \delta}{\cos \delta + i(\eta_2/\eta_1) \sin \delta} \quad (2.93)$$

and from (2.93) and (2.91) we can calculate the reflectance.

$$\begin{bmatrix} B \\ C \end{bmatrix}$$

is known as the characteristic matrix of the assembly.

2.4 The reflectance of an assembly of thin films

Let another film be added to the single film of the previous section so that the final interface is now denoted by c , as shown in figure 2.7. The characteristic matrix of the film nearest the substrate is

$$\begin{bmatrix} \cos \delta_2 & (i \sin \delta_2)/\eta_2 \\ i\eta_2 \sin \delta_2 & \cos \delta_2 \end{bmatrix} \quad (2.94)$$

and from equation (2.89)

$$\begin{bmatrix} E_b \\ H_b \end{bmatrix} = \begin{bmatrix} \cos \delta_2 & (i \sin \delta_2)/\eta_2 \\ i\eta_2 \sin \delta_2 & \cos \delta_2 \end{bmatrix} \begin{bmatrix} E_c \\ H_c \end{bmatrix}. \quad (2.95)$$

We can apply equation (2.89) again to give the parameters at interface a, i.e.

$$\begin{bmatrix} E_a \\ H_a \end{bmatrix} = \begin{bmatrix} \cos \delta_1 & (i \sin \delta_1)/\eta_1 \\ i\eta_1 \sin \delta_1 & \cos \delta_1 \end{bmatrix} \begin{bmatrix} \cos \delta_2 & (i \sin \delta_2)/\eta_2 \\ i\eta_2 \sin \delta_2 & \cos \delta_2 \end{bmatrix} \begin{bmatrix} E_c \\ H_c \end{bmatrix}$$

and the characteristic matrix of the assembly, by analogy with equation (2.92) is,

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_1 & (i \sin \delta_1)/\eta_1 \\ i\eta_1 \sin \delta_1 & \cos \delta_1 \end{bmatrix} \begin{bmatrix} \cos \delta_2 & (i \sin \delta_2)/\eta_2 \\ i\eta_2 \sin \delta_2 & \cos \delta_2 \end{bmatrix} \begin{bmatrix} 1 \\ \eta_3 \end{bmatrix}.$$

Y is, as before, C/B , and the amplitude reflection coefficient and the reflectance are, from (2.91),

$$\begin{aligned} \rho &= \frac{\eta_0 - Y}{\eta_0 + Y} \\ R &= \left(\frac{\eta_0 - Y}{\eta_0 + Y} \right) \left(\frac{\eta_0 - Y}{\eta_0 + Y} \right)^*. \end{aligned} \quad (2.95)$$

This result can be immediately extended to the general case of an assembly of q layers, when the characteristic matrix is simply the product of the individual matrices taken in the correct order, i.e.

$$\begin{bmatrix} B \\ C \end{bmatrix} = \left\{ \prod_{r=1}^q \begin{bmatrix} \cos \delta_r & (i \sin \delta_r)/\eta_r \\ i\eta_r \sin \delta_r & \cos \delta_r \end{bmatrix} \right\} \begin{bmatrix} 1 \\ \eta_m \end{bmatrix}, \quad (2.96)$$

where

$$\delta_r = \frac{2\pi N_r d_r \cos \vartheta_r}{\lambda}$$

$$\begin{aligned} \eta_r &= \mathcal{Y} N_r \cos \vartheta_r && \text{for s-polarisation (TE)} \\ \eta_r &= \mathcal{Y} N_r / \cos \vartheta_r && \text{for p-polarisation (TM)} \end{aligned}$$

and where we have now used the suffix m to denote the substrate or emergent medium

$$\begin{aligned} \eta_m &= \mathcal{Y} N_m \cos \vartheta_m && \text{for s-polarisation (TE)} \\ \eta_m &= \mathcal{Y} N_m / \cos \vartheta_m && \text{for p-polarisation (TM).} \end{aligned}$$

If ϑ_0 , the angle of incidence, is given, the values of ϑ_r can be found from Snell's law, i.e.

$$N_0 \sin \vartheta_0 = N_r \sin \vartheta_r = N_m \sin \vartheta_m. \quad (2.97)$$

The expression (2.96) is of prime importance in optical thin-film work and forms the basis of almost all calculations.

A useful property of the characteristic matrix of a thin film is that the determinant is unity. This means that the determinant of the product of any number of these matrices is also unity.

It avoids difficulties over signs and quadrants if, in the case of absorbing media, the arrangement used for computing phase thicknesses and admittances is:

$$\delta_r = (2\pi/\lambda)d_r(n_r^2 - k_r^2 - n_0^2 \sin^2 \vartheta_0 - 2in_rk_r)^{1/2}, \quad (2.98)$$

the correct solution being in the fourth quadrant. Then

$$\eta_{rs} = \mathcal{Y}(n_r^2 - k_r^2 - n_0^2 \sin^2 \vartheta_0 - 2in_rk_r)^{1/2} \quad (2.99)$$

also in the fourth quadrant, and

$$\eta_{rp} = \frac{y_r^2}{\eta_{rs}} = \frac{\mathcal{Y}^2(n_r - ik_r)^2}{\eta_{rs}}. \quad (2.100)$$

It is useful to examine the phase shift associated with the reflected beam. Let $Y = a + ib$. Then with η_0 real

$$\begin{aligned} \rho &= \frac{\eta_0 - a - ib}{\eta_0 + a + ib} \\ &= \frac{(\eta_0^2 - a^2 - b^2) - i(2b\eta_0)}{(\eta_0 + a)^2 + b^2}, \end{aligned}$$

i.e.

$$\tan \varphi = \frac{(-2b\eta_0)}{(\eta_0^2 - a^2 - b^2)}, \quad (2.101)$$

where φ is the phase shift. This must be interpreted, of course, on the basis of the sign convention we have already established in figure 2.3. It is important to preserve the signs of the numerator and denominator separately as shown, otherwise the quadrant cannot be uniquely specified. The rule is simple. It is the quadrant in which the vector associated with ρ lies and the following scheme can be derived by treating the denominator as the x coordinate and the numerator as the y coordinate.

Numerator	+	+	-	-
Denominator	+	-	+	-
Quadrant	1 st	2 nd	4 th	3 rd

Note that the reference surface for the calculation of phase shift on reflection is the front surface of the multilayer.

2.5 Reflectance, transmittance and absorptance

Sufficient information is included in equation (2.96) to allow the transmittance and absorptance of a thin-film assembly to be calculated. For this to have a physical meaning, as we have already seen, the incident medium should be transparent, that is, η_0 must be real. The substrate need not be transparent, but the transmittance calculated will be the transmittance into, rather than through, the substrate.

First of all, we calculate the net irradiance at the exit side of the assembly, which we take as the k th interface. This is given by

$$I_k = \frac{1}{2} \operatorname{Re}(E_k H_k^*),$$

where, once again, we are dealing with the component of irradiance normal to the interfaces.

$$\begin{aligned} I_k &= \frac{1}{2} \operatorname{Re}(E_k \eta_m^* E_k^*) \\ &= \frac{1}{2} \operatorname{Re}(\eta_m^*) E_k E_k^*. \end{aligned} \tag{2.102}$$

If the characteristic matrix of the assembly is

$$\begin{bmatrix} B \\ C \end{bmatrix}$$

then the net irradiance at the entrance to the assembly is

$$I_a = \frac{1}{2} \operatorname{Re}(BC^*) E_k E_k^*. \tag{2.103}$$

Let the incident irradiance be denoted by I_i , then equation (2.103) represents the irradiance actually entering the assembly, which is $(1 - R)I_i$:

$$(1 - R)I_i = \frac{1}{2} \operatorname{Re}(BC^*) E_k E_k^*,$$

i.e.

$$I_i = \frac{\operatorname{Re}(BC^*) E_k E_k^*}{2(1 - R)}.$$

Equation (2.102) represents the irradiance leaving the assembly and entering the substrate and so the transmittance T is

$$T = \frac{I_k}{I_i} = \frac{\operatorname{Re}(\eta_m)(1 - R)}{\operatorname{Re}(BC^*)}. \tag{2.104}$$

The absorptance A in the multilayer is connected with R and T by the relationship

$$1 = R + T + A$$

so that

$$A = 1 - R - T = (1 - R) \left(1 - \frac{\text{Re}(\eta_m)}{\text{Re}(BC^*)} \right). \quad (2.105)$$

In the absence of absorption in any of the layers it can readily be shown that the above expressions are consistent with $A = 0$ and $T + R = 1$, for the individual film matrices will have determinants of unity and the product of any number of these matrices will also have a determinant of unity. The product of the matrices can be expressed as

$$\begin{bmatrix} \alpha & i\beta \\ i\gamma & \delta \end{bmatrix}$$

where $\alpha\delta + \gamma\beta = 1$ and, because there is no absorption, α, β, γ and δ are all real.

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \alpha & i\beta \\ i\gamma & \delta \end{bmatrix} \begin{bmatrix} 1 \\ \eta_m \end{bmatrix} = \begin{bmatrix} \alpha + i\beta\eta_m \\ \delta\eta_m + i\gamma \end{bmatrix}$$

$$\begin{aligned} \text{Re}(BC^*) &= \text{Re}[(\alpha + i\beta\eta_m)(\delta\eta_m - i\gamma)] = (\alpha\delta + \gamma\beta)\text{Re}(\eta_m) \\ &= \text{Re}(\eta_m) \end{aligned}$$

and the result follows.

We can manipulate equations (2.104) and (2.105) into slightly better forms. From equation (2.91)

$$R = \left(\frac{\eta_0 B - C}{\eta_0 B + C} \right) \left(\frac{\eta_0 B - C}{\eta_0 B + C} \right)^* \quad (2.106)$$

so that

$$(1 - R) = \frac{2\eta_0(BC^* + B^*C)}{(\eta_0 B + C)(\eta_0 B + C)^*}. \quad (2.107)$$

Inserting this result into equation (2.104) we obtain

$$T = \frac{4\eta_0 \text{Re}(\eta_m)}{(\eta_0 B + C)(\eta_0 B + C)^*} \quad (2.108)$$

and in equation (2.105)

$$A = \frac{4\eta_0 \text{Re}(BC^* - \eta_m)}{(\eta_0 B + C)(\eta_0 B + C)^*}. \quad (2.109)$$

Equations (2.106), (2.108) and (2.109) are the most useful forms of the expressions for R , T and A .

An important quantity which we shall discuss in a later section of this chapter is $T/(1 - R)$, known as the potential transmittance ψ . From equation (2.104)

$$\psi = \frac{T}{(1 - R)} = \frac{\operatorname{Re}(\eta_m)}{\operatorname{Re}(BC^*)}. \quad (2.110)$$

The phase change on reflection (equation (2.101)) can also be put in a form compatible with equations (2.106) to (2.109).

$$\varphi = \arctan\left(\frac{\operatorname{Im}[\eta_m(BC^* - CB^*)]}{(\eta_m^2 BB^* - CC^*)}\right). \quad (2.111)$$

The quadrant of φ is given by the same scheme of signs of numerator and denominator as equation (2.101). The phase change on reflection is measured at the front surface of the multilayer.

Phase shift on transmission is sometimes important. This can be obtained in a way similar to the phase shift on reflection. We denote the phase shift by ζ and we define it as the difference in phase between the resultant transmitted wave as it enters the emergent medium and the incident wave exactly at the front surface, that is as it enters the multilayer. The electric field amplitude at the emergent surface has been normalised to unity and so the phase may be taken as zero. Then we simply have to find the expression, which will involve B and C , for the incident amplitude. These are the normalised total tangential electric and magnetic fields. So we can write

$$\begin{aligned} E_i + E_r &= B \\ \eta_0 E_i - \eta_0 E_r &= C. \end{aligned}$$

Then we eliminate E_r to give

$$E_i = \frac{1}{2}\left(B + \frac{C}{\eta_0}\right)$$

and the amplitude transmission coefficient as

$$\tau = \frac{2\eta_0}{(\eta_0 B + C)} = \frac{2\eta_0(\eta_0 B + C)^*}{(\eta_0 B + C)(\eta_0 B + C)^*}$$

so that

$$\zeta = \arctan\left[\frac{-\operatorname{Im}(\eta_0 B + C)}{\operatorname{Re}(\eta_0 B + C)}\right]. \quad (2.112)$$

Again it is important to keep the signs of the numerator and the denominator separate. The quadrant is then given by the same arrangement of signs as equation (2.101).

2.6 Units

We have been using the International System of Units (SI) in the work so far. In this system y , η and Y are measured in siemens. Much thin-film literature has been written in Gaussian units. In Gaussian units, \mathcal{Y} , the admittance of free space, is unity and so, since $y = NY$, y (the optical admittance) and N (the refractive index) are numerically equal at normal incidence, although N is a number without units. The position is different in SI units, where \mathcal{Y} is 2.6544×10^{-3} S. We could, if we choose, measure y and η in units of \mathcal{Y} siemens, which we can call free space units, and in this case y becomes numerically equal to N , just as in the Gaussian system. This is a perfectly valid procedure, and all the expressions for ratioed quantities, notably reflectance, transmittance, absorptance and potential transmittance, are unchanged. This applies particularly to equations (2.96) and (2.106)–(2.110). We must simply take due care when calculating absolute rather than relative irradiance and also when deriving the magnetic field. In particular, equation (2.89) becomes

$$\begin{bmatrix} E_a \\ H_a/\mathcal{Y} \end{bmatrix} = \begin{bmatrix} \cos \delta & (i \sin \delta)/\eta_1 \\ i\eta_1 \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} E_b \\ H_b/\mathcal{Y} \end{bmatrix}, \quad (2.113)$$

where η is now measured in free space units. In most cases in this book, either arrangement can be used. In some cases, particularly where we are using graphical techniques, we shall use free space units, because otherwise the scales become quite cumbersome.

2.7 Summary of important results

We have now covered all the basic theory necessary for the understanding of the remainder of the book. It has been a somewhat long and involved discussion and so we now summarise the principal results. The statement numbers refer to those in the text where the particular quantities were originally introduced.

Refractive index is defined as the ratio of the velocity of light in free space c to the velocity of light in the medium v . When the refractive index is real it is denoted by n but it is frequently complex and then is denoted by N .

$$N = c/v = n - ik. \quad (2.17)$$

N is often called the complex refractive index, n the real refractive index (or often simply refractive index) and k the extinction coefficient. N is always a function of λ .

k is related to the absorption coefficient α by

$$\alpha = 4\pi k/\lambda. \quad (2.33)$$

Light waves are electromagnetic and a homogeneous, plane, plane-polarised harmonic (or monochromatic) wave may be represented by expressions of the

form

$$\mathbf{E} = \mathcal{E} \exp[i\omega t - (2\pi N/\lambda)x + \varphi], \quad (2.20)$$

where x is the distance along the direction of propagation, \mathbf{E} is the electric field, \mathcal{E} the electric amplitude and φ an arbitrary phase. A similar expression holds for \mathbf{H} , the magnetic field:

$$\mathbf{H} = \mathcal{H} \exp[i\omega t - (2\pi N/\lambda)x + \varphi'], \quad (2.114)$$

where φ , φ' and N are not independent. The physical significance is attached to the real parts of the above expressions.

The phase change suffered by the wave on traversing a distance d of the medium is, therefore,

$$-\frac{2\pi Nd}{\lambda} = -\frac{2\pi nd}{\lambda} + i\frac{2\pi kd}{\lambda} \quad (2.115)$$

and the imaginary part can be interpreted as a reduction in amplitude (by substituting in equation (2.20)).

The optical admittance is defined as the ratio of the magnetic and electric fields

$$y = H/E \quad (2.23)-(2.28)$$

and y is usually complex. In free space, y is real and is denoted by \mathcal{Y} :

$$\mathcal{Y} = 2.6544 \times 10^{-3} \text{ S.} \quad (2.116)$$

The optical admittance of a medium is connected with the refractive index by

$$y = N\mathcal{Y}. \quad (2.117)$$

(In Gaussian units \mathcal{Y} is unity and y and N are numerically the same. In SI units we can make y and N numerically equal by expressing y in units of \mathcal{Y} , i.e. free space units. All expressions for reflectance, transmittance etc involving ratios will remain valid, but care must be taken when computing absolute irradiances, although these are not often needed in thin-film optics, except where damage studies are involved.)

The irradiance of the light, defined as the mean rate of flow of energy per unit area carried by the wave, is given by

$$I = \frac{1}{2}\text{Re}(EH^*). \quad (2.31)$$

This can also be written

$$I = \frac{1}{2}n\mathcal{Y}EE^*, \quad (2.118)$$

where $*$ denotes the complex conjugate.

At a boundary between two media, denoted by suffix 0 for the incident medium and by suffix 1 for the exit medium, the incident beam is split into a reflected beam and a transmitted beam. For normal incidence we have

$$\rho = \frac{\mathcal{E}_r}{\mathcal{E}_i} = \frac{y_0 - y_1}{y_0 + y_1} = \frac{(n_0 - ik_0) \mathcal{Y} - (n_1 - ik_1) \mathcal{Y}}{(n_0 - ik_0) \mathcal{Y} + (n_1 - ik_1) \mathcal{Y}}$$

$$= \frac{(n_0 - n_1) - i(k_0 - k_1)}{(n_0 + n_1) - i(k_0 + k_1)} \quad (2.70)$$

$$\tau = \frac{\mathcal{E}_t}{\mathcal{E}_i} = \frac{2y_0}{y_0 - y_1} = \frac{2(n_0 - ik_0) \mathcal{Y}}{(n_0 - ik_0) \mathcal{Y} + (n_1 - ik_1) \mathcal{Y}}$$

$$= \frac{2(n_0 - ik_0)}{(n_0 + n_1) - i(k_0 + k_1)}, \quad (2.71)$$

where ρ is the amplitude reflection coefficient and τ the amplitude transmission coefficient.

There are fundamental difficulties associated with the definitions of reflectance and transmittance unless the incident medium is absorption-free, i.e. N_0 and y_0 are real. For that case

$$R = \rho\rho^* = \left(\frac{y_0 - y_1}{y_0 + y_1} \right) \left(\frac{y_0 - y_1}{y_0 + y_1} \right)^* \quad (2.74)$$

$$T = \frac{4y_0 \operatorname{Re}(y_1)}{(y_0 + y_1)(y_0 + y_1)^*}. \quad (2.75)$$

Oblique incidence calculations are simpler if the wave is split into two plane-polarised components, one with the electric vector in the plane of incidence, known as p-polarised (or TM, for transverse magnetic field) and one with the electric vector normal to the plane of incidence, known as s-polarised (or TE, for transverse electric field). The propagation of each of these two waves can be treated quite independently of the other. Calculations are further simplified if only energy flows normal to the boundaries and electric and magnetic fields parallel to the boundaries are considered, because then we have a formulation which is equivalent to a homogeneous wave.

We must introduce the idea of a tilted optical admittance η , which is given by

$$\eta_p = \frac{N\mathcal{Y}}{\cos \vartheta} \quad (\text{for p-waves}) \quad (2.80)$$

$$\eta_s = N\mathcal{Y} \cos \vartheta \quad (\text{for s-waves}),$$

where N and ϑ denote either N_0 and ϑ_0 or N_1 and ϑ_1 as appropriate. ϑ_1 is given by Snell's law, in which complex angles may be included:

$$N_0 \sin \vartheta_0 = N_1 \sin \vartheta_1. \quad (2.119)$$

Denoting η_p or η_s by η we have, for either plane of polarisation,

$$\rho = \frac{\eta_0 - \eta_1}{\eta_0 + \eta_1} \quad (2.81)$$

$$\tau = \frac{2\eta_0}{\eta_0 + \eta_1}. \quad (2.82)$$

If η_0 is real, we can write

$$R = \left(\frac{\eta_0 - \eta_1}{\eta_0 + \eta_1} \right) \left(\frac{\eta_0 - \eta_1}{\eta_0 + \eta_1} \right)^* \quad (2.83)$$

$$T = \frac{4\eta_0 \operatorname{Re}(\eta_1)}{(\eta_0 + \eta_1)(\eta_0 + \eta_1)^*}. \quad (2.84)$$

The phase shift experienced by the wave as it traverses a distance d normal to the boundary is then given by $-2\pi Nd \cos \vartheta/\lambda$.

The reflectance of an assembly of thin films is calculated through the concept of optical admittance. We replace the multilayer by a single surface which presents an admittance Y , which is the ratio of the total tangential magnetic and electric fields and is given by

$$Y = C/B, \quad (2.120)$$

where

$$\begin{bmatrix} B \\ C \end{bmatrix} = \left\{ \prod_{r=1}^q \begin{bmatrix} \cos \delta_r & (\sin \delta_r)/\eta_r \\ i\eta_r \sin \delta_r & \cos \delta_r \end{bmatrix} \right\} \begin{bmatrix} 1 \\ \eta_m \end{bmatrix}, \quad (2.96)$$

$\delta_r = 2\pi Nd \cos \vartheta/\lambda$ and η_m = substrate admittance.

The order of multiplication is important. If q is the layer next to the substrate then the order is

$$\begin{bmatrix} B \\ C \end{bmatrix} = [M_1][M_2] \dots [M_q] \begin{bmatrix} 1 \\ \eta_m \end{bmatrix}. \quad (2.121)$$

M_1 indicates the matrix associated with layer 1, and so on. Y and η are in the same units. If η is in siemens then so also is Y , or if η is in free space units (i.e. units of \mathcal{Y}) then Y will be in free space units also. As in the case of a single surface, η_0 must be real for reflectance and transmittance to have a valid meaning. With that proviso, then

$$R = \left(\frac{\eta_0 B - C}{\eta_0 B + C} \right) \left(\frac{\eta_0 B - C}{\eta_0 B + C} \right)^* \quad (2.106)$$

$$T = \frac{4\eta_0 \operatorname{Re}(\eta_m)}{(\eta_0 B + C)(\eta_0 B + C)^*} \quad (2.108)$$

$$A = \frac{4\eta_0 \operatorname{Re}(BC^* - \eta_m)}{(\eta_0 B + C)(\eta_0 B + C)^*} \quad (2.109)$$

$$\psi = \text{potential transmittance} = \frac{T}{(1 - R)} = \frac{\operatorname{Re}(\eta_m)}{\operatorname{Re}(BC^*)}. \quad (2.110)$$

Phase shift on reflection, measured at the front surface of the multilayer, is given by

$$\varphi = \arctan \left(\frac{\operatorname{Im} [\eta_m (BC^* - CB^*)]}{(\eta_m^2 BB^* - CC^*)} \right) \quad (2.111)$$

and that on transmission, measured between the emergent wave as it leaves the multilayer and the incident wave as it enters, by

$$\zeta = \arctan \left[\frac{-\operatorname{Im} (\eta_0 B + C)}{\operatorname{Re} (\eta_0 B + C)} \right]. \quad (2.112)$$

The signs of the numerator and denominator in these expressions must be preserved separately. Then the quadrants are given by the arrangement in the table:

Numerator	+	+	-	-
Denominator	+	-	+	-
Quadrant	1 st	2 nd	4 th	3 rd

In spite of the apparent simplicity of expression (2.96), numerical calculations without some automatic aid are tedious in the extreme. Even with the help of a calculator, the labour involved in determining the performance of an assembly of more than a very few transparent layers at one or two wavelengths is completely discouraging. At the very least, a programmable calculator of reasonable capacity is required. Extended calculations are usually carried out on a computer.

However, insight into the properties of thin-film assemblies cannot easily be gained simply by feeding the calculations into a computer, and insight is necessary if filters are to be designed and if their limitations in use are to be fully understood. Studies have been made of the properties of the characteristic matrices and some results which are particularly helpful in this context have been obtained. Approximate methods, especially graphical ones, have also been found useful.

2.8 Potential transmittance

The potential transmittance of a layer or an assembly of layers is the ratio of the irradiance leaving by the rear, or exit, interface to that entering by the front interface. The concept was introduced by Berning and Turner [6] and we will make considerable use of it in designing metal-dielectric filters and in calculating losses in all-dielectric multilayers. Potential transmittance is denoted by ψ and is given by

$$\psi = \frac{I_{\text{exit}}}{I_{\text{enter}}} = \frac{T}{(1 - R)}, \quad (2.122)$$

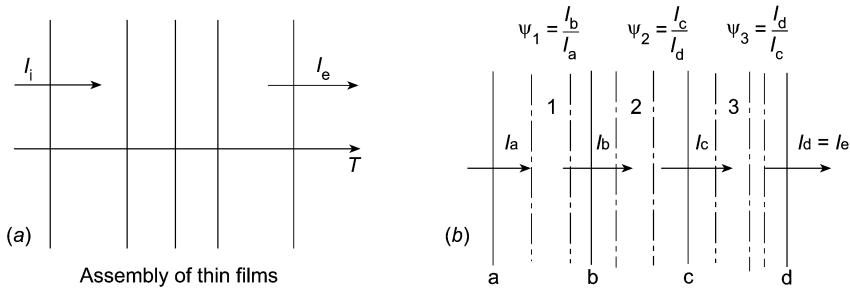


Figure 2.8. (a) An assembly of thin films. (b) The potential transmittance of an assembly of thin film consisting of a number of subunits.

that is the ratio between the irradiance leaving the assembly and the net irradiance actually entering. For the entire system, the net irradiance actually entering is the difference between the incident and reflected irradiances.

The potential transmittance of a series of subassemblies of layers is simply the product of the individual potential transmittances. Figure 2.8 shows a series of film subunits making up a complete system. Clearly

$$\psi = \frac{I_e}{I_i} = \frac{I_d}{I_a} = \frac{I_b}{I_a} \frac{I_c}{I_b} \frac{I_d}{I_c} = \psi_1 \psi_2 \psi_3. \quad (2.123)$$

The potential transmittance is fixed by the parameters of the layer, or combination of layers, involved, and by the characteristics of the structure at the exit interface, and it represents the transmittance which this particular combination would give if there were no reflection losses. Thus, it is a measure of the maximum transmittance which could be expected from the arrangement. By definition, the potential transmittance is unaffected by any transparent structure deposited over the front surface—which can affect the transmittance as distinct from the potential transmittance—and to ensure that the transmittance is equal to the potential transmittance the layers added to the front surface must maximise the irradiance actually entering the assembly. This implies reducing the reflectance of the complete assembly to zero or, in other words, adding an antireflection coating. The potential transmittance is, however, affected by any changes in the structure at the exit interface and it is possible to maximise the potential transmittance of a subassembly in this way.

We now show that the parameters of the layer, or subassembly of layers, together with the optical admittance at the rear surface, are sufficient to define the potential transmittance. Let the complete multilayer performance be given by

$$\begin{bmatrix} B \\ C \end{bmatrix} = [M_1][M_2] \dots [M_a][M_b][M_c] \dots [M_p][M_q] \begin{bmatrix} 1 \\ \eta_m \end{bmatrix},$$

where we want to calculate the potential transmittance of the subassembly $[M_a][M_b][M_c]$. Let the product of the matrices to the right of the subassembly

be given by

$$\begin{bmatrix} B_e \\ C_e \end{bmatrix}.$$

Now, if

$$\begin{bmatrix} B_i \\ C_i \end{bmatrix} = [M_a][M_b][M_c] \begin{bmatrix} B_e \\ C_e \end{bmatrix}, \quad (2.124)$$

then

$$\psi = \frac{\operatorname{Re}(B_e C_e^*)}{\operatorname{Re}(B_i C_i^*)}. \quad (2.125)$$

By dividing equation (2.124) by B_e we have

$$\begin{bmatrix} B'_i \\ C'_i \end{bmatrix} = [M_a][M_b][M_c] \begin{bmatrix} 1 \\ Y_e \end{bmatrix},$$

where $Y_e = C_e/B_e$, $B'_i = B_i/B_e$, $C'_i = C_i/C_e$ and the potential transmittance is

$$\begin{aligned} \psi &= \frac{\operatorname{Re}(Y_e)}{\operatorname{Re}(B'_i C'^*_i)} \\ &= \frac{\operatorname{Re}(C_e/B_e)}{\operatorname{Re}[(B_i/B_e)(C_i^*/B_e^*)]} = \frac{B_e B_e^* \operatorname{Re}(C_e/B_e)}{\operatorname{Re}(B_i C_i^*)} \\ &= \frac{\operatorname{Re}(B_e^* C_e)}{\operatorname{Re}(B_i C_i^*)} = \frac{\operatorname{Re}(B_e C_e^*)}{\operatorname{Re}(B_i C_i^*)}, \end{aligned}$$

which is identical to equation (2.125). Thus the potential transmittance of any subassembly is determined solely by the characteristics of the layer or layers of the subassembly together with the optical admittance of the structure at the exit interface.

Further expressions involving potential transmittance will be derived as they are required.

2.9 Quarter- and half-wave optical thicknesses

The characteristic matrix of a dielectric thin film takes on a very simple form if the optical thickness is an integral number of quarter- or half-waves. That is, if

$$\delta = m(\pi/4) \quad m = 0, 1, 2, 3 \dots$$

For m even, $\cos \delta = \pm 1$ and $\sin \delta = 0$, so that the layer is an integral number of half wavelengths thick, and the matrix becomes

$$\pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

This is the unity matrix and can have no effect on the reflectance or transmittance of an assembly. It is as if the layer were completely absent. This is a particularly useful result and, because of it, half-wave layers are sometimes referred to as absentee layers. In the computation of the properties of any assembly, layers which are an integral number of half wavelengths thick can be omitted completely without altering the result. Of course this is true only at the wavelength for which the layers are half-waves.

For m odd, $\sin \delta = \pm 1$ and $\cos \delta = 0$, so that the layer is an odd number of quarter wavelengths thick, and the matrix becomes

$$\pm \begin{bmatrix} 0 & i/\eta \\ i\eta & 0 \end{bmatrix}.$$

This is not quite as simple as the half-wave case, but such a matrix is still easy to handle in calculations. In particular, if a substrate or combination of thin films has an admittance of Y , then addition of an odd number of quarter-waves of admittance η alters the admittance of the assembly to η^2/Y . This makes the properties of a succession of quarter-wave layers very easy to calculate. The admittance of, say, a stack of five quarter-wave layers is

$$Y = \frac{\eta_1^2 \eta_3^2 \eta_5^2}{\eta_2^2 \eta_4^2 \eta_m},$$

where the symbols have their usual meanings.

Because of the simplicity of assemblies involving quarter- and half-wave optical thicknesses, designs are often specified in terms of fractions of quarter-waves at a reference wavelength. Usually only two, or perhaps three, different materials are involved in designs and a convenient shorthand notation for quarter-wave optical thicknesses is H , M or L where H refers to the highest of the three indices, M the intermediate and L the lowest. Half-waves are denoted by HH , MM , LL or $2H$, $2M$ and so on.

2.10 A theorem on the transmittance of a thin-film assembly

The transmittance of a thin-film assembly is independent of the direction of propagation of the light. This applies regardless of whether or not the layers are absorbing.

A proof of this result, due to Abelès [7, 8], who was responsible for the development of the matrix approach to the analysis of thin films, follows quickly from the properties of the matrices.

Let the matrices of the various layers in the assembly be denoted by

$$[M_1], [M_2], \dots [M_q]$$

and let the two massive media on either side be transparent. The two products of the matrices corresponding to the two possible directions of propagation can be written as

$$[M] = [M_1][M_2][M_3], \dots [M_q]$$

and

$$[M'] = [M_q][M_{q-1}] \dots [M_2][M_1].$$

Now, because the form of the matrices is such that the diagonal terms are equal, regardless of whether there is absorption or not, we can show that if we write

$$[M] = [a_{ij}] \quad \text{and} \quad [M'] = [a'_{ij}]$$

then

$$a_{ij} = a'_{ij} \quad (i \neq j), \quad a_{11} = a'_{22} \text{ and } a_{22} = a'_{11}.$$

This can be proved simply by induction.

We denote the medium on one side of the assembly by η_0 and on the other by η_m , where η_0 is next to layer 1. In the case of the first direction the characteristic matrix is given by (equation (2.96))

$$\begin{bmatrix} B \\ C \end{bmatrix} = [M] \begin{bmatrix} 1 \\ \eta_m \end{bmatrix}$$

and

$$B = a_{11} + a_{12}\eta_m \quad C = a_{21} + a_{22}\eta_m.$$

In the second case

$$B = a'_{11} + a'_{12}\eta_0 = a_{22} + a_{12}\eta_0$$

$$C = a'_{21} + a'_{22}\eta_0 = a_{21} + a_{11}\eta_0.$$

The two expressions for the transmittance of the assembly are then, from equation (2.108),

$$T = \frac{4\eta_0\eta_m}{|\eta_0(a_{11} + a_{12}\eta_m) + a_{21} + a_{22}\eta_m|^2}$$

$$T' = \frac{4\eta_m\eta_0}{|\eta_m(a_{22} + a_{12}\eta_0) + a_{21} + a_{11}\eta_0|^2}$$

which are identical.

This rule does not, of course, apply to the reflectance of an assembly, which will necessarily be the same on both sides of the assembly only if there is no absorption in any of the layers.

Amongst other things, this expression shows that the one-way mirror, which allows light to travel through it in one direction only, cannot be constructed from simple optical thin films. The common so-called one-way mirror has a high reflectance with some transmittance and relies for its operation on an appreciable difference in the illumination conditions existing on either side.

2.11 Admittance loci

This section is devoted to the admittance diagram. The admittance diagram in common with the Smith chart and the reflection circle diagram, described later, is a graphical technique based on an exact solution of the appropriate equations. We imagine that the multilayer is gradually built up on the substrate layer by layer, immersed all the time in the final incident medium. As each layer in turn increases from zero thickness to its final value, the admittance of the multilayer at that stage of its construction is calculated and the locus is plotted. Alternatively, we may imagine the multilayer as already constructed and then a reference plane is slid continuously through the layers and the locus of admittance of the structure up to that plane plotted. Either of these views is equally valid and the results are identical. (Note that only the first possibility applies to the reflection circle diagram and only the second to the Smith chart.) The loci for dielectric layers take the form of a series of circular arcs or even complete circles, each corresponding to a single layer, which are connected at points corresponding to the interfaces between the different layers. Perfect metals are also represented by arcs of circles. Absorbing materials give spiral loci. Although the technique can be used for quantitative calculation it cannot compete even with a small programmable calculator, and its great value is in the visualisation of the characteristics of a particular multilayer.

As the reference plane moves from the surface of the substrate to the front surface of the multilayer, let us calculate and plot the variation of the input optical admittance at the reference plane.

Equation (2.96) is

$$\begin{bmatrix} B \\ C \end{bmatrix} = \left\{ \prod_{r=1}^q \begin{bmatrix} \cos \delta_r & (i \sin \delta_r) / \eta_r \\ i \eta_r \sin \delta_r & \cos \delta_r \end{bmatrix} \right\} \begin{bmatrix} 1 \\ \eta_m \end{bmatrix},$$

where $Y = C/B$ is the input optical admittance of the assembly. For the r th layer we can write

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_r & (i \sin \delta_r) / \eta_r \\ i \eta_r \sin \delta_r & \cos \delta_r \end{bmatrix} \begin{bmatrix} B' \\ C' \end{bmatrix}$$

and since it is optical admittance we are interested in we can divide throughout by B' to give

$$\begin{bmatrix} B/B' \\ C/B' \end{bmatrix} = \begin{bmatrix} \cos \delta_r & (i \sin \delta_r) / \eta_r \\ i \eta_r \sin \delta_r & \cos \delta_r \end{bmatrix} \begin{bmatrix} 1 \\ Y' \end{bmatrix},$$

where $Y' = C/B'$ represents the admittance of the structure at the exit side of the layer. We now find the locus of the input admittance

$$Y = \frac{C}{B} = \frac{C/B'}{B/B'}.$$

Let

$$Y = x + iy$$

and

$$Y' = \alpha + i\beta$$

and let the layer in question be dielectric so that η_r and δ_r are both real. Then

$$\begin{aligned} Y = x + iy &= \frac{(\alpha + i\beta) \cos \delta_r + i\eta_r \sin \delta_r}{\cos \delta_r + (\alpha + i\beta)(i \sin \delta_r)/\eta_r} \\ &= \frac{\alpha \cos \delta_r + i(\beta \cos \delta_r + \eta_r \sin \delta_r)}{[\cos \delta_r - (\beta/\eta_r) \sin \delta_r] + i(\alpha/\eta_r) \sin \delta_r}. \end{aligned}$$

Equating real and imaginary parts

$$x[\cos \delta_r - (\beta/\eta_r) \sin \delta_r] - (y\alpha/\eta_r) \sin \delta_r = \alpha \cos \delta_r \quad (2.126)$$

$$y[\cos \delta_r - (\beta/\eta_r) \sin \delta_r] + (x\alpha/\eta_r) \sin \delta_r = \beta \cos \delta_r + \eta_r \sin \delta_r. \quad (2.127)$$

Eliminating δ_r yields

$$x^2 + y^2 - x[(\alpha^2 + \beta^2 + \eta_r^2)/\alpha] + \eta_r^2 = 0 \quad (2.128)$$

which is the equation of a circle with centre $((\alpha^2 + \beta^2 + \eta_r^2)/2\alpha, 0)$, i.e. on the real axis and with radius such that it passes through the point (α, β) , i.e. its starting point. The circle is traced out in a clockwise direction, which can be shown by setting $\beta = 0$ in equation (2.127).

We can plot the locus in the complex plane in the same way as the locus of the amplitude reflection coefficient (section 2.15.5).

The scale of δ_r can also be plotted on the diagram. Let $\beta = 0$ and then, from equations (2.126) and (2.127),

$$\begin{aligned} x - (y\alpha/\eta_r) \tan \delta_r &= \alpha \\ y + (x\alpha/\eta_r) \tan \delta_r &= \eta_r \tan \delta_r. \end{aligned}$$

Eliminating α , we have

$$x^2 + y^2 - y(\tan \delta_r - 1/\tan \delta_r) - \eta_r^2 = 0. \quad (2.129)$$

This is a circle with centre

$$(0, (\eta_r/2)(\tan \delta_r - 1/\tan \delta_r)),$$

i.e. on the imaginary axis and passing through the point $(\eta_r, 0)$. The simplest contours of equal δ_r are $\delta_r = 0, \pi/2, \pi, 3\pi/2, \dots$, which coincide with the real axis, and $\delta_r = \pi/4, 3\pi/4, 5\pi/4, \dots$, which is the circle with centre the origin

and which passes through the point $(\eta_r, 0)$. For layers which start at a point not on the real axis, the same set of contours of equal δ_r will still apply, with a correction to the value of δ_r that each represents.

Figure 2.9(a) shows the locus of a film that is deposited on a transparent substrate of admittance α . The starting point is $(\alpha, 0)$ and, as the thickness is increased to a quarter-wave, a semicircle is traced out clockwise which reintersects the real axis at the point $(\eta_r^2/\alpha, 0)$. A second quarter-wave completes the circle. We could have had any point on the locus as a starting point without changing its form. The only difference would have been an offset in the scale of δ_r .

We could add isoreflectance contours to the diagram if we wished. These are circles with centres on the real axis, centres and radii being given by

$$(\eta_0(1+R)/(1-R), 0) \quad \text{and} \quad 2\eta_0(R)^{1/2}/(1-R), \quad (2.130)$$

respectively, where η_0 is the admittance of the incident medium.

The phase of the reflectance can also be important and isophase contours are not unlike the contours of constant δ_r . We can carry through a similar procedure to determine the contours and the most important ones are $0, \pi/2, \pi$, and $3\pi/2$, that is, the boundaries between the quadrants. The boundary between the first and fourth and between the second and third is simply the real axis, while that between the first and second and the third and fourth is a circle with centre the origin which passes through the point $(\eta_0, 0)$. These contours are shown in figure 2.9(b) where the various quadrants are labelled.

For the purpose of drawing an admittance diagram, it is most convenient to set η in units of \mathcal{Y} , the admittance of free space. The optical admittances will then have the same numerical value as the refractive indices (at normal incidence only, of course).

The method can be illustrated by the same example to be used in the amplitude reflection coefficient loci of figure 2.23

Air|*HL*|Glass

where glass has index 1.52, air 1.0, and *H* and *L* are quarter-waves of zinc sulphide ($n = 2.35$) and cryolite ($n = 1.35$), respectively.

In free space units, the starting admittance is simply 1.52, the admittance of glass. The termination of the first layer, since it is a quarter-wave, will be at an admittance of $2.35^2/1.52 = 3.633$ on the real axis, and of the second, which is also a quarter-wave, at $1.35^2/3.633 = 0.5016$ on the real axis. The circles are traced out clockwise and each is a semicircle with centre on the real axis. Figure 2.10 shows the complete locus.

Metal and other absorbing layers can also be included, although we find the calculations sufficiently involved to require the assistance of a computer. Figure 2.11 shows two loci applying to metal layers, one starting from an admittance of 1.0 and the other from 1.52 (free space units). The higher the

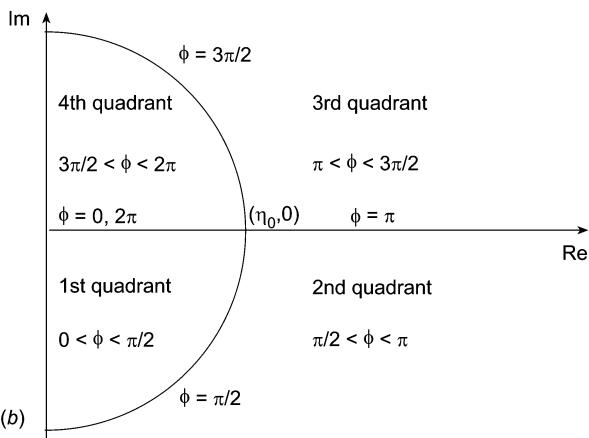
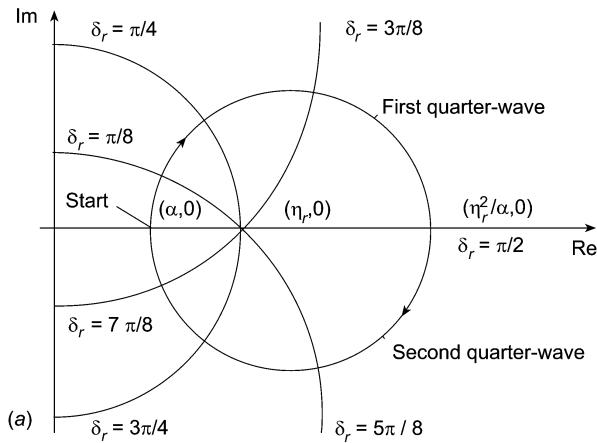


Figure 2.9. (a) Admittance locus of a single dielectric film. The locus is a circle centred on the real axis and described clockwise. The film of characteristic admittance y is assumed to be deposited over a substrate or structure with real admittance α . Note that the product of the admittance of the two points of intersection of the locus with the real axis is always y^2 , the square of the characteristic admittance of the film. Equi-phase thickness contours have also been added to the diagram. (b) Contours of constant phase shift on reflection ϕ can be added to the admittance diagram. These contours are all circles with centres on the imaginary axis and passing through the point on the real axis corresponding to the admittance η_0 of the incident medium. The four most important contours correspond to 0, $\pi/2$, π , $3\pi/2$, and these are represented by portions of the real axis and the circle centred on the origin and passing through the point η_0 . These are indicated on the diagram and the regions corresponding to the various quadrants of ϕ are marked.

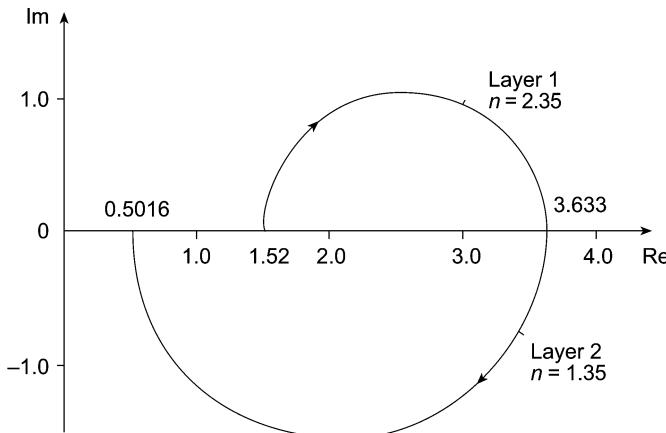


Figure 2.10. The admittance of the coating: Air| $H L$ |Glass, with L a quarter-wave of index 1.35, H of 2.35. The indices of air and glass are 1.00 and 1.52, respectively. This is the same coating as in figure 2.23; note the similarity in shape to that figure.

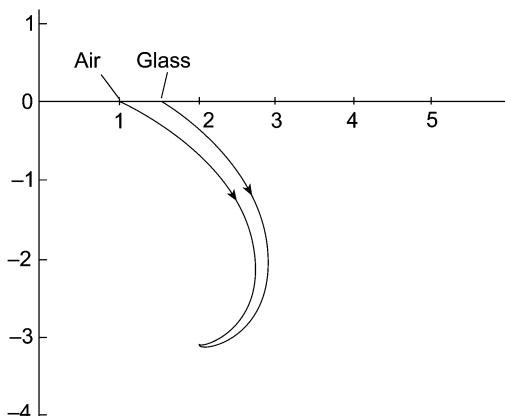


Figure 2.11. Admittance loci corresponding to a metal such as chromium with $n - ik = 2 - i3$. Loci are shown for starting points 1.00 and 1.52, corresponding to air and glass respectively. Note that the initial direction towards the lower right of the diagram implies that in the case of the internal reflectance of the film deposited on glass (i.e. air as substrate and glass as incident medium and the left of the two loci) the reflectance initially falls and then rises, whereas the external reflectance (glass as substrate and air as incident medium and the right of the two admittance loci) always increases, even for very thin layers. When the layers are very thick, they terminate at the point $2 - i3$, so that the film is optically indistinguishable from the bulk material.

ratio k/n for the metal, the nearer the locus is to a circle with centre the origin. In the case of figure 2.11 the locus is somewhat distorted from the ideal case, with a loop bowing out along the direction of the real axis. If we were to add isoreflectance contours to the diagram, corresponding to admittances of 1.52 for the starting admittance of 1.0, and of 1.0 for the starting admittance of 1.52, so that the loci correspond to internal and external reflection from such a metal layer on glass in air, we would see that the observed reduction in internal reflectance when the metal is very thin is predicted by the diagram as well as the constantly increasing external reflectance for the same range of thicknesses (we can see such an effect in figure 4.7). Metals with still lower ratios of k/n depart still further from the ideal circle and in fact those starting at 1.0 can initially loop into the first quadrant so that they actually cut the real axis again, even sometimes at the point 1.52 to give zero internal reflectance.

We have gained much in simplicity by choosing to deal in terms of optical admittance throughout the assembly. It has not affected in any way our ability to calculate either the amplitude reflection coefficient or reflectance. Transmittance is another matter. Strictly we need to preserve the values of B and C in the matrix calculation; the optical admittance is not sufficient. For dielectric assemblies, we know that the transmittance is given by $(1 - R)$, but for assemblies containing absorbing layers, subsidiary calculations are necessary. For many purposes, reflectance is sufficient and, since the graphical technique is used for visualisation rather than calculation, a lack of transmission information is not a serious defect. Nevertheless there are concepts that do yield useful information about transmittance and about losses in layers, directly from the admittance diagram. These are dealt with in the following section.

2.12 Electric field and losses in the admittance diagram

The optical properties of any material are determined largely by the electrons and their interaction with electromagnetic disturbances. Any optical material is made up of atoms or molecules consisting of heavy positively charged masses surrounded by negatively charged electrons. These electrons are light and mobile compared with the heavy positively charged nuclei. An electric field can exert a force on a charged particle even while it is stationary, but a magnetic field can interact only when the charged particle moves, and for any significant interaction, the particle must be moving at an appreciable fraction of the speed of light. At the very high frequencies of optical waves the magnetic interaction is virtually zero. We have already used the fact that the relative permeability is unity in setting up the basic theory. The interaction between light and a material is, therefore, entirely through the electric field. Where the electric field amplitude is high the potential for interaction is high. When thin-film optical coatings are illuminated by light, standing wave patterns form which can exhibit considerable variations in electric field amplitude both in terms of wavelength and position

within the coating. The admittance diagram permits a simple technique for assessing these amplitude variations and from them deductions about losses can be made, sometimes with surprising results.

In this discussion we limit ourselves to normal incidence. Oblique incidence represents only a very slight extension.

The basic matrix technique for the calculation of the properties of an optical coating actually already contains the electric field and so only a slight modification is required to extract it. The matrix expression, with the usual meaning for the symbols, is

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta & \frac{i \sin \delta}{y} \\ iy \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} 1 \\ y_{\text{exit}} \end{bmatrix}.$$

In this expression B and C and the corresponding terms in the other column matrix are normalised total tangential electric and magnetic fields. The admittances, too, are normalised so that they are in free space units rather than SI units. The first thing we do, therefore, is to restore the expressions to their fundamental form.

$$\begin{bmatrix} E' \\ H' \end{bmatrix} = \begin{bmatrix} \cos \delta & \frac{i \sin \delta}{y} \\ iy \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} E_{\text{exit}} \\ H_{\text{exit}} \end{bmatrix}.$$

Here y is in free space units and so to change it to SI units we must write

$$y = (n - ik)\mathcal{Y},$$

where \mathcal{Y} is the admittance of free space. E and H indicate the complex tangential amplitudes which include the relative phase.

To have absolute values for the total tangential electric field amplitude through the multilayer, it remains simply to give an absolute value to one of the E s. This can be done in a number of ways. The easiest is to put a value on the final tangential component at the emergent interface, that is the interface with the substrate. This is related to the incident irradiance through the transmittance. If the incident irradiance is I_{inc} then

$$\frac{1}{2} \operatorname{Re}(E_{\text{exit}} \cdot H_{\text{exit}}^*) = T \cdot I_{\text{inc}}$$

but

$$H_{\text{exit}} = y_{\text{exit}} E_{\text{exit}}$$

and so

$$\frac{1}{2} \operatorname{Re}(E_{\text{exit}} \cdot y_{\text{exit}}^* E_{\text{exit}}^*) = T \cdot I_{\text{inc}}.$$

Now

$$E \cdot E^* = \mathcal{E}^2$$

giving, with a little manipulation,

$$E_{\text{exit}} = \mathcal{E}_{\text{exit}} = \sqrt{\frac{2T \cdot I_{\text{inc}}}{y_{\text{exit}}}},$$

where y_{exit} must be in SI units, that is siemens.

If the multilayer system is completely free of absorption then there is a simple connection between the variation of admittance through the multilayer, which is the quantity we plot in the admittance diagram, and the electric field amplitude.

The admittance at any point in the multilayer is simply the ratio of the total tangential magnetic and electric fields. These total tangential fields also yield the total net irradiance transmitted by the multilayer. Since this multilayer is free of losses, the transmitted irradiance is constant through the multilayer. Putting all this together gives

$$\begin{aligned} I_{\text{out}} &= \frac{1}{2} \operatorname{Re}(E \cdot H^*) \\ &= \frac{1}{2} \operatorname{Re}(E \cdot Y^* E^*) \\ &= \frac{1}{2} \mathcal{E}^2 \cdot \operatorname{Re}(Y) \end{aligned}$$

i.e.

$$\mathcal{E} = \sqrt{\frac{2I_{\text{out}}}{\operatorname{Re}(Y)}} = \sqrt{\frac{2T \cdot I_{\text{inc}}}{\operatorname{Re}(Y)}} \propto \frac{1}{\sqrt{\operatorname{Re}(Y)}}. \quad (2.131)$$

Contours of constant electric field are therefore lines, normal to the real axis in the admittance diagram. If we put Y in free space units then (2.131) becomes:

$$\mathcal{E} = 27.46 \sqrt{\frac{T \cdot I_{\text{inc}}}{\operatorname{Re}(Y)}} \text{ V m}^{-1}. \quad (2.132)$$

Now let us consider a very thin slice of absorbing material embedded in a multilayer. What can we say about the absorption of this slice? The result is contained in the expression:

$$\begin{bmatrix} E' \\ H' \end{bmatrix} = \begin{bmatrix} \cos \delta & \frac{i \sin \delta}{y} \\ iy \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} E \\ H \end{bmatrix},$$

where the input and exit irradiances are given by

$$I_{\text{in}} = \frac{1}{2} \operatorname{Re}(E' \cdot H'^*) \text{ and } I_{\text{exit}} = \frac{1}{2} \operatorname{Re}(E \cdot H^*).$$

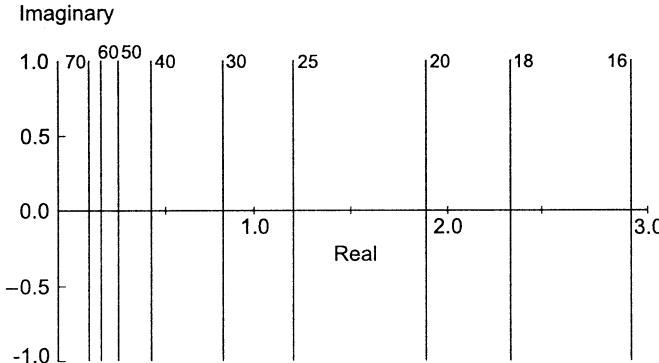


Figure 2.12. Lines of constant electric field amplitude for dielectric materials in the admittance diagram. The figures are in volts per metre if the transmitted irradiance is 1 W m^{-2} .

The irradiance lost by absorption in the layer is the difference between these two quantities. Now let the layer be extremely thin. Since the layer is absorbing, δ is given by

$$\delta = \frac{2\pi (n - ik) d}{\lambda} = \alpha - i\beta. \quad (2.133)$$

Equation (2.133) defines the quantities α and β . By extremely thin, we mean that d/λ should be sufficiently small to make both α and β vanishingly small, whatever the size of either n or k . Then,

$$\begin{aligned} \begin{bmatrix} E' \\ H' \end{bmatrix} &= \begin{bmatrix} \cos(\alpha - i\beta) & \frac{i \sin(\alpha - i\beta)}{y} \\ iy \sin(\alpha - i\beta) & \cos(\alpha - i\beta) \end{bmatrix} \begin{bmatrix} E \\ H \end{bmatrix} \\ &= \begin{bmatrix} 1 & \frac{i(\alpha - i\beta)}{(n - ik)\gamma} \\ i(\alpha - i\beta)(n - ik)\gamma & 1 \end{bmatrix} \begin{bmatrix} E \\ H \end{bmatrix} \\ &= \begin{bmatrix} E + \frac{i(\alpha - i\beta)H}{(n - ik)\gamma} \\ i(\alpha - i\beta)(n - ik)\gamma E + H \end{bmatrix}, \end{aligned}$$

where we are including terms up to the first order only in α and β .

The irradiance at the entrance to this thin layer will then be given by

$$\begin{aligned} I_{\text{in}} &= \frac{1}{2} \operatorname{Re} \left[\left\{ E + \frac{i(\alpha - i\beta)H}{(n - ik)\gamma} \right\} \cdot \{i(\alpha - i\beta)(n - ik)\gamma E + H\}^* \right] \\ &= \frac{1}{2} \operatorname{Re} [E \cdot H^* + E \cdot \{-i(\alpha + i\beta)(n + ik)\gamma E^*\}] \\ &\quad + \frac{1}{2} \operatorname{Re} \left[\frac{i(\alpha - i\beta)H \cdot H^*}{(n - ik)\gamma} \right]. \end{aligned} \quad (2.134)$$

The second of the two terms in (2.134) simplifies to

$$\begin{aligned} \frac{1}{2}\operatorname{Re}\left[\frac{i(\alpha - i\beta)H.H^*}{(n - ik)\gamma}\right] &= \frac{1}{2}\operatorname{Re}\left[\frac{i(\alpha - i\beta)(n + ik)H.H^*}{(n^2 + k^2)\gamma}\right] \\ &= \frac{1}{2}\operatorname{Re}\left[\frac{\{\beta n - \alpha k + i(\alpha n + \beta k)\}H.H^*}{(n^2 + k^2)\gamma}\right] \\ &= \frac{1}{2}\left[\frac{(\beta n - \alpha k)H.H^*}{(n^2 + k^2)\gamma}\right]. \end{aligned}$$

However,

$$\beta n - \alpha k = \frac{2\pi kd}{\lambda}n - \frac{2\pi nd}{\lambda}k = 0.$$

The first term gives

$$\begin{aligned} I_{\text{in}} &= \frac{1}{2}\operatorname{Re}[E \cdot H^* + E \cdot \{-i(\alpha + i\beta)(n + ik)\gamma E^*\}] \\ &= \frac{1}{2}\operatorname{Re}[E \cdot H^*] + \frac{1}{2}[(\alpha k + \beta n)\gamma E \cdot E^*] \end{aligned}$$

where

$$\alpha k + \beta n = \frac{4\pi nkd}{\lambda} \text{ and } E \cdot E^* = \mathcal{E}^2.$$

The irradiance that has been absorbed is therefore given by the difference between the irradiance incident on the thickness element, I_{in} , and that emerging on the exit side, I_{exit} , and this is

$$I_{\text{absorbed}} = \frac{2\pi nkd}{\lambda} \cdot \gamma \cdot \mathcal{E}^2. \quad (2.135)$$

The magnitude of the absorbed energy is directly proportional to the product of n and k . Both must be nonzero for absorption to occur. The absorption will be small both for a metal with vanishingly small n and a dielectric with vanishingly small k . The factor involving n and k may be thought of as a phase thickness multiplied by k or as a quantity β multiplied by n .

Now we need to consider the contribution to the absorption A of the multilayer. This is a little more difficult and we need to introduce a further concept that will be used in subsequent chapters.

Potential transmittance, ψ , of any element of a coating system is defined as the ratio of the output to the input irradiances, the input being the net irradiance rather than the incident. Potential transmittance has several advantages over transmittance when dealing with absorbing systems because it completely avoids any problems associated with the mixed Poynting vector in absorbing media.

The potential transmittance of a complete system is simply the product of the individual potential transmittances.

$$\psi = \frac{I_{\text{exit}}}{I_{\text{in}}}$$

$$\psi_{\text{system}} = \psi_1 \cdot \psi_2 \cdot \psi_3 \cdot \psi_4 \cdot \psi_5 \dots \psi_q$$

with the eventual overall transmittance given by

$$T = (1 - R) \cdot \psi_{\text{system}}.$$

The potential transmittance of the thin elemental film is given by

$$\psi = \frac{I_{\text{exit}}}{I_{\text{in}}} = 1 - \frac{I_{\text{absorbed}}}{I_{\text{in}}} = 1 - \mathcal{A},$$

where \mathcal{A} is the potential absorptance. But

$$I_{\text{in}} = \frac{1}{2} \mathcal{Y} \cdot \text{Re}(Y) \cdot \mathcal{E}^2$$

where Y is given in free space units. Then

$$\psi = 1 - \mathcal{A} = 1 - \frac{2\pi nkd}{\lambda} \cdot \frac{2}{\text{Re}(Y)}. \quad (2.136)$$

This result allows interpretation of an admittance locus in terms of potential absorption.

To move from potential absorption to absorption is straightforward when the absorption is confined to a very thin layer, the rest of the multilayer being essentially transparent. Then the absorption, A , is given by:

$$A = (1 - R)\mathcal{A}.$$

If, however, the absorption is distributed through the layer then the calculation is rather more involved. Normally the absorption would be calculated by the normal matrix expression for the entire film and would be completely accurate. We, however, are looking for a way of estimating the absorption and its variation through a layer given the locus in the admittance diagram or the electric field distribution. Let us assume that the absorption is rather small. The layer may be considered as a succession of slices of equal optical thickness and extinction coefficient, and so the first factor in the expression for A is a constant. Each slice has a potential absorptance that depends on the real part of the optical admittance following equation (2.136). The potential transmittance is given by the product of the individual potential transmittances,

$$\begin{aligned} \psi &= \psi_1 \cdot \psi_2 \cdot \psi_3 \cdot \psi_4 \dots \\ &= (1 - \mathcal{A}_1) \cdot (1 - \mathcal{A}_2) \cdot (1 - \mathcal{A}_3) \dots \\ &= 1 - (\mathcal{A}_1 + \mathcal{A}_2 + \mathcal{A}_3 + \mathcal{A}_4 + \dots) + \mathcal{A}_1 \mathcal{A}_2 + \dots \end{aligned}$$

Provided the potential absorptances are small enough the product terms can be neglected and then the total potential absorptance is given by the sum of the individual absorptances,

$$\mathcal{A} = \mathcal{A}_1 + \mathcal{A}_2 + \mathcal{A}_3 + \mathcal{A}_4 + \dots \quad (2.137)$$

In terms of an integral this can be written as

$$\mathcal{A} = \sum_j \mathcal{A}_j = \int_{\delta} \frac{2k}{\text{Re}(Y)} d\delta = \int_{\beta} \frac{2n}{\text{Re}(Y)} d\beta. \quad (2.138)$$

If an accurate answer is required we will always turn to the computer and a very simple rapid calculation. For understanding the result, usually we would like to know what to do either to increase or decrease the absorptance or to find sensitive regions where contamination or scattering roughness is especially to be avoided. To answer such questions usually a rough answer that shows trends is all that is needed.

2.13 The vector method

The vector method is a valuable technique, especially in design work associated with antireflection coatings. Two assumptions are involved: first, that there is no absorption in the layers, and second, that the behaviour of a multilayer can be determined by considering one reflection of the incident wave at each interface only. The errors involved in using this method can, in some cases, be significant, especially where high overall reflectance from the multilayer exists, but they are small in most types of antireflection coating.

Consider the assembly sketched in figure 2.9. If there is no absorption in the layers, then $N_r = n_r$ and $k_r = 0$. The amplitude reflection coefficient at each interface is given by

$$\rho = \frac{n_{r-1} - n_r}{n_{r-1} + n_r}$$

which may be positive or negative depending on the relative magnitudes of n_{r-1} and n_r .

The phase thicknesses of the layers are given by $\delta_1, \delta_2, \dots$, where

$$\delta_r = 2\pi n_r d_r / \lambda.$$

A quarter-wave optical thickness is represented by 90° and a half-wave by 180° .

As the diagram shows, the resultant amplitude reflection coefficient is given by the vector sum of the coefficients for each interface, where each is associated with the appropriate phase lag corresponding to the passage of the wave from the front surface to the interface and back to the front surface again.

$$\begin{aligned} \rho = \rho_a &+ \rho_b \exp(-2i\delta_1) + \rho_c \exp[-2i(\delta_1 + \delta_2)] \\ &+ \rho_d \exp[-2i(\delta_1 + \delta_2 + \delta_3)] + \dots \end{aligned}$$

The sum can be found analytically, or, as is more usual, graphically. The graphical case is easier because the angles between successive vectors are merely $2\delta_1$, $2\delta_2$, $2\delta_3$, and so on.

The calculation of the angles for any wavelength is simplified if, as is usual, the optical thicknesses of the layers are given in terms of quarter-wave optical thicknesses at a reference wavelength λ_0 . If the optical thickness of the r th layer is t_r quarter-waves at λ_0 , then the value of δ_r at λ is just $\delta_r = (90^\circ t_r \lambda_0 / \lambda)$ degrees of arc.

In practice it will be found extremely easy to confuse angles and directions, particularly where negative reflectances are involved. The task of drawing the vector diagram is greatly eased by plotting first the vectors with directions on a polar diagram and then transferring the vectors to a vector polygon rather than attempting to draw the vector polygon straight away. An important point to remember is that the resultant vector represents the amplitude reflection coefficient and its length must be squared in order to give the reflectance.

A typical arrangement is shown in figure 2.13. The vector method is used to a considerable extent in chapter 3, which deals with antireflection coatings.

2.14 Incoherent reflection at two or more surfaces

So far, we have treated substrates as being one-sided slabs of material of infinite depth. In almost all practical cases, the substrate will have finite depth with rear surfaces that reflect some of the energy and affect the performance of the assembly.

The depth of the substrate will usually be much greater than the wavelength of the light and variations in the flatness and parallelism of the two surfaces will be appreciable fractions of a wavelength. Generally the incident light will not be particularly well collimated. Under these conditions it will not be possible with a finite aperture to observe interference effects between light reflected at the front and rear surfaces of the substrate, and because of this the substrate is known as thick. The waves reflected successively at the front and back surfaces add incoherently instead of coherently. The resultant is the sum of the various intensities instead of the vector sum of the amplitudes. It can be shown that incoherent addition yields the same result as the averaging of the coherent result over any moderate geometrical area or wavelength interval, or range of angles of incidence, such that an appreciable number of fringes is included in the interval.

The symbols used are illustrated in figure 2.14. Waves are reflected successively at the front and rear surfaces. The sums of the irradiances are given by

$$\begin{aligned} R &= R_a^+ + T_a^+ R_b^+ T_a^- \left[1 + R_a^- R_b^+ + (R_a^- R_b^+)^2 + \dots \right] \\ &= R_a^+ + [T_a^+ R_b^+ T_a^- / (1 - R_a^- R_b^+)], \end{aligned}$$

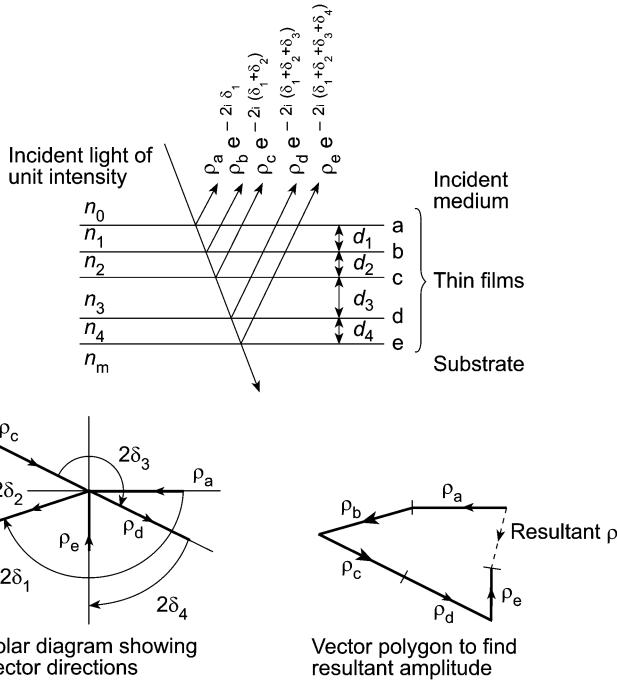


Figure 2.13. The vector method. The lengths of the vectors and the phase angles are given by

$$\begin{aligned}\rho_a &= (n_0 - n_1) / (n_0 + n_1) & \delta_1 &= 2\pi n_1 d_1 / \lambda \\ \rho_b &= (n_1 - n_2) / (n_1 + n_2) & \delta_2 &= 2\pi n_2 d_2 / \lambda\end{aligned}$$

etc. Note that the sign of the expression for the vector lengths is important and must be included. In the diagram \$\rho_a, \rho_c\$ and \$\rho_e\$, are shown as of negative sign. Note also that the angles between successive vectors are phase lags, so that the sense of all the angles in the polar diagram, \$\delta_1, \delta_2\$, etc, is also negative.

i.e. since \$T^+\$ and \$T^- are always identical

$$T_a^+ = T_a^- = T_a$$

and so

$$R = \frac{R_a^+ + R_b^+ (T_a^2 - R_a^- R_b^+)}{1 - R_a^- R_b^+}.$$

If there is no absorption in the layers,

$$R_a^+ = R_a^- = R_a \quad \text{and} \quad 1 = R_a + T_a$$

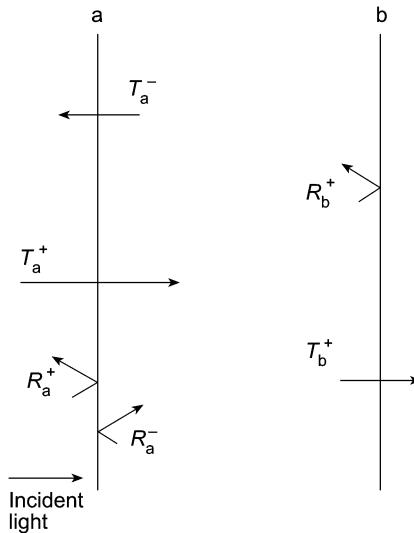


Figure 2.14. Symbols used in calculation of incoherent reflection at two or more surfaces.

so that

$$R = \frac{R_a + R_b - 2R_a R_b}{1 - R_a R_b}.$$

Similarly

$$\begin{aligned} T &= T_a^+ T_b^+ \left[1 + R_a^- R_b^+ + (R_a^- R_b^+)^2 + \dots \right] \\ &= \frac{T_a T_b}{1 - R_a^- R_b^+} \end{aligned}$$

and again, if there is no absorption,

$$T = \frac{T_a T_b}{1 - R_a R_b} \quad (2.139)$$

or

$$T = \left(\frac{1}{T_a} + \frac{1}{T_b} - 1 \right)^{-1} \quad (2.140)$$

since

$$R_a = 1 - T_a \quad R_b = 1 - T_b.$$

A nomogram for solving equation (2.140) can easily be constructed. Two axes at right angles are laid out on a sheet of graph paper and, taking the point of

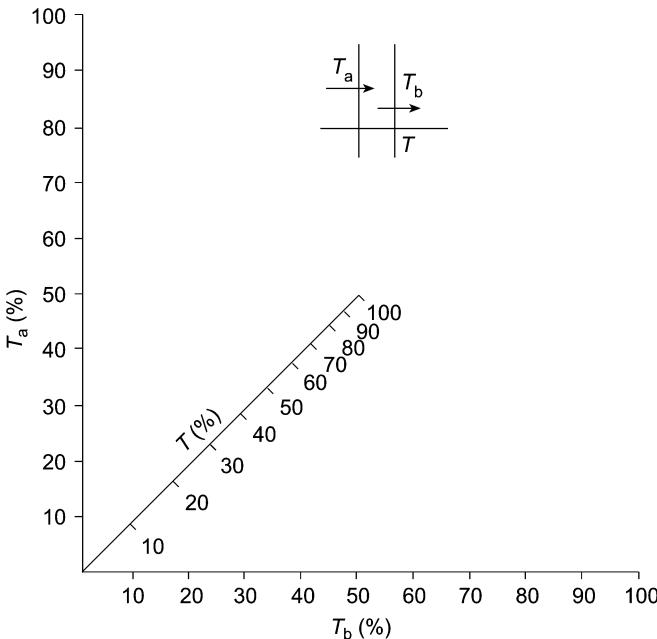


Figure 2.15. A nomogram for calculating the overall transmittance of a thick transparent plate given the transmittance of each individual surface.

intersection as the zero, two linear equal scales of transmittance are marked out on the axes. One of these is labelled T_a and the other T_b . The angle between T_a and T_b is bisected by a third axis which is to have the T scale marked out on it. To do this, a straight edge is placed so that it passes through the 100% transmittance value on, say, the T_a axis and any chosen transmittance on the T_b axis. The value of T to be associated with the point where the straight edge crosses the T axis is then that of the intercept with the T_b axis. The entire scale can be marked out in this way. A completed nomogram of this type is shown in figure 2.15

In the absence of absorption, the analysis can be very simply extended to further surfaces. Consider the case of two substrates, i.e. four surfaces. These we can label T_a , T_b , T_c and T_d . Then, from equation (2.140), we have for the first substrate

$$T_1 = \left(\frac{1}{T_a} + \frac{1}{T_b} - 1 \right)^{-1},$$

i.e.

$$\frac{1}{T_1} = \frac{1}{T_a} + \frac{1}{T_b} - 1$$

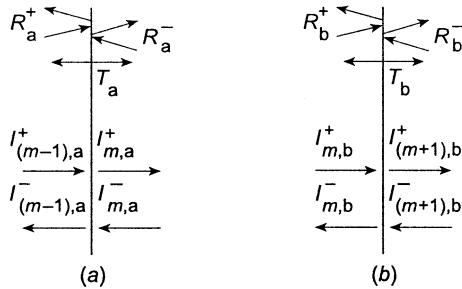


Figure 2.16. Symbols defining two successive coatings with intervening medium in a stack.

and similarly for the second

$$\frac{1}{T_2} = \frac{1}{T_c} + \frac{1}{T_d} - 1.$$

The transmittance through all four surfaces is then obtained by applying equation (2.140) once again:

$$\frac{1}{T} = \frac{1}{T_1} + \frac{1}{T_2} - 1,$$

i.e.

$$T = \left(\frac{1}{T_a} + \frac{1}{T_b} + \frac{1}{T_c} + \frac{1}{T_d} - 3 \right)^{-1}. \quad (2.141)$$

The iterative nature of these calculations can be clumsy when dealing with a succession of surfaces. A technique based on a study by Baumeister *et al* [9] yields a rather more useful matrix form of the calculation. The emphasis is placed on the flows of irradiance. Absorption in the media between the coated surfaces is supposedly sufficiently small so that the coupling problem mentioned earlier is negligible. The symbols are defined in figure 2.16.

The direction of the light is denoted by the usual plus and minus signs. *a* and *b* are two coatings separated by a medium *m* with internal transmittance T_{mint} . The final medium will be the emergent medium and there, the negative-going irradiance will be zero. The procedure to be outlined will derive the values of $I_{m,a}^+$ and $I_{m,a}^-$ from $I_{(m+1),b}^+$ and $I_{(m+1),b}^-$. The rest is straightforward.

The irradiances on either side of the coating with label *b* are related through the equations

$$\begin{aligned} I_{(m+1),b}^+ &= T_b I_{m,b}^+ + R_b^- I_{(m+1),b}^- \\ I_{m,b}^- &= R_b^+ I_{m,b}^+ + T_b I_{(m+1),b}^-. \end{aligned}$$

These can be manipulated into the form

$$\begin{aligned} I_{m b}^- &= \frac{1}{T_b} \left\{ R_b^+ I_{(m+1) b}^+ + \left(T_b^2 - R_b^- R_b^+ \right) I_{(m+1) b}^- \right\} \\ I_{m b}^+ &= \frac{1}{T_b} \left\{ I_{(m+1) b}^+ - R_b^- I_{(m+1) b}^- \right\} \end{aligned}$$

and in matrix form this is

$$\begin{bmatrix} I_{m b}^- \\ I_{m b}^+ \end{bmatrix} = \begin{bmatrix} \frac{(T_b^2 - R_b^- R_b^+)}{T_b} & \frac{R_b^+}{T_b} \\ \frac{-R_b^-}{T_b} & \frac{1}{T_b} \end{bmatrix} \begin{bmatrix} I_{(m+1) b}^- \\ I_{(m+1) b}^+ \end{bmatrix}. \quad (2.142)$$

The conversion through the medium is given by

$$\begin{bmatrix} I_{m a}^- \\ I_{m a}^+ \end{bmatrix} = \begin{bmatrix} T_{m \text{ int}} & 0 \\ 0 & \frac{1}{T_{m \text{ int}}} \end{bmatrix} \begin{bmatrix} I_{m b}^- \\ I_{m b}^+ \end{bmatrix}. \quad (2.143)$$

Equations (2.142) and (2.143) can be applied to the various coatings and intervening media in succession.

2.15 Other techniques

Great progress was made in the subject of thin-film optics well before computers became both exceedingly powerful and generally available. Many techniques for assisting in the creation and assessment of designs were developed at a time when accurate extended calculations were so time consuming as to be out of the question. Their usefulness has not ceased with the advent of the personal computer because they bring an insight that is completely lacking in pure numerical calculations. Some of these techniques we will use from time to time in the remainder of the book. Others are commonly encountered in the literature of the subject. The fact that we collect a number of them together under the appellation of ‘other’ should not be taken as an indication of a reduced usefulness or ranking but rather as an admission that there is a limit to the size of this book. There are many others that we have simply been completely unable to include.

2.15.1 The Herpin index

An extremely important result for filter design is derived in chapter 6, which deals with edge filters. Briefly, this is the fact that any symmetrical product of three thin-film matrices can be replaced by a single matrix which has the same form as that of a single film and which therefore possesses an equivalent thickness and an equivalent optical admittance. Of course, this is a mathematical device rather than a case of true physical equivalence, but the result is of considerable use in giving an insight into the properties of a great number of filter designs which can be split into a series of symmetrical combinations. The method also allows

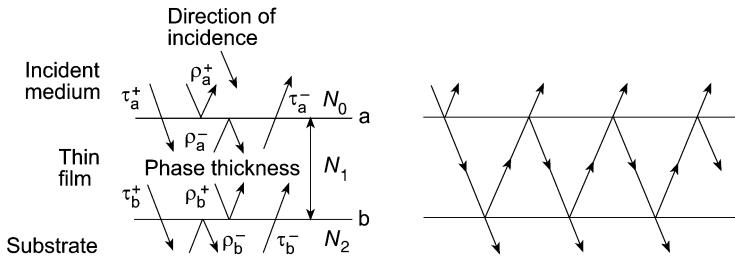


Figure 2.17. Parameters in the multiple beam summation.

the replacement, under certain conditions, of a layer of intermediate index by a symmetrical combination of high- and low-index material. This is especially useful in the design of antireflection coatings, which frequently require quarter-wave thicknesses of unobtainable intermediate indices. These difficult layers can be replaced by symmetrical combinations of existing materials with the additional advantage of limiting the total number of materials required for the structure.

The equivalent admittance is frequently known as the Herpin index, after the originator, and the symmetrical combination as an Epstein period, after the author of two of the most important early papers dealing with the application of the result to the design of filters.

The detailed derivation of the relevant formulae is left until chapter 6, which will make considerable use of the concept.

2.15.2 Alternative method of calculation

The success of the vector method prompts one to ask whether it can be made more accurate by considering second and subsequent reflections at the various boundaries instead of just one. In fact, an alternative solution of the thin-film problem can be obtained in this way and this was the earlier way of formulating film properties dating back to Poisson (chapter 1). It is simpler to consider normal incidence only. The expressions can be adapted for non-normal incidence quite simply when the materials are transparent and with some difficulty when they are absorbing. We consider first the case of a single film. Figure 2.17 defines the various parameters.

The resultant amplitude reflection coefficient is given by

$$\begin{aligned} \rho^+ &= \rho_a^+ + \tau_a^+ \rho_b^+ \tau_a^- e^{-2i\delta} + \tau_a^+ \rho_b^+ \rho_a^- \rho_b^+ \tau_a^- e^{-4i\delta} \\ &\quad + \tau_a^+ \rho_b^+ \rho_a^- \rho_b^+ \rho_a^- \rho_b^+ \tau_a^- e^{-6i\delta} \\ &= \rho_a^+ + \frac{\rho_b^+ \tau_a^+ \tau_a^- e^{-2i\delta}}{1 - \rho_b^+ \rho_a^- e^{-2i\delta}}. \end{aligned}$$

However,

$$\tau_a^+ \tau_a^- = \frac{4N_0 N_1}{(N_0 + N_1)^2} = 1 - \rho$$

and $\rho_a^- = -\rho_a^+$ so that

$$\rho^+ = \frac{\rho_a^+ + \rho_b^+ e^{-2i\delta}}{1 + \rho_b^+ \rho_a^+ e^{-2i\delta}}. \quad (2.144)$$

Similarly

$$\tau^+ = \tau_a^+ \tau_b^+ \rho_b^+ e^{-i\delta} + \tau_a^+ \rho_b^+ \rho_a^- \tau_b^+ e^{-3i\delta} + \tau_a^+ \rho_b^+ \rho_a^- \rho_b^+ \rho_a^- \tau_b^+ e^{-5i\delta}$$

which reduces to

$$\begin{aligned} \tau^+ &= \frac{\tau_a^+ \tau_b^+ e^{-i\delta}}{1 - \rho_a^- \rho_b^+ e^{-2i\delta}} \\ &= \frac{\tau_a^+ \tau_b^+ e^{-i\delta}}{1 + \rho_a^+ \rho_b^+ e^{-2i\delta}}. \end{aligned} \quad (2.145)$$

These expressions can be used in calculations of assemblies of more than one film by applying them successively, first to the final two interfaces which can then be replaced by a single interface with the resultant coefficients, and then to this equivalent interface and the third last interface, and so on.

The resultant amplitude transmission and reflection coefficients τ^+ and ρ^+ can be converted into transmittance and reflectance using the expressions

$$\begin{aligned} R &= (\rho^+) (\rho^+)^* \\ T &= \frac{n_2}{n_0} (\tau^+) (\tau^+)^*. \end{aligned}$$

n_2 and n_0 are the refractive indices of the substrate, or exit medium, and the incident medium, respectively. For these expressions to be meaningful we must, as before, restrict the incident medium to be transparent so that $N_0 = n_0$. No such restriction applies to the exit medium which can have complex $N_2 = n_1 - ik_2$, the real part being used in the above expression for T .

It is also possible to develop a matrix approach along these lines. The electric field vectors E_0^+ and E_0^- in medium 0 at interface a can be expressed in terms of E_1^+ and E_1^- in film 1 at interface b (see figure 2.18).

$$\begin{bmatrix} E_0^+ \\ E_0^- \end{bmatrix} = \frac{1}{\tau_a^+} \begin{bmatrix} e^{i\delta_1} & \rho_a^+ e^{-i\delta_1} \\ \rho_a^+ e^{i\delta_1} & e^{-i\delta_1} \end{bmatrix} \begin{bmatrix} E_1^+ \\ E_1^- \end{bmatrix}. \quad (2.146)$$

If E_2^+ is the tangential component of amplitude in medium 2, then, since there is only a positive-going wave in that medium

$$\begin{bmatrix} E_1^+ \\ E_1^- \end{bmatrix} = \frac{1}{\tau_b^+} \begin{bmatrix} 1 \\ \rho_b^+ \end{bmatrix} E_2^+. \quad (2.147)$$

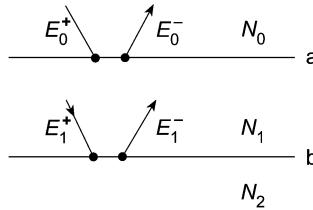


Figure 2.18. The positive- and negative-going waves at the two interfaces.

Equations (2.146) and (2.147) can be extended in the normal way to cover the case of many layers. The only point to watch is that ρ_a^+ and τ_a^+ must refer to the coefficients of the boundary in the correct medium. That is, all the reflection coefficients ρ , and transmission coefficients τ , must be calculated for the boundaries as they exist in the multilayer. Thus, if we take an existing multilayer and add an extra layer, not only do we add an extra interface but we alter the amplitude reflection and transmission coefficients of what now becomes the second last interface. Thus two layers must be recomputed and not just one.

If absorption is included, the formulae remain the same but the parameters ρ , τ and δ become complex.

2.15.3 Smith's method of multilayer design

In 1958, Smith [10], then of the University of Reading, published a useful design method based on equation (2.145). The technique is also known as the method of effective interfaces. It consists of choosing any layer in the multilayer and then considering multiple reflections within it, the reflection and transmission coefficients at its boundaries being the resultant coefficients of the complete structures on either side. The method of summing multiple beams is, of course, quite old and the novel feature of the present technique is the way in which it is applied. Although the technique described by Smith was principally concerned with dielectric multilayers, it can be extended to deal with absorbing layers. As before, we limit ourselves, in the derivation, to normal incidence. When the layers are transparent, the expressions can be extended to oblique incidence without major difficulty. The notation is illustrated in figure 2.19.

From equation (2.145)

$$\tau^+ = \frac{\tau_a^+ \tau_b^+ e^{-i\delta}}{1 - \rho_a^- \rho_b^+ e^{-2i\delta}}$$

where

$$\delta = 2\pi Nd/\lambda.$$

Now $N = n - ik$ and we can write δ as

$$\delta = 2\pi(n - ik)d/\lambda = \alpha + i\beta$$

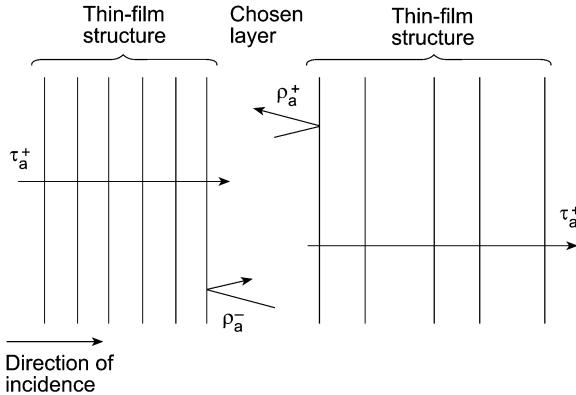


Figure 2.19. The quantities associated with the effective interfaces in Smith's technique.

and

$$e^{-i\delta} = e^{-\beta} e^{-i\alpha}$$

where $\alpha = 2\pi nd/\lambda$, the phase thickness of the layer, and $\beta = 2\pi kd/\lambda$. Now

$$T = \frac{n_m}{n_0} (\tau^+) (\tau^+)^*,$$

where n_m is the real part of the exit medium index and n_0 is the refractive index of the incident medium.

$$T = \frac{n_m}{n_0} \frac{(\tau_a^+) (\tau_a^+)^* (\tau_b^+) (\tau_b^+)^* e^{-2\beta}}{(1 - \rho_a^- \rho_b^+ e^{-2\beta} e^{-2i\alpha}) (1 - \rho_a^- \rho_b^+ e^{-2\beta} e^{-2i\alpha})^*}.$$

Now, let

$$\begin{aligned} \tau_a^+ &= |\tau_a^+| e^{i\varphi_a'} & \rho_a^- &= |\rho_a^-| e^{i\varphi_a} \\ \tau_b^+ &= |\tau_b^+| e^{i\varphi_b'} & \rho_b^+ &= |\rho_b^+| e^{i\varphi_b}. \end{aligned}$$

Then,

$$T = \frac{n_m}{n_0} \times \frac{|\tau_a^+|^2 |\tau_b^+|^2 e^{-2\beta}}{\left(1 - |\rho_a^-|^2 |\rho_b^+|^2 e^{i(\varphi_a + \varphi_b)} e^{-2\beta} e^{-2i\alpha}\right) \left(1 - |\rho_a^-|^2 |\rho_b^+|^2 e^{-i(\varphi_a + \varphi_b)} e^{-2\beta} e^{2i\alpha}\right)},$$

i.e.

$$T = \frac{n_m}{n_0} \frac{|\tau_a^+|^2 |\tau_b^+|^2 e^{-2\beta}}{\left[1 - |\rho_a^-|^2 |\rho_b^+|^2 e^{-4\beta} - 2 |\rho_a^-| |\rho_b^+| e^{-2\beta} \cos(\varphi_a + \varphi_b - 2\alpha)\right]}. \quad (2.148)$$

A marginally more convenient form of the expression can be obtained by substituting $1 - 2 \sin^2[(\varphi_a + \varphi_b)/2 - \alpha]$ for $\cos(\varphi_a + \varphi_b - 2\alpha)$, and with some rearrangement

$$T = \frac{n_m}{n_0} \frac{|\tau_a^+|^2 |\tau_b^+|^2 e^{-2\beta}}{(1 - |\rho_a^-| |\rho_b^+| e^{-2\beta})^2} \cdot \left[1 + \frac{4 |\rho_a^-| |\rho_b^+| e^{-2\beta}}{(1 - |\rho_a^-| |\rho_b^+| e^{-2\beta})^2} \right. \\ \times \sin^2 \left(\frac{\varphi_a + \varphi_b}{2} - \frac{2\pi nd}{\lambda} \right) \left. \right]^{-1}. \quad (2.149)$$

If there is no absorption in the chosen layer, i.e. $\beta = 0$, then the restrictions on reflectances in absorbing media no longer apply and we can write

$$\begin{aligned} T_a &= \frac{n}{n_0} |\tau_a^+|^2 & R_a^- &= |\rho_a^-|^2 \\ T_b &= \frac{n_m}{n} |\tau_b^+|^2 & R_b^- &= |\rho_b^+|^2 \end{aligned}$$

$$T = \frac{T_a T_b}{\left[1 - (R_a^- R_b^+)^{1/2} \right]^2} \cdot \left[1 + \frac{4 R_a^- R_b^+}{\left[1 - (R_a^- R_b^+)^{1/2} \right]^2} \right. \\ \times \sin^2 \left(\frac{\varphi_a + \varphi_b}{2} - \frac{2\pi nd}{\lambda} \right) \left. \right]^{-1} \quad (2.150)$$

which is the more usually quoted version.

The usefulness of this method is mainly in providing an insight into the properties of a particular type of filter, and it is of considerable value in design. It is certainly not the easiest method of determining the performance of a given multilayer—this is best tackled by a straightforward application of the matrix method. What equations (2.149) or (2.150) do is to make it possible to isolate a layer, or a combination of several layers, and to examine the influence which these layers and any changes in them have on the performance of the filter as a whole. Smith's original paper includes a large number of examples of this approach and repays close study.

2.15.4 The Smith chart

The Smith chart is one of a number of different devices of the same broad type that were originally intended to simplify calculation. The Smith chart is the one which appears most frequently in the literature and so it is included here, although little use is made of it in the remainder of the book. The method depends on three properties of a thin-film structure.

1. Since the tangential components of \mathbf{E} and \mathbf{H} are continuous across a boundary, so also is the equivalent admittance. This has been implied in the

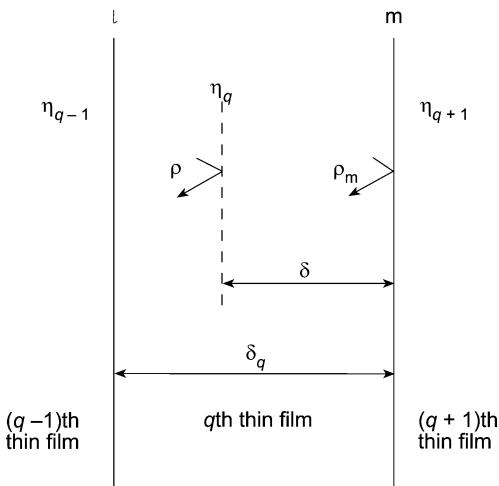


Figure 2.20. Parameters used in the Smith chart description.

section dealing with the matrix method, but has not, perhaps, been explicitly stated there.

2. In any thin film, for example layer q in figure 2.20, the amplitude reflectance ρ at any plane within the layer is related to that at the edge of the layer remote from the incident wave ρ_m by

$$\rho = \rho_m e^{-2i\delta}, \quad (2.151)$$

where δ is the phase thickness of that part of the layer between the far boundary m and the plane in question.

This second point is almost self-evident, but may be shown by putting $\rho_a^+ = 0$ in equation (2.145), since the boundary under consideration is an imaginary one between two media of identical admittance.

3. The amplitude reflection coefficient of any thin-film assembly, with optical admittance at the front surface of Y , is given by equation (2.144), i.e.

$$\rho = \frac{\eta_0 - Y}{\eta_0 + Y} = \frac{1 - Y/\eta_0}{1 + Y/\eta_0}, \quad (2.152)$$

where η_0 is the admittance of the incident medium. Y/η_0 is sometimes known as the reduced admittance.

The procedure for calculating the effect of any layer in a thin-film assembly by using these properties is as follows.

- (i) ρ_m , the amplitude reflection coefficient at the boundary of the layer remote from the side of incidence, is given.

- (ii) The amplitude reflection coefficient within the layer just inside boundary 1 is then given by equation (2.151):

$$\rho = \rho_m e^{-2i\delta_q}. \quad (2.153)$$

- (iii) The optical admittance just inside boundary 1 is given by equation (2.152):

$$\rho = \frac{1 - Y/\eta_q}{1 + Y/\eta_q}, \quad (2.154)$$

i.e.

$$\frac{Y}{\eta_q} = \frac{1 - \rho}{1 + \rho}. \quad (2.155)$$

- (iv) The optical admittance on the incident side of boundary 1 is still Y because of condition 1 above. The reduced admittance is Y/η_{q-1} where

$$\frac{Y}{\eta_{q-1}} = \frac{\eta_q}{\eta_{q-1}} \cdot \frac{Y}{\eta_q}. \quad (2.156)$$

- (v) The amplitude reflection coefficient ρ_1 on the incident side of boundary 1 is given by

$$\rho_1 = \frac{1 - Y/\eta_{q-1}}{1 + Y/\eta_{q-1}}. \quad (2.157)$$

Calculation of the amplitude reflection coefficient of any thin-film assembly is merely the successive application of equations (2.153)–(2.157) to each layer in the system, starting with that at the end of the assembly remote from the incident wave.

The calculation can be carried out in any convenient way, and can even be used as the basis for a computer program. The problem is similar to one found in the study of high-frequency transmission lines and a simple graphical approach has been devised. The most awkward parts of the calculation are in equations (2.155) and (2.157). A chart connecting values of X and Z , where

$$X = \frac{1 - Z}{1 + Z} \quad (2.158)$$

is shown in figure 2.21 and is known as a Smith chart after the originator P H Smith (not to be confused with the S D Smith of the previous section). Z is plotted in polar coordinates on the diagram and the corresponding real and imaginary parts of X are read off from the sets of orthogonal circles. A slide rule is capable of the other part of the calculation, the multiplication by η_q/η_{q-1} .

A scale is provided around the outside of the chart to enable the calculation involved in equation (2.153) to be very simply carried out by rotating the point corresponding to ρ_m around the centre of the chart through the appropriate angle $2\delta_q$. The scale is calibrated in terms of optical thickness measured in fractions of a wavelength, taking into account that the angle is actually $2 \times \delta_q$.

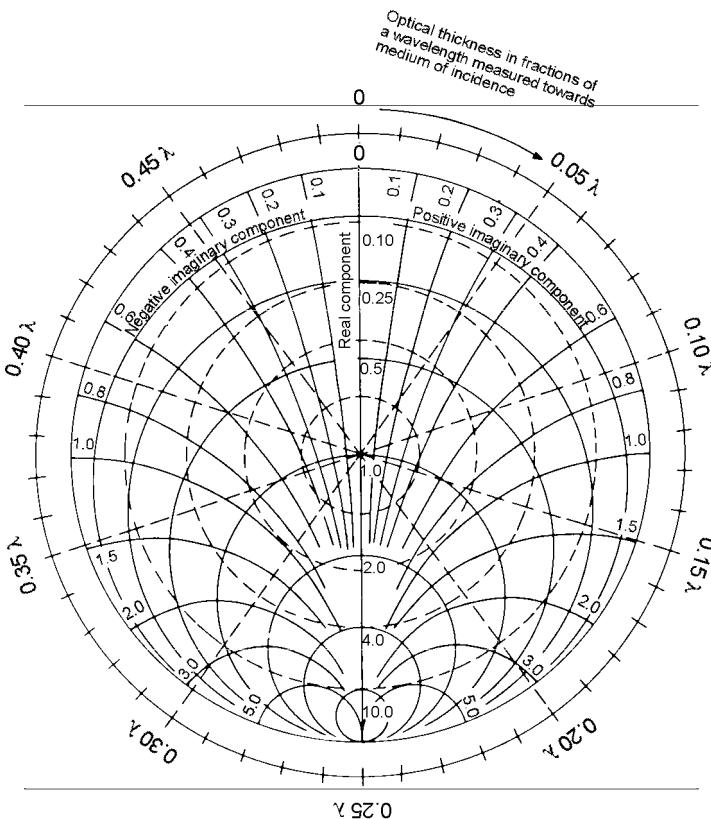


Figure 2.21. The Smith chart. Dashed circles are circles of constant amplitude reflection coefficient ρ . From the smallest to the largest they correspond to $\rho = 0.2, 0.4, 0.6, 0.8$ and 1.0 , the outer solid circle. Solid circles are circles of constant real part and constant imaginary part of the reduced optical admittance. Note: an optical thickness of 0.25λ corresponds to a phase thickness of 90° . (This Smith chart was constructed using the details given in Jackson W 1951 *High Frequency Transmission Lines* 3rd edn (London: Methuen) pp 129 and 146.)

2.15.5 Reflection circle diagrams

This technique, sometimes referred to simply as a circle diagram, was described by Berning [4] and its use in coating design was considerably developed and described in much detail by Apfel [11]. According to Apfel, Frank Rock originated this technique in the mid-1950s. The technique results in diagrams that have an appearance similar to that of the admittance diagram.

The scale and shape of the diagram is similar to that of the Smith chart and, indeed, the identical set of coordinates and prepared graph paper may be used for

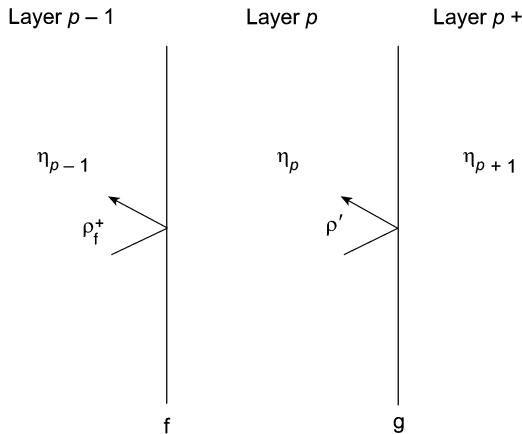


Figure 2.22. Quantities in the method of reflection circles.

both. This leads to a confusion of the two techniques with the name Smith chart being applied to the circle diagram. They are really quite different. The Smith chart slides a reference plane through an already existing multilayer and plots the net amplitude reflection coefficient at the plane. There are discontinuities in the locus, therefore, when an interface is crossed. Dielectric loci are circles centred at the origin. The circle diagram assumes that the multilayer is under construction so that the incident medium for the amplitude reflection coefficient is the incident medium for the entire multilayer. This results also in circles but there are no discontinuities in the resulting locus and the individual dielectric circles are no longer centred at the origin.

Equation (2.144) gives an expression for calculating the change in amplitude reflection coefficient resulting from the addition of a single layer:

$$\rho^+ = \frac{\rho_a^+ + \rho_b^+ e^{-2i\delta}}{1 + \rho_b^+ \rho_a^+ e^{-2i\delta}}.$$

We can calculate the properties of a multilayer by successive applications of this formula, as has already been indicated. Let us imagine that we have arrived at the p th layer in the calculation. The quantities involved are indicated in figure 2.22. ρ_f^+ is the amplitude reflection coefficient of the $(p - 1)$ th layer at the outer interface, which we have labelled f.

$$\rho_f^+ = \frac{\eta_{p-1} - \eta_p}{\eta_{p-1} + \eta_p}.$$

ρ' in figure 2.22 is the resultant amplitude reflection coefficient at the inner interface of the p th layer due to the entire structure on that side and is not to be confused with ρ_g , the amplitude reflection coefficient of the q th interface. The

resultant amplitude reflection coefficient ρ at the f th interface is given by

$$\rho = \frac{\rho_f^+ + \rho' e^{-2i\delta}}{1 + \rho_f^+ \rho' e^{-2i\delta}}. \quad (2.159)$$

Provided we are dealing with dielectric materials ρ_f^+ will be real. ρ' may be complex but we can include any phase angle due to ρ' in the factor $e^{-2i\delta}$. Let us plot the locus of ρ in the complex plane as δ varies. To simplify the analysis, we can replace ρ by $x + iy$ and $\rho' e^{-2i\delta}$ by $\alpha + i\beta$, where

$$(\alpha^2 + \beta^2)^{1/2} = |\rho'|.$$

Then

$$x + iy = \frac{\rho_f^+ + \alpha + i\beta}{1 + \rho_f^+ (\alpha + i\beta)}.$$

Multiplying both sides by the denominator of the right-hand side and then equating real and imaginary parts of the resulting expressions yields

$$\begin{aligned} x(1 + \rho_f^+ \alpha) - y\rho_f^+ \beta &= \rho_f^+ + \alpha \\ y(1 + \rho_f^+ \alpha) + x\rho_f^+ \beta &= \beta, \end{aligned}$$

i.e.

$$\begin{aligned} (x - \rho_f^+) &= \alpha x (1 - x\rho_f^+) + \beta y \rho_f^+ \\ y &= -\alpha y \rho_f^+ + \beta (1 - x\rho_f^+). \end{aligned}$$

To find the locus, we square and add these equations to give

$$\begin{aligned} (x - \rho_f^+)^2 + y^2 &= (\alpha^2 + \beta^2) \left[(1 - x\rho_f^+)^2 + (\rho_f^+ y)^2 \right] \\ &= |\rho'|^2 \left[(1 - x\rho_f^+)^2 + (\rho_f^+ y)^2 \right] \end{aligned}$$

which can be manipulated to

$$x^2 \left(1 - |\rho'|^2 \rho_f^{+2} \right) + y^2 \left(1 - |\rho'|^2 \rho_f^{+2} \right) - 2x\rho_f^+ \left(1 - |\rho'|^2 \right) + \rho_f^{+2} - |\rho'|^2 = 0. \quad (2.160)$$

This is the equation of a circle with centre

$$\left(\frac{\rho_f^+ (1 - |\rho'|^2)}{(1 - |\rho'|^2 \rho_f^{+2})}, 0 \right),$$

i.e. on the real axis, and radius

$$\frac{|\rho'| \left(1 - \rho_f^{+2} \right)}{\left(1 - |\rho'|^2 \rho_f^{+2} \right)}.$$

The locus of the reflection coefficient, as the layer thickness is allowed to increase steadily from zero, is therefore a circle. A half-wave layer traces out a complete circle, while a quarter-wave layer, if it starts on the real axis, will trace out a semicircle; otherwise it will be slightly more or less than a semicircle, depending on the exact starting point. In all cases, the circle is traced clockwise.

The locus corresponding to a single layer is straightforward. The plotting of the locus corresponding to two or more layers is slightly more complicated. The form of the locus of each layer is an arc of a circle traced from the terminal point of the previous layer. The complication arises from the subsidiary calculation which must be performed each time to calculate the current value of ρ' from the terminal value of the previous layer. An example will serve to illustrate the point.

Let us consider a glass substrate of index 1.52, on which is deposited first a layer of zinc sulphide of index 2.35 and thickness of one quarter-wave, followed by a layer of cryolite of index 1.35 and of thickness also one quarter-wave. Air, of index 1.0, is the incident medium.

Calculation of the circles is most easily performed by using equation (2.159) to calculate the terminal points. The starting point is known and that, together with the fact that the centre is on the real axis, completes the specification of the circles.

The values of ρ_f^+ and ρ' for the first layer are

$$\rho_f^+ = \frac{1.0 - 2.35}{1.0 + 2.35} = -0.4030$$

$$\rho' = \frac{2.35 - 1.52}{2.35 + 1.52} = 0.2144.$$

The starting point for the layer is

$$\rho = \frac{\rho_f^+ + \rho'}{1 + \rho_f^+ \rho'} = -0.2063$$

which corresponds to the amplitude reflection coefficient of bare glass in air.

For a quarter-wave layer $e^{-2i\delta} = -1$ and so the terminal value of ρ is given by

$$\rho = \frac{\rho_f^+ - \rho'}{1 - \rho_f^+ \rho'} = -0.5683$$

and the locus up to this point is a semicircle. This value of ρ corresponds to the amplitude reflection coefficient of a quarter-wave of zinc sulphide on glass in air. To continue the locus into the next layer, we need new values of ρ_f^+ and ρ' .

$(\rho_f^+)_\text{new}$ is straightforward, being the external reflection coefficient at an air–cryolite boundary:

$$(\rho_f^+)_\text{new} = \frac{1.0 - 1.35}{1.0 + 1.35} = -0.1489.$$

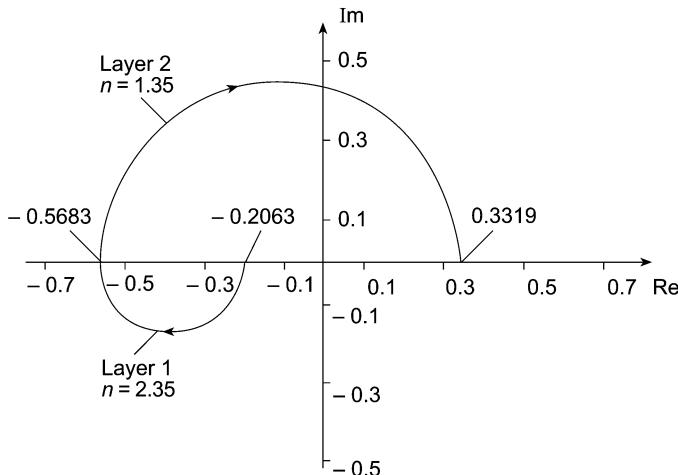


Figure 2.23. Reflection circles, or amplitude reflection locus, for the coating: Air| LH |Glass, where L indicates a quarter-wave of index 1.35, H of 2.35, and the indices of air and glass are 1.00 and 1.52, respectively.

$(\rho')_{\text{new}}$ is more difficult. This is the amplitude reflection coefficient which the substrate plus a quarter-wave of zinc sulphide will have, no longer in a medium of air, but in one of cryolite. It can be calculated either using the normal matrix method or simply by inverting the equation

$$\rho = (\rho)_{\text{old}} = \frac{(\rho_f^+)_{{\text{new}}} + (\rho')_{{\text{new}}}}{1 + (\rho_f^+)_{{\text{new}}} (\rho')_{{\text{new}}}}$$

which must be satisfied if the start of the new layer is to coincide with $(\rho)_{\text{old}}$, the termination of the old.

$$(\rho')_{{\text{new}}} = \frac{(\rho)_{\text{old}} - (\rho_f^+)_{{\text{new}}}}{1 - (\rho)_{\text{old}} (\rho_f^+)_{{\text{new}}}}$$

and in this case $(\rho)_{\text{old}}$ is -0.5683 , so that

$$(\rho')_{{\text{new}}} = \frac{-0.5683 - (-0.1489)}{1 - (-0.5683) (-0.1489)} = -0.4582.$$

The new locus, which is another semicircle, then starts at the point -0.5683 on the real axis and terminates at

$$\rho = \frac{(\rho_f^+)_{{\text{new}}} - (\rho')_{{\text{new}}}}{1 - (\rho_f^+)_{{\text{new}}} (\rho')_{{\text{new}}}} = 0.3319.$$

The loci are shown in figure 2.23.

The advantage of the technique over the Smith chart is especially that the locus is a continuous one, since the termination of each layer is the starting point for the next. All possible loci corresponding to a particular refractive index form a set of nested circles centred on the real axis of the diagram. Enough of these circles can be drawn to form a separate template or overlay for each of the materials involved in a design and these can considerably ease the task of drawing the diagram.

Since the method of the Smith chart is based on the real and imaginary axes of the amplitude reflection coefficient, the loci can actually be drawn on the same diagram as a Smith chart. Strictly, in that case, the chart should not be referred to as a Smith chart because it is not being used in that way.

Many examples of the use of this technique in design are given by Apfel [11] who has also extended it to include absorbing layers such as metals.

References

- [1] Yeh P 1988 *Optical Waves in Layered Media* (New York: Wiley)
- [2] Hodgkinson I J and Wu Q h 1997 *Birefringent Thin Films and Polarizing Elements* 1st edn (Singapore: World Scientific)
- [3] Born M and Wolf E 1975 *Principles of Optics* 5th edn (Oxford: Pergamon)
- [4] Berning P H 1963 Theory and calculations of optical thin films *Physics of Thin Films* ed G Hass (New York: Academic) pp 69–121
- [5] Macleod H A 1995 Antireflection coatings on absorbing substrates *38th Annual Technical Conference Chicago (Society of Vacuum Coaters)* pp 172–5
- [6] Berning P H and Turner A F 1957 Induced transmission in absorbing films applied to band pass filter design *J. Opt. Soc. Am.* **47** 230–9
- [7] Abelès F 1950 Recherches sur la propagation des ondes électromagnétiques sinusoïdales dans les milieux stratifiés. Applications aux couches minces *Ann. Phys., Paris, 12ième Serie* **5** 596–640
- [8] Abelès F 1950 Recherches sur la propagation des ondes électromagnétiques sinusoïdales dans les milieux stratifiés. Applications aux couches minces *Ann. Phys., Paris, 12ième Serie* **5** 706–84
- [9] Baumeister P, Hahn R and Harrison D 1972 The radiant transmittance of tandem arrays of filters *Opt. Acta* **19** 853–64
- [10] Smith S D 1958 Design of multilayer filters by considering two effective interfaces *J. Opt. Soc. Am.* **48** 43–50
- [11] Apfel J H 1972 Graphics in optical coating design *Appl. Opt.* **11** 1303–12

Chapter 3

Antireflection coatings

As has already been mentioned in chapter 1, antireflection coatings were the principal objective of much of the early work in thin-film optics. Of all the possible applications, antireflection coatings have had the greatest impact on technical optics, and even today, in sheer volume of production, they still exceed all other types of coating. In some applications, antireflection coatings are simply required for the reduction of surface reflection. In others, not only must surface reflection be reduced, but the transmittance must also be increased. The crown glass elements in a compound lens have a transmittance of only 96% per untreated surface, while the flint components can have a surface transmittance as low as 90%. The net transmittance of even a modest number of untreated elements in series can therefore be quite low. Additionally, part of the light reflected at the various surfaces eventually reaches the focal plane, where it appears as ghosts or as a veiling glare, thus reducing the contrast of the images. This is especially true of the zoom lenses used in television or photography, where 20 or more elements may be included, and which would be completely unusable without antireflection coatings.

Antireflection coatings can range from a simple single layer having virtually zero reflectance at just one wavelength, to a multilayer system of more than a dozen layers, having virtually zero reflectance over a range of several octaves. The type used in any particular application will depend on a variety of factors, including the substrate material, the wavelength region, the required performance and the cost.

In the visible region, crown glass, which has a refractive index of around 1.52, is most commonly used. As we shall see, this presents a very different problem from infrared materials, which can have very much higher refractive indices. It is convenient, therefore, to split what follows into antireflection coatings for low-index substrates and antireflection coatings for high-index substrates, corresponding roughly to the visible and infrared. Since, from the point of view of design, antireflection coatings for high-index substrates are more straightforward, they are considered first.

There is no systematic method for the design of antireflection coatings. Trial and error, assisted by approximate techniques (frequently one or other of the graphical methods mentioned in chapter 2) backed up by accurate computer calculation, are frequently employed. Very promising designs can be further improved by computer refinement. Several different approaches are used in this chapter, partly to illustrate their use and partly because they are complementary. All the performance curves have been computed by application of the matrix method. In most cases, the materials are considered to be completely transparent.

The vast majority of antireflection coatings are required for matching an optical element into air. Air has an index of around 1.0003 at standard temperature and pressure which, for practical purposes, can be considered as unity. The earliest antireflection coatings were on glass for use in the visible region of the spectrum. As shall become apparent later, a single-layer antireflection coating on glass, for the centre of the visible region, has a distinct magenta tinge when examined visually in reflection. This gives an appearance not unlike tarnish, and indeed in chapter 1 we mentioned the beneficial effects of the tarnish layer on aged flint objectives, and so the term ‘bloom’, in the sense of tarnish, has been used in this connection. The action of applying the coating is referred to as ‘blooming’ and the element is said to be ‘bloomed’.

3.1 Antireflection coatings on high-index substrates

The term high index in this context cannot be defined precisely in the sense of a range with a definite lower bound. It simply means that the substrate has an index sufficiently higher than the available thin-film materials to enable the design of high-performance antireflection coatings consisting entirely, or almost entirely, of layers with indices lower than that of the substrate. These high-index substrates are principally of use in the infrared. Semiconductors, such as germanium, with an index of around 4.0, giving a reflection loss of around 36% per surface, and silicon, with an index around 3.5 and reflection loss of 31%, are common, and it would be completely impossible to use them in the vast majority of applications without some form of antireflection coating. For many purposes, the reduction of a 30% reflection loss to one of a few per cent would be considered adequate. It is only in a limited number of applications where the reflection loss must be reduced to less than 1%.

3.1.1 The single-layer antireflection coating

The simplest form of antireflection coating is a single layer. Consider figure 3.1. Here we have a vector diagram which, since two interfaces are involved, contains two vectors, each representing the amplitude reflection coefficient at an interface.

If the incident medium is air, then, provided the index of the film is lower than the index of the substrate, the reflection coefficient at each interface will be negative, denoting a phase change of 180° . The resultant locus is a circle with a

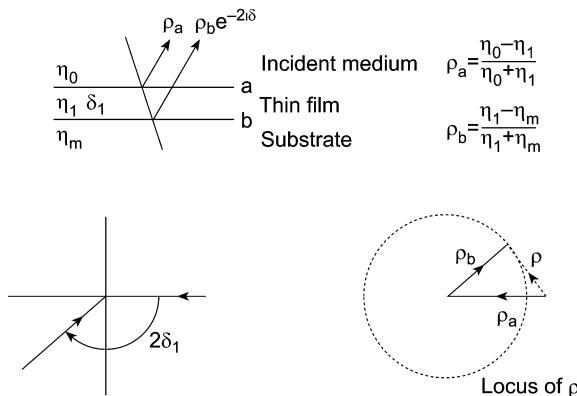


Figure 3.1. Vector diagram of a single-layer antireflection coating.

minimum at the wavelength for which the phase thickness of the layer is 90° , that is, a quarter-wave optical thickness, when the two vectors are completely opposed. Complete cancellation at this wavelength, that is, zero reflectance, will occur if the vectors are of equal length. This condition, in the notation of figure 3.1, is

$$\frac{y_0 - y_1}{y_0 + y_1} = \frac{y_1 - y_m}{y_1 + y_m}$$

which requires

$$\frac{y_1}{y_0} = \frac{y_m}{y_1}$$

or

$$y_1 = (y_0 y_m)^{1/2}, \quad (3.1)$$

which at optical frequencies can also be written

$$n_1 = (n_0 n_m)^{1/2}.$$

At oblique incidence, the admittances, y , in (3.1) should be replaced by the appropriate tilted values, η .

Although this result was derived by an approximate technique, the result is exactly correct. We recall that in chapter 2 it was shown that the optical admittance of a substrate coated with a quarter-wave optical thickness is

$$Y = y_1^2/y_m,$$

where y_1 is the admittance of the film material and y_m that of the substrate. The reflectance is therefore given by

$$R = \left(\frac{y_0 - Y}{y_0 + Y} \right)^2 = \left(\frac{y_0 - y_1^2/y_m}{y_0 + y_1^2/y_m} \right)^2.$$

This is an exact result and clearly the reflectance is zero if y_1 is given by (3.1).

The condition for a perfect single-layer antireflection coating is, therefore, a quarter-wave optical thickness of material with optical admittance equal to the square root of the product of the admittances of substrate and medium. It is seldom possible to find a material of exactly the optical admittance which is required. If there is a small error, ε , in y_1 such that

$$y_1 = (1 + \varepsilon) (y_0 y_m)^{1/2}$$

then

$$R = \left(\frac{-2\varepsilon - \varepsilon^2}{2 + 2\varepsilon + \varepsilon^2} \right)^2 \approx \varepsilon^2$$

provided that ε is small. A 10% error in y_1 , therefore, leads to a residual reflectance of 1%.

Zinc sulphide has an index of around 2.2 at 2 μm and 2.15 at 15 μm . It has sufficient transparency for use as a quarter-wave antireflection coating over the range 0.4–25 μm . Germanium, silicon, gallium arsenide, indium arsenide and indium antimonide can all be treated satisfactorily by a single layer of zinc sulphide. The procedure to be followed for hard, rugged zinc sulphide films is described in a paper by Cox and Hass [1]. The substrate should be maintained at around 150 °C during coating and cleaned by a glow discharge immediately before coating. The transmittance of a germanium plate with a single-layer zinc sulphide antireflection coating is shown in figure 3.2.

Zinc sulphide, even deposited under the best conditions, can deteriorate after prolonged exposure to humid atmospheres. Somewhat harder and more robust coatings are produced with cerium oxide or silicon monoxide. Cerium oxide, when deposited at a substrate temperature of 200 °C or more, forms very hard and durable films of refractive index 2.2 at 2 μm . Unfortunately, in common with many other materials it displays a slight absorption band at 3 μm owing to adsorbed water vapour. Silicon monoxide does not show this water vapour band to the same degree, and so Cox and Hass have recommended this material as the most satisfactory for coating germanium and silicon in the near infrared. The index of silicon monoxide evaporated in a good vacuum at a high rate is around 1.9. The transmittance of a silicon plate coated on both sides with silicon monoxide is shown in figure 3.3.

So far, we have considered only normal incidence in the numerical calculations which we have made. At angles of incidence other than normal, the behaviour is similar, but the effective phase thickness of the layer is reduced as the incidence increases due to the cosine term in the phase thickness

$$\delta = (2\pi n d \cos \vartheta) / \lambda$$

and so the optimum wavelength is shorter. For the optical admittance we must use the appropriate η_p or η_s , and, as these are different, polarisation effects become

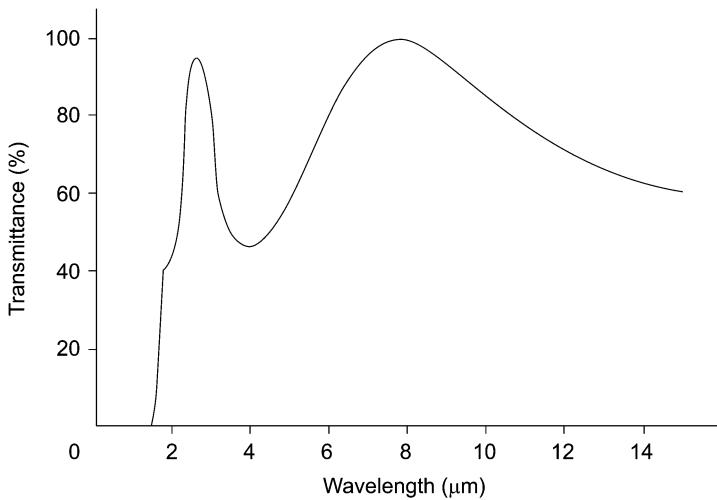


Figure 3.2. Transmittance of a germanium plate bloomed on both sides with zinc sulphide for 8 μm . (Courtesy of Sir Howard Grubb, Parsons & Co. Ltd.)

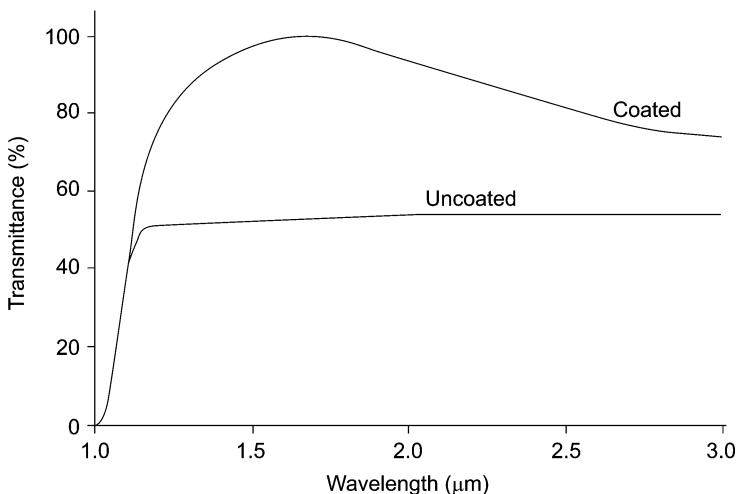


Figure 3.3. Transmittance of a 1.5-mm thick silicon plate with and without antireflection coatings of silicon monoxide, a quarter-wavelength thick at 1.7 μm . (After Cox and Hass [1].)

evident. For high-index substrates and coatings the effects are much less than for the low-index coatings for the visible region, as we shall see later. Figure 3.4 shows the calculated variation with angle of incidence of the performance of a

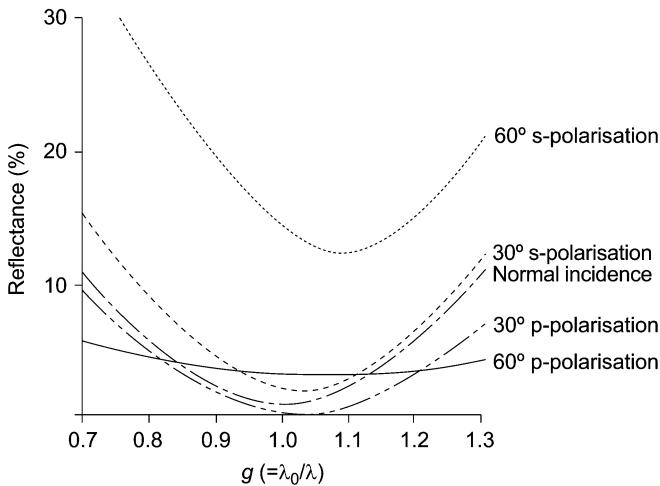


Figure 3.4. Calculated performance at various angles of incidence of a zinc sulphide coating ($n = 2.2$) on a germanium substrate ($n = 4.0$).

zinc sulphide coating ($n = 2.2$) on a germanium substrate ($n = 4.0$).

Such calculations are relatively straightforward. If we use the matrix method, the characteristic matrix of a single film on a substrate is given by

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_1 & \frac{i \sin \delta_1}{\eta_1} \\ i \eta_1 \sin \delta_1 & \cos \delta_1 \end{bmatrix} \begin{bmatrix} 1 \\ \eta_m \end{bmatrix},$$

i.e.

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_1 + i (\eta_m / \eta_1) \sin \delta_1 \\ \eta_m \cos \delta_1 + i \eta_1 \sin \delta_1 \end{bmatrix},$$

where the symbols have the meanings, defined in chapter 2,

$$\left. \begin{array}{l} \eta_p = y / \cos \vartheta \\ \eta_s = y \cos \vartheta \\ \delta_1 = (2\pi n_1 d_1 \cos \vartheta_1) / \lambda \end{array} \right\} \text{for each material}$$

and where

$$n_0 \sin \vartheta_0 = n_1 \sin \vartheta_1 = n_m \sin \vartheta_m.$$

If λ_0 is the wavelength for which the layer is a quarter-wave optical thickness at normal incidence, then $n_1 d_1 = \lambda_0 / 4$ and

$$\delta_1 = \frac{\pi}{2} \left(\frac{\lambda_0}{\lambda} \right) \cos \vartheta_1$$

so that the new optimum wavelength is $\lambda_0 \cos \vartheta_1$.

The amplitude reflection coefficient is

$$\begin{aligned}\rho &= \frac{\eta_0 - Y}{\eta_0 + Y} = \frac{\eta_0 - C/B}{\eta_0 + C/B} \\ &= \frac{(\eta_0 - \eta_m) \cos \delta_1 + i[(\eta_0 \eta_m / \eta_1) - \eta_1] \sin \delta_1}{(\eta_0 + \eta_m) \cos \delta_1 + i[(\eta_0 \eta_m / \eta_1) + \eta_1] \sin \delta_1}\end{aligned}\quad (3.2)$$

and the reflectance

$$R = \frac{(\eta_0 - \eta_m)^2 \cos^2 \delta_1 + i[(\eta_0 \eta_m / \eta_1) - \eta_1]^2 \sin^2 \delta_1}{(\eta_0 + \eta_m)^2 \cos^2 \delta_1 + i[(\eta_0 \eta_m / \eta_1) + \eta_1]^2 \sin^2 \delta_1}. \quad (3.3)$$

This expression is deceptively simple. An increase in the number of layers or a move to an absorbing system immediately increases the complexity to a degree that is completely discouraging.

It is instructive to prepare an admittance diagram (figure 3.5) for the single-layer coating. We recall that admittance loci were discussed in chapter 2. We consider normal incidence only and use free space units for the admittances so that they are numerically equal to the refractive indices. The locus for a single layer is a circle and in this case it begins at the point 4.0 on the real axis, corresponding to the admittance of the germanium substrate. The centre of the circle is on the real axis and the circle cuts the real axis again at the point $2.2^2/4.0 = 1.21$, corresponding to a quarter-wave optical thickness. Note especially that since the two points of intersection with the real axis are defined we do not need to calculate the position of the centre. We can mark a scale of δ_1 along the locus. Since $\delta_1 = 2\pi n_1 d_1 / \lambda$, we can either assume λ constant and replace the scale with one of optical thickness, or, provided that we assume that the refractive index remains constant with wavelength, for a given layer optical thickness we can mark the scale in terms of g ($= \lambda_0 / \lambda$). These various scales have been added. The scale of g assumes that λ_0 is the wavelength for which the layer has an optical thickness of one quarter-wave.

This is a particularly simple admittance locus and it is included principally to illustrate the method. We will make some use of admittance diagrams in this chapter. Normally these will be drawn for one value of wavelength and for one value of optical thickness for each layer.

3.1.2 Double-layer antireflection coatings

The disadvantage of the single-layer coating, as far as the design is concerned, is the limited number of adjustable parameters. We can see from the admittance locus of figure 3.5 that only where the locus passes through the point $(1, 0)$ will zero reflectance be obtained (or more generally when the locus passes through the point $(y_0, 0)$) and this must correspond to a semicircle or a quarter-wave optical thickness (or, strictly, an odd integral multiple thereof). The refractive index, or

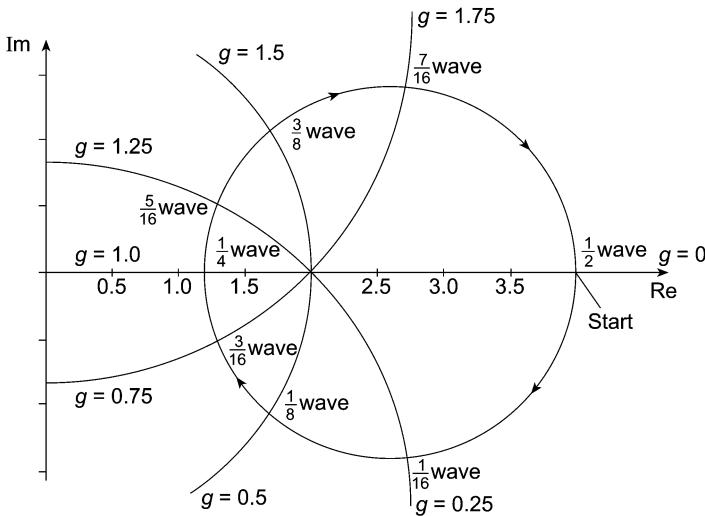


Figure 3.5. Admittance diagram for a single-layer zinc sulphide ($n = 2.2$) coating on germanium ($n = 4.0$).

optical admittance, of the layer is also uniquely determined as $y_1 = (y_0 y_m)^{1/2}$. There is thus no room for manoeuvre in the design of a single-layer coating. In practice, the refractive index is not a parameter that can be varied at will. Materials suitable for use as thin films are limited in number and the designer has to use what is available. A more rewarding approach, therefore, is to use more layers, specifying obtainable refractive indices for all layers at the start, and to achieve zero reflectance by varying the thickness. Then, too, there is the limitation that the single-layer coating can give zero reflectance at one wavelength only and low reflectance over a narrow region. A wider region of high performance demands additional layers.

Much of this design work nowadays is carried out by automatic methods and this is a perfectly sensible and efficient development. Automatic methods are briefly described elsewhere in this book. They are particularly valuable for antireflection coatings and are strongly recommended. Here, however, we are concerned also with the understanding of the structures of the coatings and particularly with the parts played by the individual layers. Without such understanding we are completely vulnerable when things go wrong and the results are not as expected. Also, automatic design techniques function more efficiently when they are furnished with good starting designs. We therefore spend much time in this chapter with some of the traditional design techniques, not so much because all are still used in actual design work, but because they require a knowledge of the structure and working of the coatings, and because they are interesting.

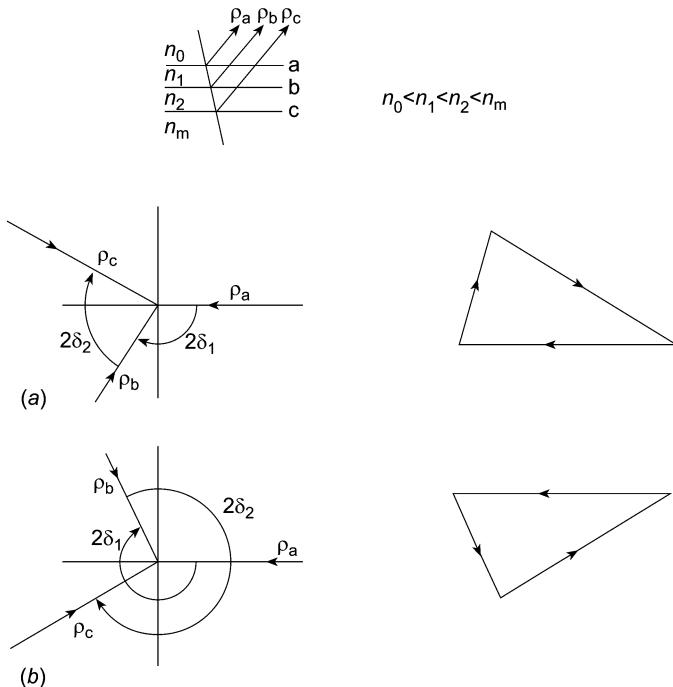


Figure 3.6. Vector diagram for a double-layer antireflection coating. The thickness of the layers can be chosen to close the vector triangle and give zero reflectance in two ways, (a) and (b).

We will consider first the problem of ensuring zero reflectance at one single wavelength and we shall attempt to achieve this with a two-layer coating. Since we are dealing with high-index substrates we look initially at combinations of layers having refractive indices lower than that of the substrate. A vector diagram of one possibility is shown in figure 3.6. Provided the vectors are not such that any one is greater in length than the sum of the other two, then there are two sets of thicknesses for which zero reflectance can be obtained at one wavelength. The thinner combination, as in figure 3.6(a), will give the broadest characteristic and should normally be chosen. In some ways, it is easier to visualise the design using an admittance plot. As usual, we plot admittance in free space units so that it is numerically the same as the refractive index. Two possible arrangements are shown in figure 3.7, which can be obtained simply by drawing the circle corresponding to index n_1 , passing through the point n_0 , and the circle corresponding to index n_2 passing through the point n_m . Provided these circles intersect, then it is possible to use them as an antireflection coating. The two sets of thicknesses correspond to the two points of intersection.

This is a very important coating with wider implications than just the

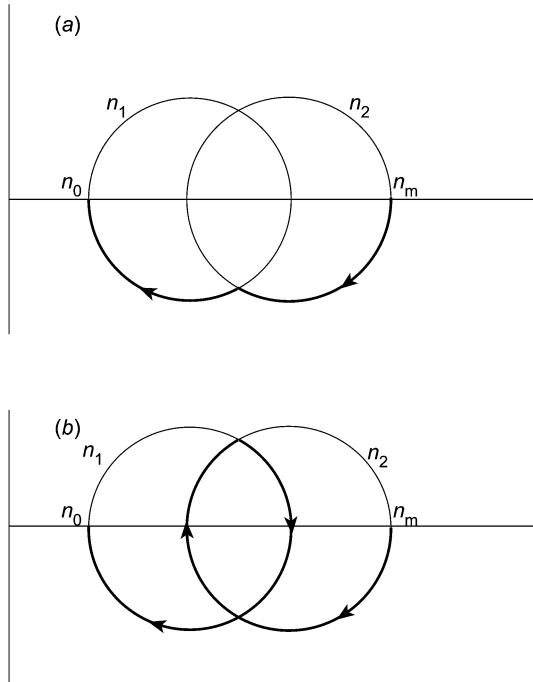


Figure 3.7. Admittance diagram for the double-layer antireflection coating. The two possible solutions are shown in (a) and (b).

blooming of a high-index substrate and so it is worth examining in greater detail. We use the matrix method and follow an analysis by Catalan [2], changing the notation to agree with the system used here. The characteristic matrix of the assembly is

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_1 & \frac{i \sin \delta_1}{y_1} \\ iy_1 \sin \delta_1 & \cos \delta_1 \end{bmatrix} \begin{bmatrix} \cos \delta_2 & \frac{i \sin \delta_2}{y_2} \\ iy_2 \sin \delta_2 & \cos \delta_2 \end{bmatrix} \begin{bmatrix} 1 \\ y_m \end{bmatrix}$$

$$= \begin{bmatrix} \cos \delta_1 [\cos \delta_2 + i(y_m/y_2) \sin \delta_2] + i \sin \delta_1 (y_m \cos \delta_2 + iy_2 \sin \delta_2) / y_1 \\ iy_1 \sin \delta_1 [\cos \delta_2 + i(y_m/y_2) \sin \delta_2] + \cos \delta_1 (y_m \cos \delta_2 + iy_2 \sin \delta_2) \end{bmatrix}.$$

The reflectance will be zero if the optical admittance Y is equal to y_0 , i.e.

$$iy_1 \sin \delta_1 [\cos \delta_2 + i(y_m/y_2) \sin \delta_2] + \cos \delta_1 (y_m \cos \delta_2 + iy_2 \sin \delta_2)$$

$$= y_0 \{ \cos \delta_1 [\cos \delta_2 + i(y_m/y_2) \sin \delta_2] + i \sin \delta_1 (y_m \cos \delta_2 + iy_2 \sin \delta_2) / y_1 \}.$$

The real and imaginary parts of these expressions must be equated separately giving

$$-(y_1 y_m / y_2) \sin \delta_1 \sin \delta_2 + y_m \cos \delta_1 \cos \delta_2$$

$$= y_0 \cos \delta_1 \cos \delta_2 - (y_0 y_2 / y_1) \sin \delta_1 \sin \delta_2$$

and

$$\begin{aligned} y_1 \sin \delta_1 \cos \delta_2 + y_2 \cos \delta_1 \sin \delta_2 \\ = (y_0 y_m / y_2) \cos \delta_1 \sin \delta_2 + (y_0 y_m / y_1) \sin \delta_1 \cos \delta_2 \end{aligned}$$

i.e.

$$\begin{aligned} \tan \delta_1 \tan \delta_2 &= (y_m - y_0)[(y_1 y_m / y_2) - (y_0 y_2 / y_1)] \\ &= y_1 y_2 (y_m - y_0)(y_1^2 y_m - y_0 y_2^2) \end{aligned} \quad (3.4)$$

and

$$\tan \delta_2 / \tan \delta_1 = y_2 (y_0 y_m - y_1^2) / [y_1 (y_2^2 - y_0 y_m)] \quad (3.5)$$

giving

$$\tan^2 \delta_1 = \frac{(y_m - y_0)(y_2^2 - y_0 y_m)y_1^2}{(y_1^2 y_m - y_0 y_2^2)(y_0 y_m - y_1^2)} \quad (3.6)$$

$$\tan^2 \delta_2 = \frac{(y_m - y_0)(y_0 y_m - y_1^2)y_2^2}{(y_1^2 y_m - y_0 y_2^2)(y_2^2 - y_0 y_m)}.$$

The values of δ_1 and δ_2 found from these equations must be correctly paired and this is most easily done either by ensuring that they also satisfy the two preceding equations or by sketching a rough admittance diagram.

For solutions to exist, or, putting it in another way, for the circles in the admittance diagram to intersect, the right-hand sides of equations (3.6) must be positive. δ_1 and δ_2 are then real. This requires that, of the expressions

$$(y_2^2 - y_0 y_m) \quad (3.7)$$

$$(y_1^2 y_m - y_0 y_2^2) \quad (3.8)$$

$$(y_0 y_m - y_1^2) \quad (3.9)$$

either all three must be positive or any two are negative and the third positive. This can be summarised in a useful diagram (figure 3.8) known as a Schuster diagram after one of the originators [3]. The bottom right-hand part of the diagram corresponds to the validity conditions given in figure 3.7.

One useful coating is given by the area at the top left-hand edge of the diagram where $y_1 \geq (y_0 y_m)^{1/2} \geq y_2$. For germanium at normal incidence in air, $(y_0 y_m)^{1/2} = 2.0$. There is no upper limit to the magnitude of y_1 , which can be conveniently chosen to be germanium with index 4.0, while y_2 can be magnesium fluoride with index 1.38, didymium fluoride with index 1.57, cerium fluoride with index 1.59, or any other similar material. The advantage of this arrangement is that the low-index film, which tends to be less robust, is protected by the high-index layer. Germanium layers are particularly good in this respect. Figure 3.9 gives an example of this type of coating. Generally, the total thickness, as in the example, is rather thinner than a quarter-wave, which adds to the durability. Cox [4] has discussed a number of different possibilities along these lines.

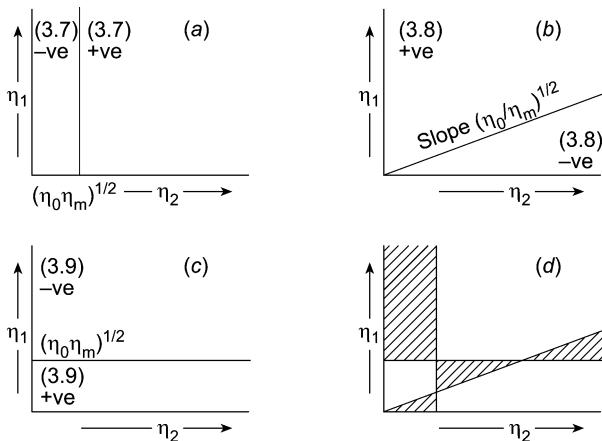


Figure 3.8. The construction of a Schuster diagram. (a), (b) and (c) are combined in one diagram in (d) and the shaded areas are those in which real solutions exist.

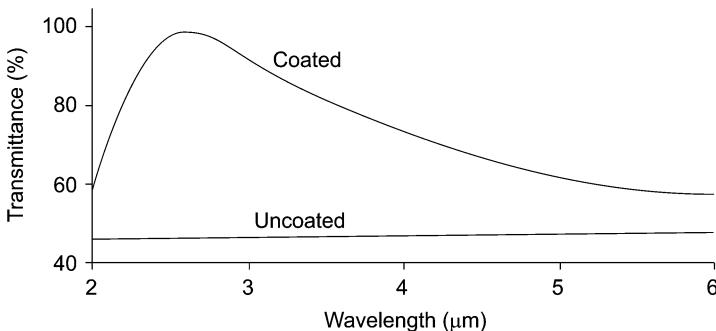


Figure 3.9. Transmittance of a germanium plate with two-layer antireflection coatings of MgF_2 , ($nd = \lambda/4$ at $1.03\ \mu\text{m}$) and germanium ($nd = \lambda/4$ at $0.61\ \mu\text{m}$), the germanium being the outermost layer. (After Cox [4].)

Unfortunately, this type of double-layer coating tends to have rather narrower useful ranges than the single-layer coating, which may itself not be broad enough for certain applications. It is possible to broaden the region of reflectance by using two, or even more, layers. A common approach is to choose layer thicknesses which are whole numbers of quarter-waves, and then to determine the refractive indices which should be used to give the desired performance.

An effective coating is one consisting of two quarter-wave layers (see figure 3.10). The appearances of the vector diagram at three different wavelengths is shown in (a), (b) and (c). At $\lambda = (3/4)\lambda_0$ and $\lambda = (3/2)\lambda_0$ the three vectors in the triangle are inclined at 60° to each other. Provided the vectors are all of

equal length, the triangles will be closed and the reflectance will be zero at these wavelengths. This condition can be written

$$\frac{y_1}{y_0} = \frac{y_2}{y_1} = \frac{y_m}{y_2}$$

and solved for y_1 and y_2 :

$$\begin{aligned} y_1^3 &= y_0^2 y_m \\ y_2^3 &= y_0 y_m^2. \end{aligned} \quad (3.10)$$

The reflectance at the reference wavelength λ_0 where the layers are quarter-waves is given by

$$\begin{aligned} R &= \left(\frac{y_0 - (y_1^2/y_2^2) y_m}{y_0 + (y_1^2/y_2^2) y_m} \right)^2 \\ &= \left(\frac{1 - (y_m/y_0)^{1/3}}{1 + (y_m/y_0)^{1/3}} \right)^2, \end{aligned}$$

a considerable improvement over the bare substrate.

For germanium of refractive index 4.0 in air, at normal incidence, the values required for the indices are $n_1 = 1.59$ and $n_2 = 2.50$ and the reflectance at λ_0 is 5.6%. The theoretical curve of this coating is shown in figure 3.11(a). Theoretical and measured curves of a similar coating on arsenic trisulphide and triselenide are given in figure 3.11(b) and (c).

The coating just described is a special case of a general coating where the layers are of equal thickness. To compute the general conditions it is easiest to return to the analysis leading up to equations (3.6).

Let δ_1 be set equal to δ_2 and denoted by δ , where we recall that if λ_0 is the wavelength for which the layers are quarter-waves then

$$\delta = \frac{\pi}{2} \left(\frac{\lambda_0}{\lambda} \right).$$

From equation (3.5)

$$y_2(y_0 y_m - y_1^2) = y_1(y_2^2 - y_0 y_m),$$

i.e.

$$y_0 y_m = y_1 y_2$$

which is a necessary condition for zero reflectance.

From equation (3.4) we find the wavelengths λ corresponding to zero reflectance

$$\tan^2 \delta = \frac{y_1 y_2 (y_m - y_0)}{y_1^2 y_m - y_0 y_2^2} = \frac{y_0 y_m (y_m - y_0)}{y_1^2 y_m - y_0 y_2^2}.$$

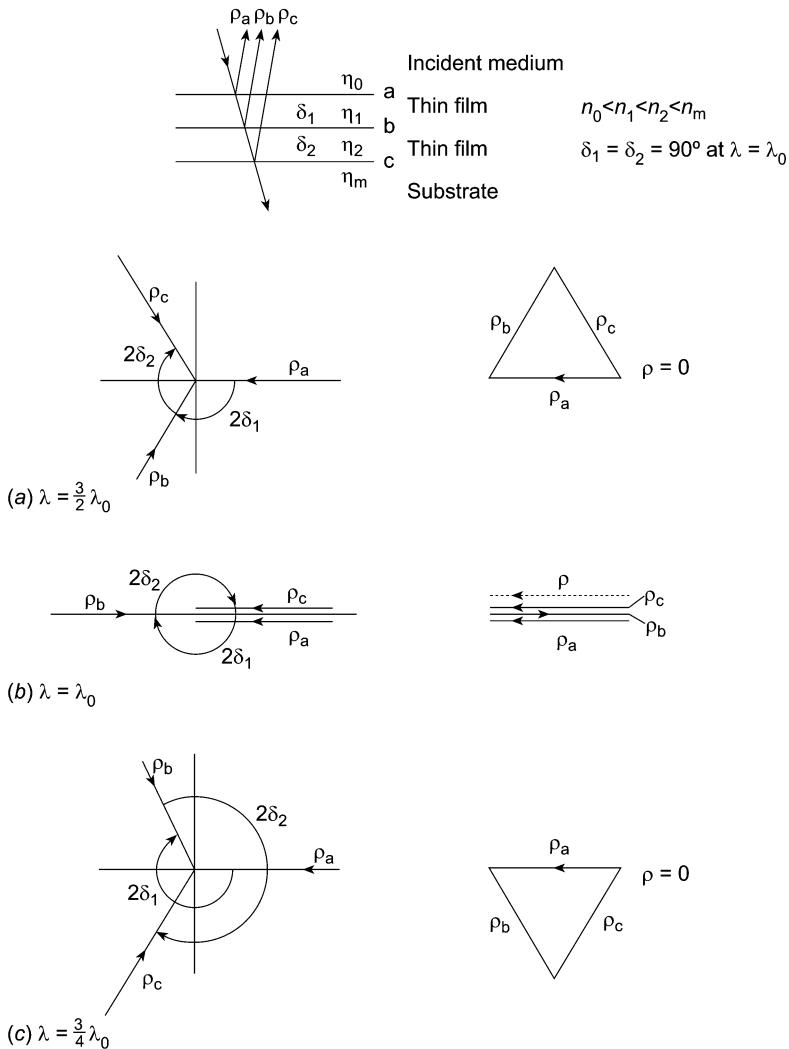


Figure 3.10. Vector diagrams for quarter-quarter antireflection coatings on a high-index substrate.

If δ is the solution in the first quadrant then there are two solutions

$$\delta = \delta' \quad \text{or} \quad \delta = \pi - \delta'$$

and the two values of λ are

$$\lambda = \left(\frac{\pi/2}{\delta} \right) \lambda_0.$$

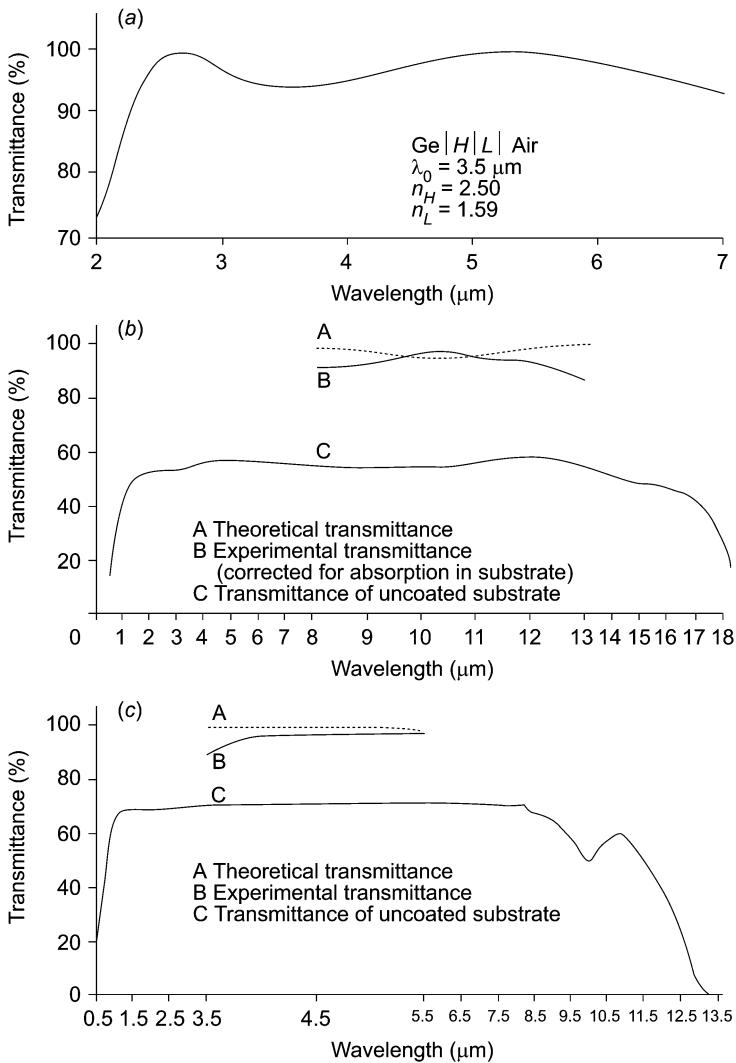


Figure 3.11. Double-layer antireflection coatings for high-index substrates. (a) Theoretical transmittance of a quarter-quarter coating on germanium (single surface). (b) Theoretical and measured transmittance of a similar coating on arsenic trisulphide glass (double surface). (c) Theoretical and measured transmittance of a similar coating on arsenic triselenide glass (double surface). ((b) and (c) by courtesy of Barr and Stroud Ltd.)

In all practical cases, y_m will be greater than y_0 and the above equation for $\tan^2 \delta$ will have a real solution provided

$$y_1^2 y_m - y_0 y_2^2 \geq 0.$$

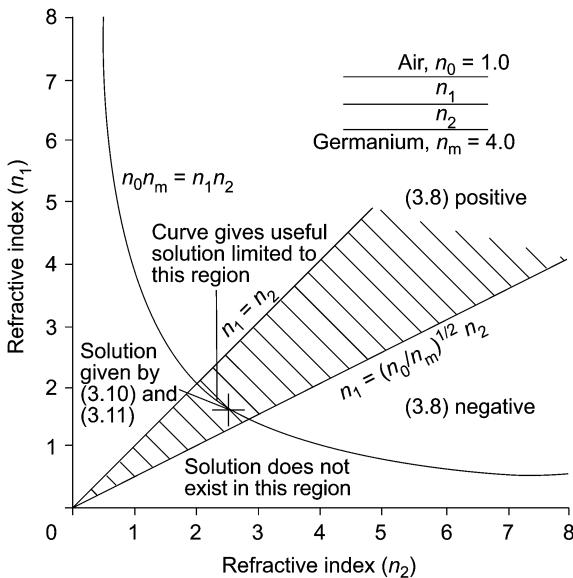


Figure 3.12. A Schuster diagram showing possible values of film indices for a quarter-quarter coating on germanium.

The left-hand side of this inequality is identical to expression (3.8).

Figure 3.12 gives the allowed values of y_1 and y_2 for germanium in air plotted on a Schuster diagram assuming normal incidence. The form of the characteristic curve of the coating is similar to that of figure 3.11. The reflectance rises to a maximum value at the reference wavelength λ_0 situated between the two zeros. The reflectance at λ_0 can be found quite simply. At this wavelength, $\delta = \pi/2$ and the layers are quarter-waves. The optical admittance is given, therefore, by

$$\frac{y_1^2}{y_2^2} y_m$$

and the reflectance by

$$R = \left(\frac{y_0 - (y_1^2/y_2^2) y_m}{y_0 + (y_1^2/y_2^2) y_m} \right)^2. \quad (3.11)$$

We are considering cases where y_m is large. For $y_1 = y_2$, the reflectance at λ_0 is that of the bare substrate. If $y_1 > y_2$ the reflectance is even higher. Thus, for the solution to be at all useful, y_1 should be less than y_2 and the region where this condition holds is indicated on the diagram.

3.1.3 Multilayer coatings

Figure 3.13 shows a vector diagram for a three-layer coating on germanium. Each layer is a quarter-wave thick at λ_0 . If $y_m > y_3 > y_2 > y_1 > y_0$ then the vectors will oppose each other, as shown, at $(2/3)\lambda_0$, λ_0 and $2\lambda_0$, and, provided the vectors are all of equal length, will completely cancel at these wavelengths, giving zero reflectance.

This coating is similar to the quarter-quarter coating of figure 3.10, but where the two zeros of the two-layer coating are situated at $(3/4)\lambda_0$ and $(3/2)\lambda_0$, those of this three-layer coating stretch from $(2/3)\lambda_0$ to $2\lambda_0$, a much broader region.

The condition for the vectors to be of equal length is

$$\frac{y_1}{y_0} = \frac{y_2}{y_1} = \frac{y_3}{y_2} = \frac{y_m}{y_3}$$

which with some manipulation becomes

$$\begin{aligned} y_1^4 &= y_0^3 y_m \\ y_2^4 &= y_0^2 y_m^2 \\ y_3^4 &= y_0 y_m^3. \end{aligned} \tag{3.12}$$

For germanium in air at normal incidence

$$n_0 = 1.00 \quad n_m = 4.00$$

and the refractive indices required for the layers are

$$\begin{aligned} n_1 &= 1.41 \\ n_2 &= 2.00 \\ n_3 &= 2.83. \end{aligned}$$

A coating which is not far removed from these theoretical figures is silicon, next to the substrate, of index 3.3, followed by cerium oxide of index 2.2, followed by magnesium fluoride, index 1.35. The performance of such a coating with $\lambda_0 = 3.5 \mu\text{m}$ is shown in figure 3.14. This coating, along with other one- and two-layer coatings for the infrared, is described by Cox *et al* [5]. The exact theory of this coating may be developed in the same way as that of the two-layer coating, but the calculations are more involved.

It is relatively easy to extend the vector method to deal with four layers, where the zeros of reflectance are found at $(5/8)\lambda_0$, $(5/6)\lambda_0$, $(5/4)\lambda_0$ and $(5/2)\lambda_0$, an even broader region than the three-layer coating. Five layers are equally straightforward. Whether or not such coatings are of practical value depends very much on the application. For many purposes the two-layer coating is quite adequate.

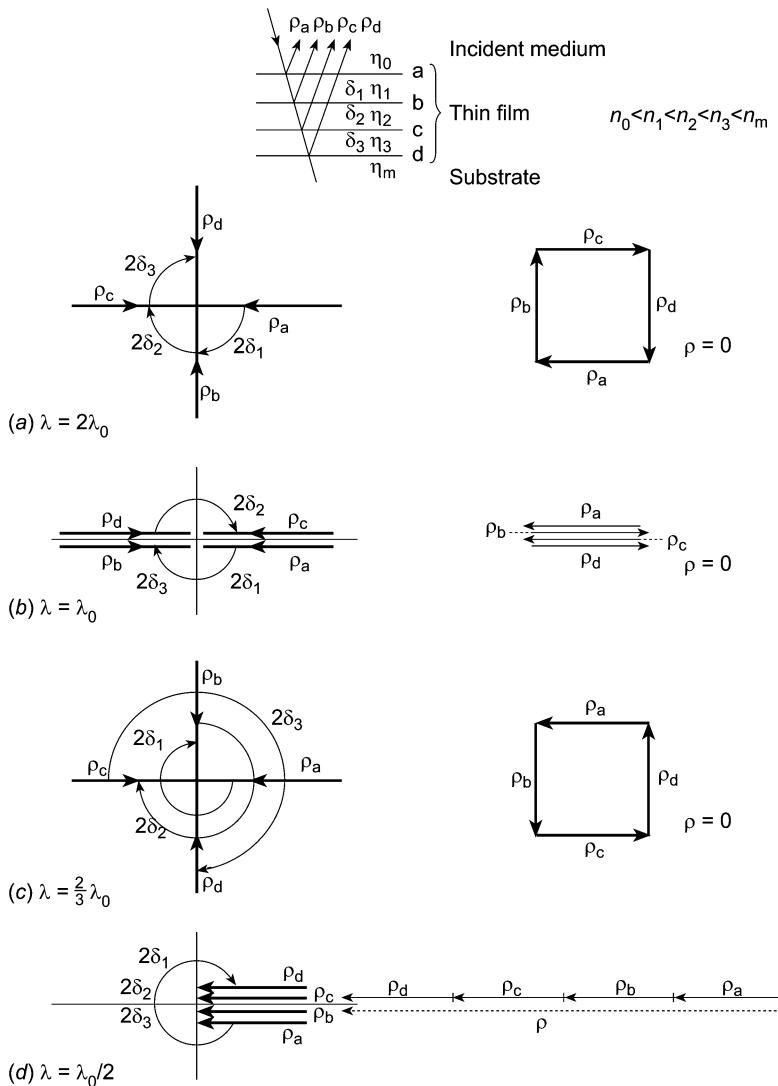


Figure 3.13. Vector diagram for a quarter-quarter-quarter coating on a high-index substrate.

The addition of an extra layer makes the exact theory of the three-layer coating very much more involved than that of the two-layer. The number of possible groups of designs is enormous. It therefore becomes profitable to employ techniques which, rather than calculate performance in detail, simply indicate arrangements which are likely to be capable of acceptable performance and

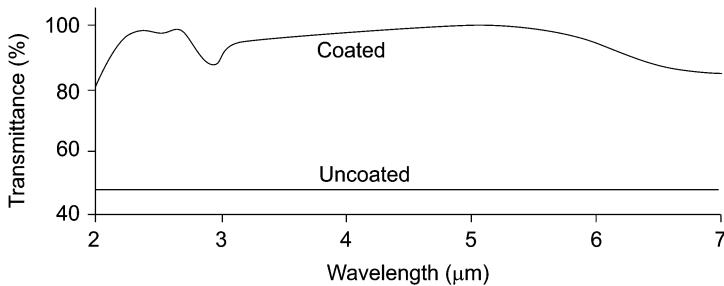


Figure 3.14. Measured transmittance of a germanium plate with coatings consisting of $\text{MgF}_2 + \text{CeO}_2 + \text{Si}$ ($n_1 d_1 = n_2 d_2 = n_3 d_3 = \lambda/4$ at $3.5 \mu\text{m}$). (After Cox *et al* [5].)

eliminate those which are not. Performance can then be accurately calculated by the procedures of chapter 2.

A particularly useful technique of this type has been developed by Musset and Thelen [6]. It is based on Smith's method, that is, the method of effective interfaces. We recall from chapter 2 that this involves the breaking down of the assembly into two subsystems. These we can label a and b. The overall transmittance of the multilayer is then given by

$$T = \left[\frac{T_a T_b}{\left(1 - R_a^{1/2} R_b^{1/2}\right)^2} \times \left[1 + \frac{4R_a^{1/2} R_b^{1/2}}{\left(1 - R_a^{1/2} R_b^{1/2}\right)^2} \sin^2\left(\frac{\varphi_a + \varphi_b - 2\delta}{2}\right) \right]^{-1} \right]. \quad (3.13)$$

We assume that there is no absorption, so that $T_a = 1 - R_a$ and $T_b = 1 - R_b$.

Both of the expressions multiplied together on the right-hand side of equation (3.13) have maximum possible values of unity, and for maximum transmittance, therefore, both must be separately maximised. The first expression

$$\frac{T_a T_b}{\left(1 - R_a^{1/2} R_b^{1/2}\right)^2}$$

will be unity if, and only if, $R_a = R_b$, while the second,

$$\left[1 + \frac{4R_a^{1/2} R_b^{1/2}}{\left(1 - R_a^{1/2} R_b^{1/2}\right)^2} \sin^2\left(\frac{\varphi_a + \varphi_b - 2\delta}{2}\right) \right]^{-1}$$

will be unity if, and only if,

$$\sin^2 \left(\frac{\varphi_a + \varphi_b - 2\delta}{2} \right) = 0.$$

The conditions for a perfect antireflection coating are then

$$R_a = R_b$$

called the amplitude condition by Musset and Thelen, and

$$\frac{\varphi_a + \varphi_b - 2\delta}{2} = m\pi$$

called the phase condition. The amplitude condition is a function of the two subsystems. The phase condition can be satisfied by adjusting the thickness of the spacer layer. The amplitude condition can, using a method devised by Musset and Thelen, be satisfied for all wavelengths, but it is difficult to satisfy the phase condition except at a limited number of discrete wavelengths. At other wavelengths the performance departs from ideal to a varying degree.

The transmittance and reflectance of a multilayer remain constant when the optical admittances are all multiplied by a constant factor or when they are all replaced by their reciprocals, in both cases keeping the optical thicknesses constant. These properties can readily be demonstrated from the structure of the characteristic matrices [7]. They enable the design of pairs of substructures having identical reflectance so that only the phase condition need be satisfied for perfect antireflection. We can, following Musset and Thelen, imagine a multilayer consisting of two subsections, a and b, as shown in figure 3.15, with a medium of admittance y_i in between. At this stage we put no restrictions on this medium in terms either of refractive index or thickness but, as we shall see, they will become defined at a later stage. Subsection a is bounded by y_m on one side and y_i on the other, while b is bounded in the same way by y_i and y_0 . We can now apply the appropriate rules for ensuring that the amplitude condition is satisfied. We set up any subsystem a and then convert it into subsystem b by retaining the optical thicknesses and either multiplying the admittances by a constant multiplier, or taking the reciprocals of the admittances and multiplying them by a constant multiplier. Systems derived by the former procedure are classified by Musset and Thelen as type I, those by the latter as type II.

For type I systems we must have

$$\begin{aligned} y_m f &= y_i \\ y_i f &= y_0 \end{aligned}$$

so that

$$y_i = (y_0 y_m)^{1/2}$$

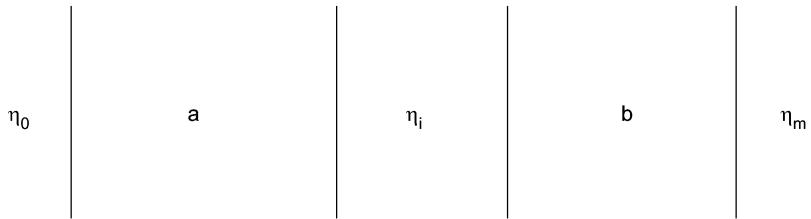


Figure 3.15. Multilayer antireflection coating consisting of two subsystems, a and b, separated by a central layer.

and

$$f = (y_0/y_m)^{1/2}.$$

In this way, any y_a gives a corresponding y_b of $y_a(y_0/y_m)^{1/2}$.

Type II systems, on the other hand, convert so that

$$\begin{aligned} f/y_m &= y_0 \\ f/y_i &= y_i \\ f/y_a &= y_b, \end{aligned}$$

i.e.

$$y_i = (y_0 y_m)^{1/2} \quad \text{and} \quad f = y_0 y_m$$

so that any y_a gives a corresponding y_b of $y_0 y_m / y_a$.

There are no restrictions on layer thickness or on the number of layers in each subsystem except that they must be equal in number, and it is simpler if quarter-wave layers are used. Once the individual subsystems a and b are established, the amplitude condition is automatically satisfied at all wavelengths and it remains to satisfy the phase condition. This involves the coupling arrangement. It is impossible to meet the phase condition at all wavelengths and the problem is so complex that it is best to take the easy way out and adopt a layer of admittance y_i with thickness zero, in which case the layer is omitted, or a quarter-wave, like the remaining layers of the assembly.

The method can be illustrated by application to the antireflection of germanium at normal incidence. In this case, $n_0 = 1.00$ and $n_m = 4.00$. Hence $n_i = (n_0 n_m)^{1/2} = 2.0$ in both type I and II systems. First of all we take, for subsystem a, a straightforward single quarter-wave matching the substrate to the coupling medium:

$$\begin{array}{c|c|c} n_1 & n_a & n_m \\ \hline 2.0 & (n_i n_m)^{1/2} & 4.0 \\ & 2.826 & \end{array}$$

Subsystem b is then, for both type I and II systems

n_0	n_b	n_i
1.0	1.414	2.0.

Putting the two subsystems together, we have either a two-layer coating if we permit the thickness of the coupling layer to shrink to zero, or a three-layer coating if the coupling layer is a quarter-wave. In the former case we have the design:

Air	1.414	2.282	Ge
1.0	$0.25\lambda_0$	$0.25\lambda_0$	4.0

and in the latter

Air	1.414	2.0	2.282	Ge
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	4.0.

The first design gives a single minimum. The second, which is similar to the three-layer design already obtained by the vector method, has a broad three-minimum characteristic (figure 3.16).

The subsystems need not be perfect matching systems for n_m to n_i and n_i to n_0 . We could, for instance, use

$$\begin{aligned} n_0 &= 1.0 \\ n_b &= (1.0 \times 4.0)^{1/3} = 1.587 \\ n_m &= 2.0 \end{aligned}$$

from the two-layer coating derived by the vector method. This gives complete two- and three-layer coatings, as follows.

Type I

Air	1.587	3.174	Ge
1.0	$0.25\lambda_0$	$0.25\lambda_0$	4.0
Air	1.587	2.0	3.174
1.0	$0.25\lambda_0$	$0.25\lambda_0$	4.0.

Type II

Air	1.587	2.520	Ge
1.0	$0.25\lambda_0$	$0.25\lambda_0$	4.0
Air	1.587	2.0	2.520
1.0	$0.25\lambda_0$	$0.25\lambda_0$	4.0.

The first of the type II designs is identical to the vector method coating. Performance curves are given in figure 3.17.

Analytical expressions for calculating the positions of the zeros and the residual reflectance maxima of two- and three-layer coatings of the above types

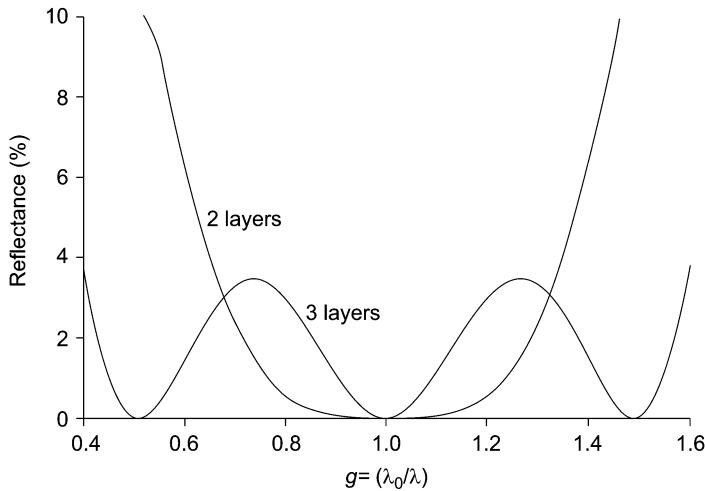


Figure 3.16. Theoretical performance of antireflection coatings on germanium designed by the method of Mussett and Thelen [6].

Two layers:	Air	1.414	2.828	Ge
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	4.00

Three layers:	Air	1.414	2.00	2.828	Ge
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	4.00.

are given by Musset and Thelen. The method can be readily extended to four and more layers.

Young [8] has developed alternative techniques for coatings consisting of quarter-wave optical thicknesses based on the correspondence between the theory of thin-film multilayers and that of microwave transmission lines. He gives a useful set of tables for the design of multilayer coatings where all thicknesses are quarter-waves. Given the bandwidth and the maximum permissible reflectance it is possible quickly to derive the coating which meets the specification with the least number of layers. The method, of course, takes no account of the possibility of achieving the given indices in practice, as with many of the other methods we have been discussing, but the optimum solution is a very useful point of departure in the design of coatings using real indices.

3.2 Antireflection coatings on low-index substrates

Although the theory developed for antireflection coatings on high-index materials applies equally well to low-index materials, the problem is made much more severe by the lack of any rugged thin-film materials of very low index.

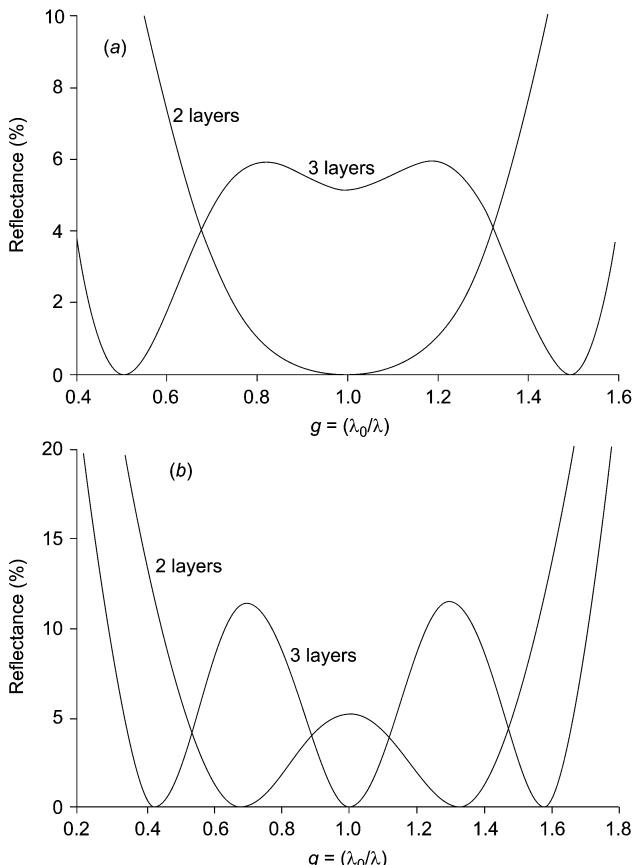


Figure 3.17. (a) Theoretical performance of type I antireflection coatings on germanium designed by the method of Mussett and Thelen [6].

Two layers:	Air	1.587	3.174	Ge
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	4.00

Three layers:	Air	1.587	2.00	3.174	Ge
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	4.00

(b) Theoretical performance of type II antireflection coatings on germanium designed by the method of Mussett and Thelen [6].

Two layers:	Air	1.587	2.520	Ge
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	4.00

Three layers:	Air	1.587	2.00	2.520	Ge
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	4.00

Magnesium fluoride, with an index of around 1.38, represents the lowest practical index that can be achieved. This immediately makes the manufacture of designs arrived at by the straightforward application of the techniques so far discussed largely impossible. Design techniques for antireflection coatings on low-index materials are less well organised and involve much more intuition and trial and error than those for high-index materials.

A very common low-index material is crown glass, and coatings are most frequently required for the visible region of the spectrum, which extends from around 400 nm to around 700 nm. Plastic materials of similar or higher refractive index are increasing in use, especially in lenses for spectacles. For the purposes of most of the coatings which we will discuss here, we will assume glass of index of 1.52, although this varies somewhat with the particular glass and also with wavelength. Although much of what follows is applied directly to the antireflection coating of crown glass, the techniques apply equally well to the coating of other low-index materials. We begin with the simplest coating, a single layer.

3.2.1 The single-layer antireflection coating

We can make use of the expressions already developed for high-index materials.

The optimum single-layer coating is a quarter-wave optical thickness for the central wavelength λ_0 with optical admittance given by

$$y_1 = (y_0 y_m)^{1/2}. \quad (3.14)$$

For crown glass in air, this represents

$$y_1 = (1.0 \times 1.52)^{1/2} = 1.23.$$

As already mentioned, the lowest useful film index which can be obtained at present is that of magnesium fluoride, around 1.38 at 500 nm. While not ideal, this does give a worthwhile improvement. The reflectance at the minimum is given by

$$R = \left(\frac{y_0 - y_1^2/y_m}{y_0 + y_1^2/y_m} \right)^2, \quad (3.15)$$

i.e. 1.3% per surface.

At angles of incidence other than normal, the phase thickness of the layer is reduced, so that for a given layer thickness the wavelength corresponding to the minimum becomes shorter. The optical admittance appropriate to the angle of incidence and the plane of polarisation should also be used in calculating the reflectance. Figure 3.18 indicates the way in which the reflectance of a single layer of magnesium fluoride on a substrate of index 1.52 can be expected to vary with angle of incidence.

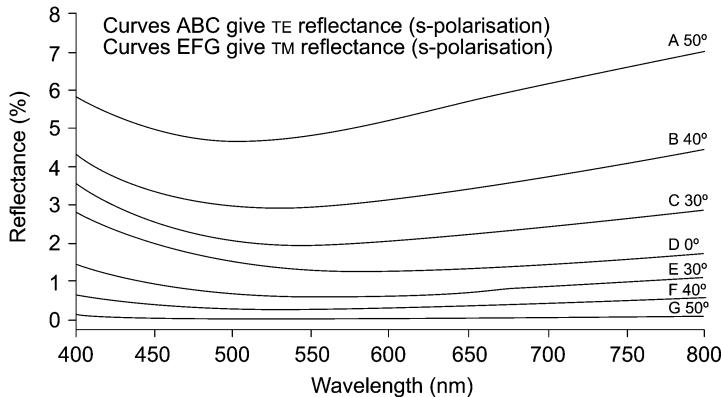


Figure 3.18. The computed reflectance at various angles of incidence of a single surface of glass of index 1.52 coated with a single layer of magnesium fluoride of index 1.38 and optical thickness at normal incidence one quarter-wave at 600 nm.

3.2.2 Two-layer antireflection coatings

The single-layer coating cannot achieve zero reflectance even at the minimum because of the absence of suitable low-index materials. Instinct suggests that a thin layer of high-index material placed next to the substrate might make it appear to have a higher index so that a subsequent layer of magnesium fluoride would be more effective. This proves to be the case. Two-layer coatings have already been considered with regard to high-index substrates and a complete analysis has been derived.

We can study the Schuster diagram (figure 3.8) for coatings on glass of index 1.52, and this is reproduced as figure 3.19. We can assume 1.38 as the lowest possible index, while a realistic upper bound to the range of possible indices is 2.45. Possible solutions are then limited to the shaded area of the diagram. This area is bounded by the lines

$$y_1 = 1.38 \quad y_2 = 2.45 \quad y_1 = y_2 (y_0/y_m)^{1/2}.$$

Solutions on the line

$$y_1 = y_2 (y_0/y_m)^{1/2}$$

will consist of two quarter-wave layers. Solutions elsewhere will consist of two layers of unequal thickness, one greater and the other less than a quarter-wave.

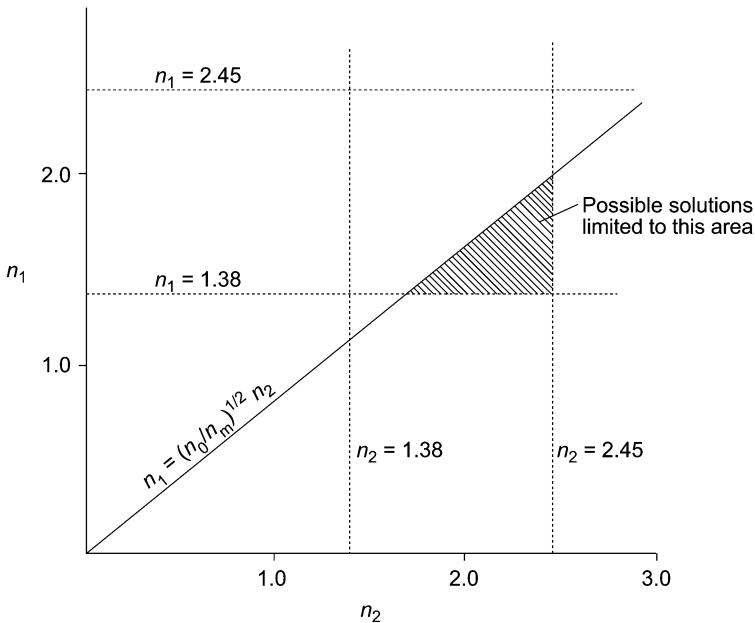


Figure 3.19. A Schuster diagram for two-layer coatings on glass ($n = 1.52$) in air ($n = 1.0$). Possible layer indices are assumed to be limited to the range 1.38–2.45.

The thicknesses are given by the expressions

$$\begin{aligned}\tan^2 \delta_1 &= \frac{(y_m - y_0)(y_2^2 - y_0 y_m)y_1^2}{(y_1^2 y_m - y_0 y_2^2)(y_0 y_m - y_1^2)} \\ \tan^2 \delta_2 &= \frac{(y_m - y_0)(y_0 y_m - y_1^2)y_2^2}{(y_1^2 y_m - y_0 y_2^2)(y_2^2 - y_0 y_m)}.\end{aligned}\quad (3.16)$$

As an example, we can take a value of 2.2 for the high-index layer, corresponding to, say, cerium oxide, and of 1.38 for the low-index layer, corresponding to magnesium fluoride. The two possible solutions are then

$$\delta_1/2\pi = 0.3208 \quad \delta_2/2\pi = 0.05877$$

and

$$\delta_1/2\pi = 0.1792 \quad \delta_2/2\pi = 0.4412,$$

respectively.

These two solutions are plotted in figure 3.20 and it can be clearly seen that the characteristic of the coating is a single minimum with a narrower bandwidth than the single layer, and that the broader of the two possible solutions is

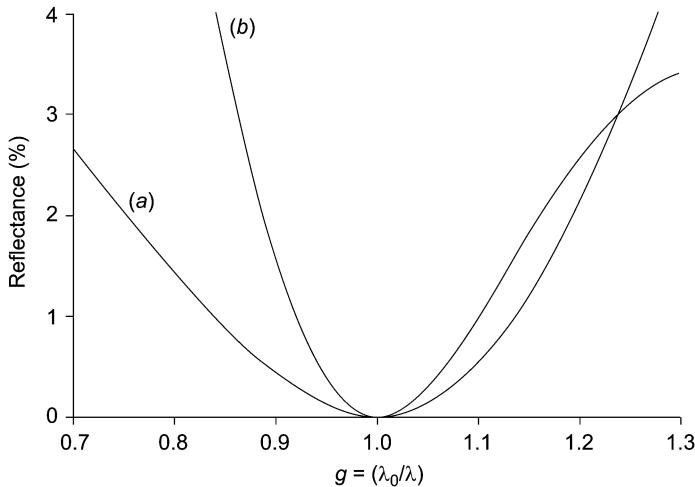


Figure 3.20. Two-layer antireflection coatings for glass.

(a)	Air	1.38	2.20	Glass
	1.00	$0.321\lambda_0$	$0.0588\lambda_0$	1.52

(b)	Air	1.38	2.20	Glass
	1.00	$0.179\lambda_0$	$0.441\lambda_0$	1.52.

(a), the broader characteristic, is usually selected. Because of the characteristic single minimum the coating is often known as a V-coat.

associated with the thinner high-index layer. The coating is also an effective one for other values of substrate index. The higher the index of the substrate, the thinner the high-index layer need be and the broader is the characteristic of the coating.

We can follow Catalan [2] and plot curves showing how the values of δ_1 and δ_2 vary with the index of the layer next to the substrate. Such curves are shown in figure 3.21 and from them several points of interest emerge. First, as already predicted by the Schuster plot, there is a region in which no solution is possible. Second, and more important, the curves flatten out as the index of the layer increases, and changes in refractive index are accompanied by only small changes in optical thickness. One of the problems in manufacturing coatings is the control of the refractive index of the layers, particularly of the high-index layers, and the curves indicate good stability of the performance of the coating in this respect.

The equations are not limited to normal incidence. Catalan has also computed, for various angles of incidence, values of reflectance of a two-layer coating consisting of bismuth oxide, with index 2.45, and magnesium fluoride,

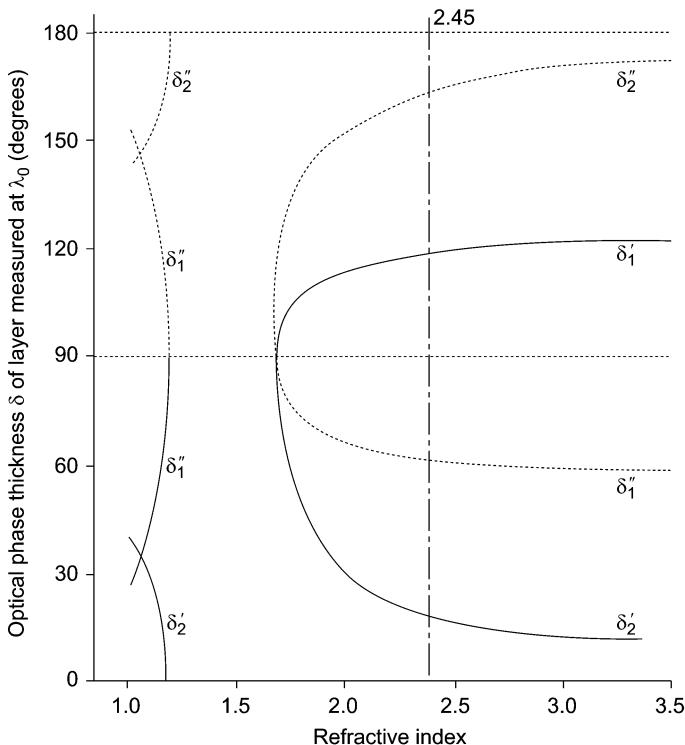


Figure 3.21. Optimum thicknesses of the layers in a double-layer antireflection coating at normal incidence. δ_1 and δ_2 , the optical phase thicknesses given by equations (3.7) and (3.8), are plotted against n_2 , the refractive index of the high-index layer. The low-index layer is assumed to be magnesium fluoride of index 1.38 and the coating is deposited on glass of index 1.50. Two pairs of solutions of (3.7) and (3.8) are possible for each set of refractive indices and are denoted by δ'_1 and δ'_2 and δ''_1 and δ''_2 . The value, 2.45, of refractive index, shown by the dashed line, corresponds to bismuth oxide and was used by Catalan in his calculations. (After Catalan [2].)

with index 1.38, on glass of index 1.5. Curves showing the variation of reflectance with angle of incidence are given in figures 3.22 and 3.23. The performance is very good up to an angle of incidence of 20° but beyond that it begins to fall off.

It may also be necessary to design coatings for angles of incidence other than normal. Turbadar [9] has considered this problem and published designs for angle of incidence of 45° . The materials were once again bismuth oxide and magnesium fluoride, of indices 2.45 and 1.38, respectively, on glass of index 1.5. Four possible solutions were given, which are reproduced as table 3.1 where the bismuth oxide is next to the glass.

A large number of performance curves of the various designs under different

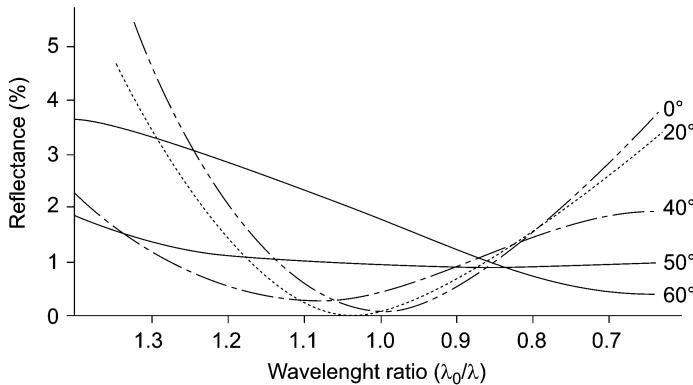


Figure 3.22. Theoretical p-reflectance (TM) as a function of wavelength ratio g ($= \lambda_0/\lambda$) of a double-layer antireflection coating. $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.45$, $n_m = 1.50$. (After Catalan [2].)

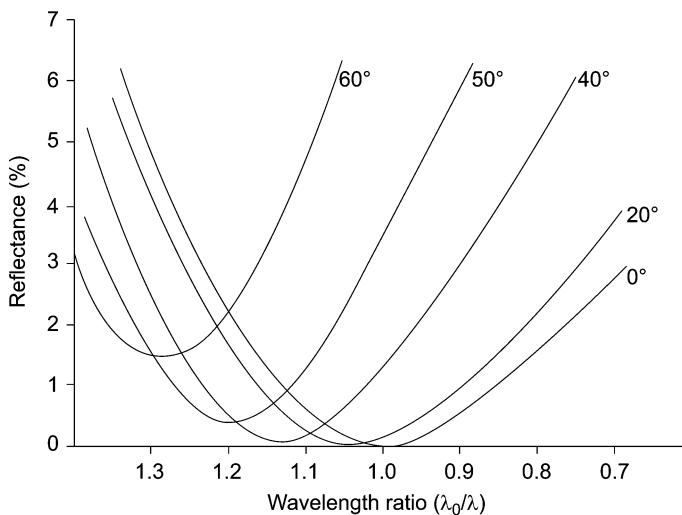


Figure 3.23. Theoretical s-reflectance (TE) as a function of wavelength ratio g ($= \lambda_0/\lambda$) of a double-layer antireflection coating. $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.45$, $n_m = 1.50$. (After Catalan [2].)

conditions, including the effect of errors, were produced. Today this is something we can do at great speed on a desktop computer. At the time this was not possible and the plots that were included of equireflectance contours over a grid of angle of incidence against wavelength were particularly valuable. The fact that they can now be more readily created does not reduce their usefulness and so they are

Table 3.1.

		Bismuth oxide	Magnesium fluoride
s-polarisation (TE wave)	S'	0.065 λ_0	0.376 λ_0
	S''	0.457 λ_0	0.206 λ_0
p-polarisation (TM wave)	P'	0.021 λ_0	0.382 λ_0
	P''	0.501 λ_0	0.201 λ_0

given in figure 3.24.

It is useful to consider an admittance plot for a two-layer coating, which can be a great help in visualising performance. The plot consists of two circles, the first corresponding to the low-index layer y_1 which passes through the point $(y_0, 0)$ if the reflectance is to be zero and which must, therefore, also pass through the point $(y_1^2/y_0, 0)$. The second circle corresponds to the high-index layer y_2 , which must pass through the point $(y_m, 0)$ corresponding to the substrate and, therefore, also through the point $(y_2^2/y_m, 0)$. Provided that these two circles intersect, then a two-layer antireflection coating of this type is possible. Such a plot is shown in figure 3.25. There are two possible arrangements of the admittance circles which will give the required zero reflectance. If we recall that a semicircle starting and finishing on the real axis corresponds to a quarter-wave, then we can see that either the high-index layer will be thinner than a quarter-wave with the low-index layer thicker, or the reverse, just as we have already established.

The special case where the layers are both quarter-waves can then be seen to occur when the y_2 circle just touches the y_1 circle internally. In that case

$$y_1^2/y_0 = y_2^2/y_m$$

or

$$y_1 = y_2 (y_0/y_m)^{1/2}$$

which is the equation of the oblique line in the Schuster plot. The admittance plot for $\lambda = \lambda_0$ and the theoretical performance curve for such a coating are shown in figure 3.26.

All the two-layer coatings considered so far exhibit one single minimum, which can be theoretically zero at $\lambda = \lambda_0$. On either side of the minimum, the reflectance rises rather more rapidly than for the single-layer coating. An alternative two-layer coating makes use of the broadening effects of a half-wave layer to produce an improvement over the single-layer performance. A half-wave layer of index higher than the substrate is inserted between the substrate and the quarter-wave low-index film. If magnesium fluoride, of index 1.38, is once again

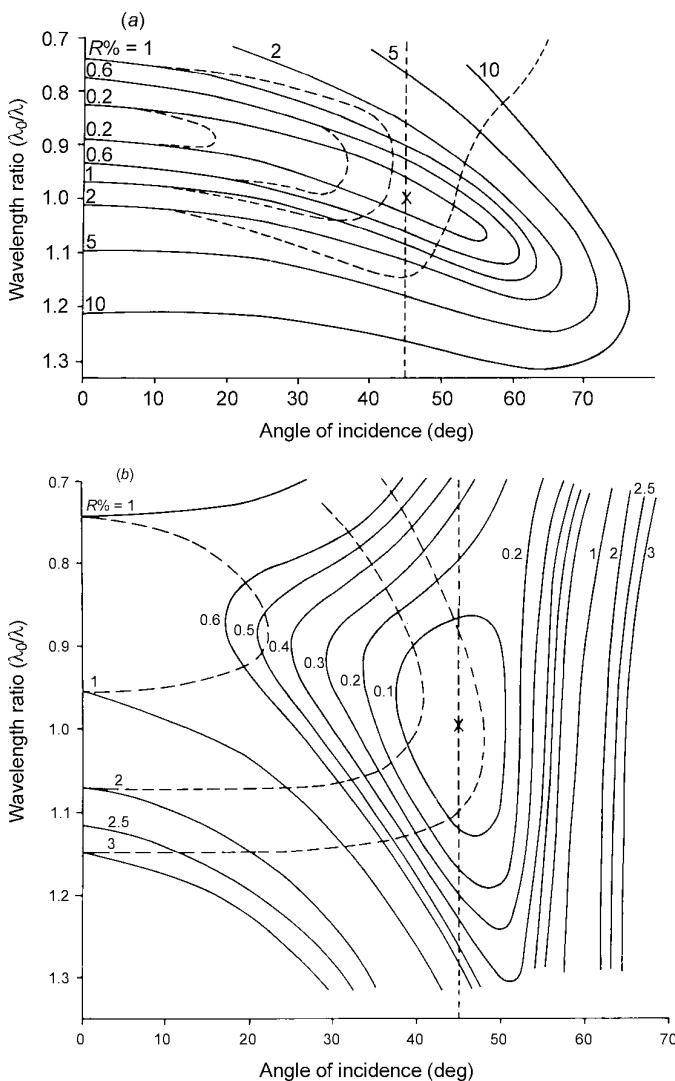


Figure 3.24. (a) Equireflectance contours for double-layer antireflection coatings on glass. $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.45$, $n_m = 1.50$, with layer thicknesses optimised for s-polarisation (TE) at 45° angle of incidence, given by S' in table 3.1. Solid curves s-reflectance (TE); dashed curves p-reflectance (TM). (After Turbadar [9].) (b) Equireflectance contours for double-layer antireflection coatings on glass. $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.45$, $n_m = 1.50$, with layer thicknesses optimised for p-polarisation (TM) at 45° angle of incidence, given by P' in table 3.1. Solid curves p-reflectance (TM); dashed curves s-reflectance (TE). (After Turbadar [9].)

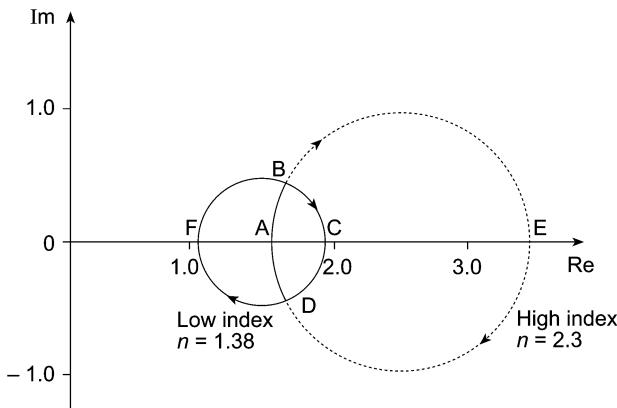


Figure 3.25. Admittance diagram showing the two possible double-layer antireflection coating designs.

chosen for the low-index film, then, for a substrate of index 1.52, the high-index layer should preferably be in the range 1.7–1.9, while, for a substrate of index 1.7, the range should be increased to 1.9–2.1. The way in which the half-wave layer acts to improve the performance can readily be understood by sketching an admittance plot, as in figure 3.27. The opening of the end of the high-index locus as the value of g decreases from 1.0 partially compensates for the shortening of the low-index locus. A similar effect exists as g increases from 1.0, when the lengthening of the low-index locus is compensated by an overlapping with the high-index locus. The half-wave layer must be of an index higher than that of the substrate, otherwise the opening of the half-wave circle would pull the low-index locus even further from the point $g = 1.0$, hence increasing the reflectance further and effectively narrowing the characteristic. The important feature of the arrangement is that, at the reference wavelength, the second quarter-wave portion of the half-wave layer and the following quarter-wave layer should have loci on the same side of the real axis.

3.2.3 Multilayer antireflection coatings

There is little further improvement in performance which can be achieved with two-layer coatings, given the limitations which exist in usable film indices. For higher performance, further layers are required.

Thetford [10] has devised a technique for designing three-layer antireflection coatings where the reflectance is zero at two wavelengths and low over a wider range than in the two-layer coating. The arrangement consists of a layer of intermediate index next to the substrate, followed by a high-index layer and finally by a low-index layer on the outside. The indices are chosen at the outset and the method yields the necessary layer thicknesses. There is an advantage in

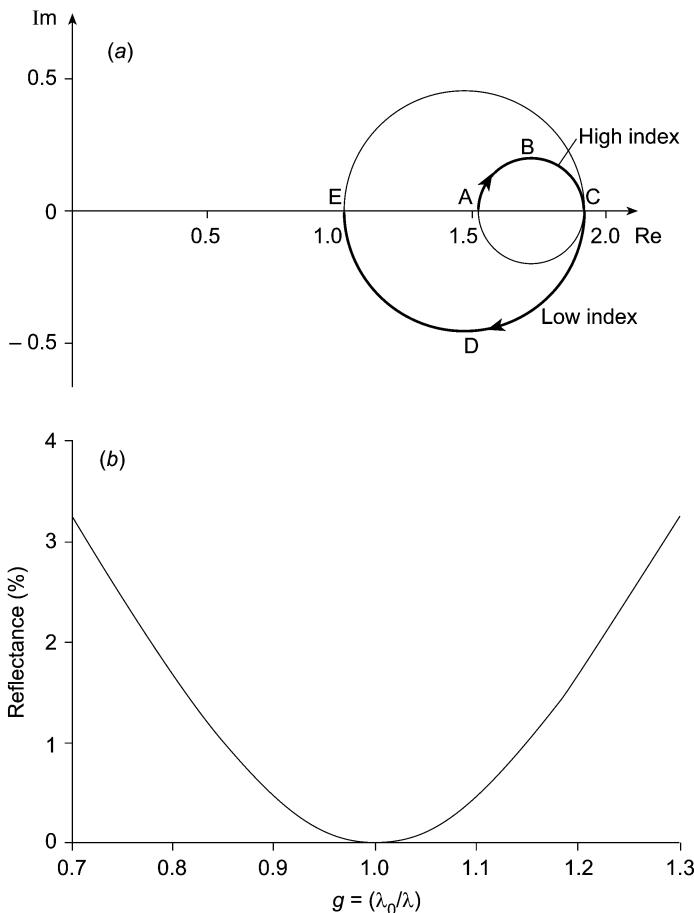


Figure 3.26. Special case of the two-layer antireflection coating where the layers become quarter-waves and the two solutions of figure 3.25 merge into one. The design is:

Air	1.38	1.70	Glass
1.00	$0.25\lambda_0$	$0.25\lambda_0$	1.52

(a) The admittance locus. (b) The theoretical performance curve.

specifying layer indices rather than thicknesses because of the limited range of materials available. Although the actual design of a coating would probably be most efficiently tackled by a process of refinement of a likely starting design, our working through the Thetford method is nevertheless worthwhile because it is an excellent example of reasoning using the vector diagram and it gives great insight.

The technique is based on both the vector method and Smith's method (the

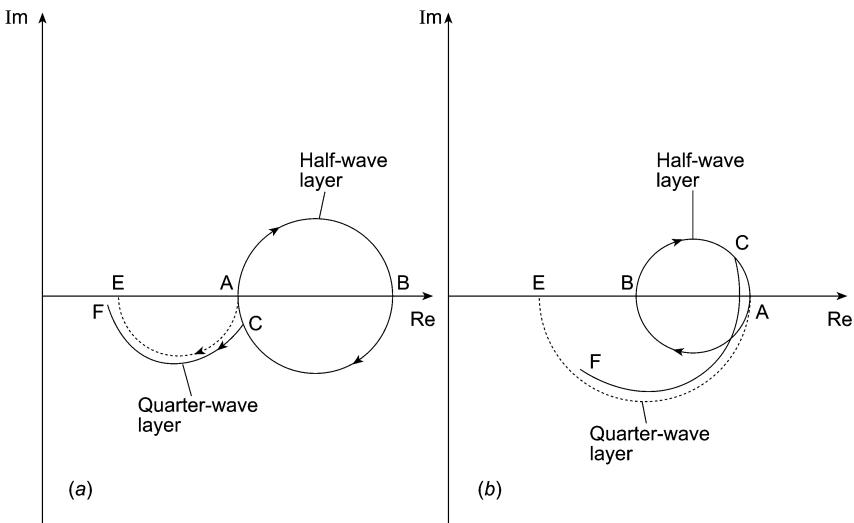


Figure 3.27. The operation of a half-wave flattening layer. The contour AE represents a low-index quarter-wave coating and in ABA a half-wave layer is inserted between it and the substrate. In (a) the half-wave is of index higher than the substrate and, as $g (= \lambda_0/\lambda)$ varies, the action of the half-wave keeps the end of the quarter-wave near the point E and the reflectance remains low. ABCF represents the locus with g somewhat less than unity. g greater than unity would give a similar effect with the point C now above the real axis and the loci slightly longer than full circle and semicircle. (b) shows the corresponding diagram for a low-index half-wave. Here the end point is dragged rapidly away from E as g varies and the reflectance rises rapidly. Flattening is therefore effective in (a) but not in (b). Note that the reflectance curve for another coating with half-wave flattening layer of design:

Air	1.38	1.90	Glass
1.00	$0.25\lambda_0$	$0.5\lambda_0$	1.52

is shown as curve (a) of figure 3.31. This latter coating is sometimes called a W-coat because of the shape of the characteristic.

method of effective interfaces). We recall that the transmittance of an assembly will be unity if, and only if, the reflectances of the structures on either side of the chosen spacer layer are equal and the thickness of the spacer layer is such that the phase change suffered by a ray of the appropriate wavelength, after having completed a round trip in the layer, being reflected once at each of the boundaries, is zero or an integral multiple of 2π . If the phase thickness of the layer is δ , then this is equivalent to saying that

$$\varphi + \varphi' - 2\delta = 2s\pi \quad s = 0, \pm 1, \pm 2, \dots \quad (3.17)$$

where φ and φ' are the phases of the amplitude reflection coefficients at the boundaries of the layer. Thetford split the assembly into two parts on either side of the middle layer and then computed the two amplitude reflection coefficients by the vector method, combining the calculations on one diagram. He chose thicknesses for the layers which made the reflectances equal at a reference wavelength. He then found expressions for the change in reflectance with wavelength for each of the two structures, and, from them, a second value of wavelength, shorter than the first, at which the reflectances were again equal. The next step was to compute the thickness of the middle layer to satisfy the phase condition at the first wavelength and hence to give zero reflectance for the complete coating at that wavelength, and then to check whether or not the phase condition was also satisfied at the second wavelength. If it was, then the reflectance of the complete coating was known to be zero at this wavelength and the design was complete. If it was not, then the procedure was repeated with slightly different initial conditions at the reference wavelength. This trial-and-error procedure turned out to be a very quick method of arriving at the final solution. The only step which remained was the accurate calculation of the performance of the design as a check.

The three-layer coating is shown in figure 3.28. Thetford's notation has been altered to fit in with the practice in this book. The vector diagrams for the two structures are shown in (b) and (c) and then combined in (d), with vectors in such a position that the resultant amplitude reflection coefficients ρ and ρ' are equal in length but not necessarily in phase. In the solution shown, both ρ and ρ' are in the fourth quadrant. It is very easy to arrive at this initial condition. All that is required is a circle with centre the origin which cuts both the loci of vectors ρ_a and ρ_d . This initial condition we can take as corresponding to our reference wavelength λ_0 . Figure 3.28(e) shows a second solution for a shorter wavelength λ_1 plotted on top of the first. The values of δ_1 and δ_3 which correspond to this solution are given by λ_0/λ_1 times the values corresponding to λ_0 , and ρ is now in the first quadrant while ρ' remains in the fourth. To find this second solution, Thetford has derived approximate expressions for the change in reflectance with change in wavelength which turn out to give surprisingly accurate results.

The reflectances corresponding to ρ and ρ' are given, from the diagram, by

$$\rho^2 = \rho_a^2 + \rho_b^2 + 2\rho_a\rho_b \cos 2\delta_1 \quad (3.18)$$

and

$$(\rho')^2 = \rho_c^2 + \rho_d^2 + 2\rho_c\rho_d \cos 2\delta_3. \quad (3.19)$$

For a reasonably small change in wavelengths we can find the corresponding change in ρ^2 and $(\rho')^2$ by differentiating equations (3.18) and (3.19), i.e.

$$\begin{aligned} \Delta(\rho^2) &= -4\rho_a\rho_b \sin 2\delta_1 \times \Delta\delta_1 \\ \Delta[(\rho')^2] &= -4\rho_c\rho_d \sin 2\delta_3 \times \Delta\delta_3. \end{aligned}$$

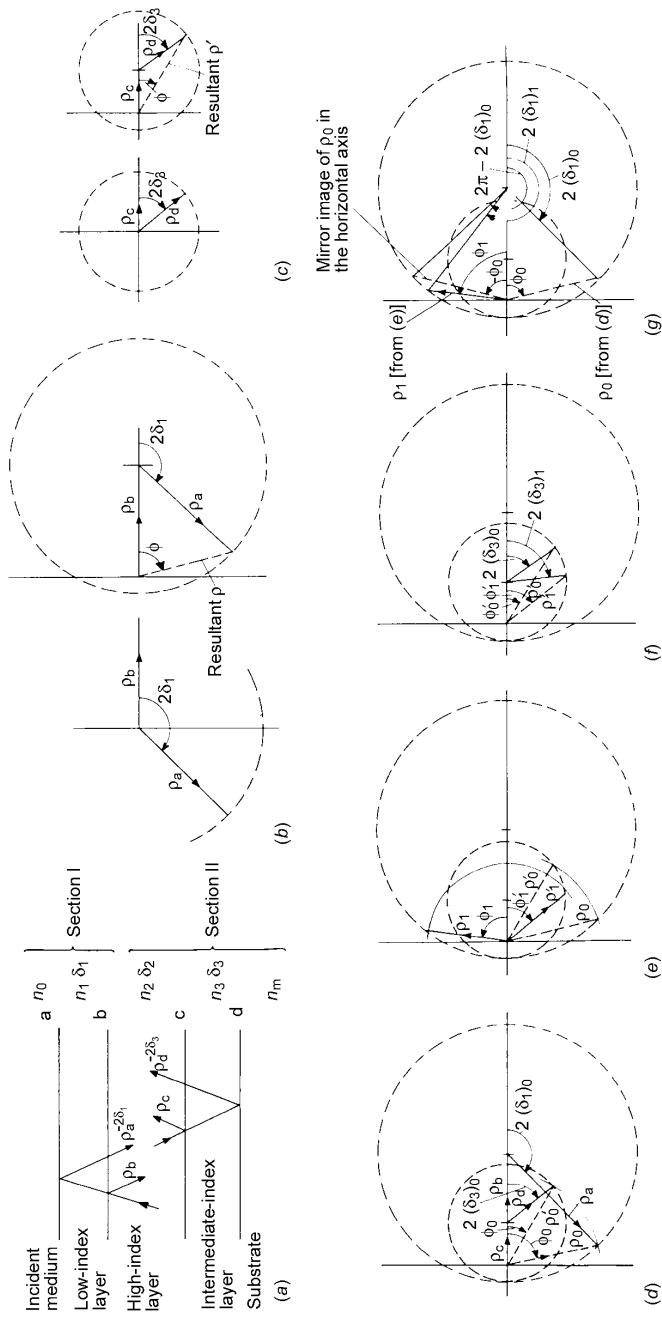


Figure 3.28. Theiford's method for antireflection coating design. (a) The three-layer coating split into two sections, I and II, on either side of the central high-index layer. (b) The vector diagram of section I. (c) The vector diagram of section I. (d) The vector diagrams of sections I and II superimposed, showing one possible orientation of ρ and ρ' so that they are of equal length. (e) An alternative orientation of ρ and ρ' where again they are of equal length. The arrangement in (d) is also shown. The vectors p_a , ρ_b , ρ_c and ρ_d have been omitted. (f) The two possible orientations of ρ from (d) and (e). Also shown is the mirror image of ρ_0 in the horizontal axis. (After Theiford [10].)

Now since the two values of $(\rho')^2$ in which we are interested are in the fourth quadrant, and well clear of any turning values, we can apply this approximate expression directly, giving

$$\Delta[(\rho')^2] = -4\rho_c\rho_d \sin 2(\delta_3)_0 \times \Delta\delta_3$$

$$\Delta\delta_3 = \left(\frac{\lambda_0}{\lambda_1} - 1 \right) (\delta_3)_0$$

for the change in $(\rho')^2$ corresponding to the shift in wavelength from λ_0 to λ_1 , where $(\delta_3)_0$ is the value at λ_0 .

ρ^2 , however, is not so simple. It passes through a turning value between the two solutions. Thetford observed that, in figure 3.28(c), the mirror image of ρ in the horizontal axis would also give the same resultant ρ^2 (although with a different phase angle), and that this would be fairly near the desired solution. This new position of ρ_a has angle $2\delta_1$, with value $2\pi - 2(\delta_1)_0$ and a change in this angle of

$$\Delta\delta_1 = \left[\left(1 + \frac{\lambda_0}{\lambda_1} \right) (\delta_1)_0 - \pi \right] \quad (3.20)$$

would swing it round exactly into the correct position. We can therefore find the change in ρ^2 that we want by using the approximate expression, but calculating it as a change of $\Delta\delta_1$ (equation (3.20)) from this fictitious position of ρ_a . $\Delta(\rho^2)$ is then given by

$$\Delta(\rho^2) = -4\rho_a\rho_b \sin [2\pi - 2(\delta_1)_0] \left[\left(1 + \frac{\lambda_0}{\lambda_1} \right) (\delta_1)_0 - \pi \right]$$

$$= 4\rho_a\rho_b \sin 2(\delta_1)_0 \left[\left(1 + \frac{\lambda_0}{\lambda_1} \right) (\delta_1)_0 - \pi \right].$$

We must now set $\Delta[(\rho')^2] = \Delta(\rho^2)$, which permits us to solve for λ_1 . Next, we investigate the phase condition and the thickness of the middle layer.

From the vector diagram for the first solution we can find the phase angles φ_0 and φ'_0 associated with ρ and ρ' and λ_0 . The necessary phase thickness of the middle layer to satisfy the condition for zero reflectance is given from equation (3.17) by

$$2(\delta_2)_0 = 2\pi + \varphi_0 + \varphi'_0$$

where we must remember to include the signs of φ_0 and φ'_0 (both negative in figure 3.28(d)) and where we have taken s as +1 to give the thinnest possible positive value for $(\delta_2)_0$. Next, from the vector diagram we find the values of phase angle φ_1 and φ'_1 associated with λ_1 . If these satisfy the expression

$$2(\delta_2)_0 \frac{\lambda_0}{\lambda_1} = 2\pi + \varphi_1 + \varphi'_1 \quad (3.21)$$

then we know we have a valid solution. The phase angles of the layers at λ_0 are then given by $(\delta_1)_0$, $(\delta_2)_0$ and $(\delta_3)_0$, respectively, and the optical thicknesses of the layers in terms of a quarter-wave at λ_0 can be found by dividing by $\pi/2$. If, however, the phase condition is not met at λ_1 then it is necessary to go back to the beginning and try a new set of solutions. In fact, a satisfactory solution will be found quickly, especially if the error in equation (3.21) is plotted against, say, $(\delta_1)_0$.

One advantage which Thetford has pointed out for this type of coating is that once the phase condition has been satisfied at both λ_0 and λ_1 it will be approximately satisfied at all wavelengths between them. This means that the design will possess a broad region of low reflectance without any pronounced peaks of high reflectance. Some of Thetford's designs are shown in figure 3.29, which also demonstrates how the characteristic varies with the index of the middle layer. This coating is clearly a considerable improvement over the two-layer coating.

It is not easy to establish analytical expressions for the ranges of n_1 , n_2 and n_3 that will give an acceptable reflectance characteristic. Generally, if the Argand diagram is not too far removed in appearance from the form of figure 3.28 where the two positions of ρ are near the minimum, which corresponds to $2\delta_1 = \pi$, then a good antireflection coating will be obtained.

If it should be a requirement that only two values of refractive index rather than three be used in the construction of the coating, then it is possible to achieve a similar performance if four layers of alternate high and low index are used. Thetford [11] has used a similar technique for the design of such a coating. He split the coating (which has a high-index layer next to the glass) at the high-index layer nearest the air, so that the high-low combination next to the glass took the place of the intermediate-index layer of the three-layer design. If the thicknesses of these two layers are fairly small, then an Argand diagram is obtained which is not too different from that for the three-layer design. Because the expressions would be much more complicated in this case, Thetford did not attempt an analytical solution, but rather arrived at a design which appeared reasonable, by trial and error. The reflectance characteristic of such a design is shown in figure 3.30. This solution was then refined by C Butler, using a computer technique, to give optimum performance. This improved coating is also shown in figure 3.30.

There are also many coatings which involve layers of either quarter-wave or half-wave optical thicknesses. A number of these can be looked upon as modifications of some of the two-layer designs already considered.

First, we take the two-layer coating consisting of a half-wave layer next to the substrate followed by a quarter-wave layer. This has a peak reflectance in the centre of the low-reflectance region. This peak corresponds to the minimum reflectance of a single-layer coating because the inner layer, being a half-wave at that wavelength, is an absentee. We can reduce the peak but retain to some extent the flattening effect of the half-wave layer by splitting it into two quarter-waves,

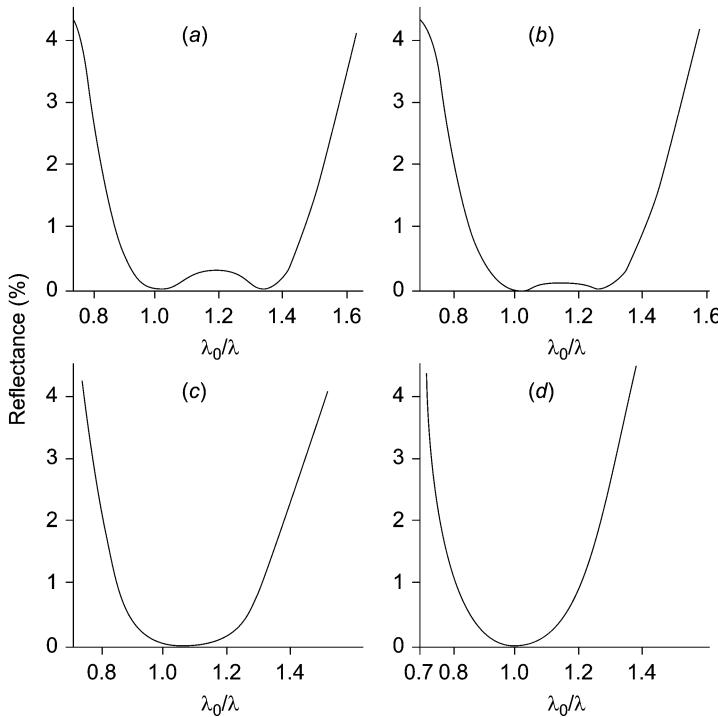


Figure 3.29. Calculated reflectance of some three-layer antireflection coatings designed by Thetford. The designs are as follows. (a) $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.00$, $n_3 = 1.80$, $n_4 = n_m = 1.52$, $n_1d_1 = 0.205\lambda_0$, $n_2d_2 = 0.336\lambda_0$, $n_3d_3 = 0.132\lambda_0$. (b) $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.10$, $n_3 = 1.80$, $n_4 = n_m = 1.52$, $n_1d_1 = 0.225\lambda_0$, $n_2d_2 = 0.359\lambda_0$, $n_3d_3 = 0.152\lambda_0$. (c) $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.20$, $n_3 = 1.80$, $n_4 = n_m = 1.52$, $n_1d_1 = 0.227\lambda_0$, $n_2d_2 = 0.338\lambda_0$, $n_3d_3 = 0.170\lambda_0$. (d) $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.40$, $n_3 = 1.80$, $n_4 = n_m = 1.52$, $n_1d_1 = 0.247\lambda_0$, $n_2d_2 = 0.445\lambda_0$, $n_3d_3 = 0.181\lambda_0$. (After Thetford [10].)

only slightly different in index. The first layer we can retain as 1.9, although it is in no way critical, and then if we make the second quarter-wave of slightly higher index, 2.0, say, the design now becoming

Air	1.38	2.0	1.9	Glass
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	1.52

we find a reduction in the reflectance at λ_0 from 1.26% to 0.38%. The characteristic remains fairly broad. Increasing the index of the central layer still further, to 2.13, i.e. a design

Air	1.38	2.13	1.9	Glass
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	1.52

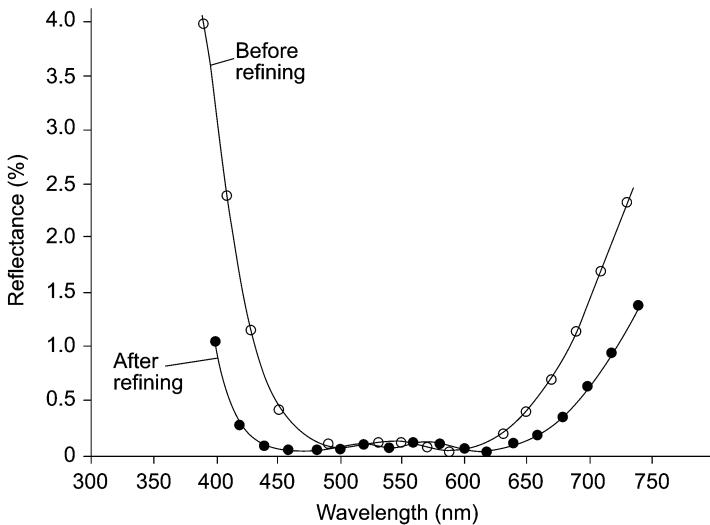


Figure 3.30. Calculated reflectance of four-layer antireflection coatings on glass showing the performance before and after the design was refined by computer. The two designs are as follows. (a) Before refining: $n_0 = 1.00$, $n_1 = n_3 = 1.38$, $n_2 = n_4 = 2.10$, $n_5 = n_m = 1.52$, $n_1 d_1 = 0.21\lambda_0$, $n_2 d_2 = 0.37\lambda_0$, $n_3 d_3 = 0.036\lambda_0$, $n_4 d_4 = 0.070\lambda_0$. (b) After refining: $n_0 = 1.00$, $n_1 = n_3 = 1.38$, $n_2 = n_4 = 2.10$, $n_5 = n_m = 1.52$, $n_1 d_1 = 0.216\lambda_0$, $n_2 d_2 = 0.458\lambda_0$, $n_3 d_3 = 0.072\lambda_0$, $n_4 d_4 = 0.049\lambda_0$. (Communicated by Thetford.)

reduces the reflectance at λ_0 to virtually zero, but the width of the coating becomes much more significantly reduced. The characteristic curves of these two coatings are shown in figure 3.31.

Yet a further increase in the width of the coating can be achieved by adding a half-wave layer of low index next to the substrate. The admittance plot is shown in figure 3.32 and we see the characteristic shape where the final part of the locus of the half-wave layer and the start of the following layer are on the same side of the real axis. A half-wave layer in the same position with index higher than the substrate would be ineffective. A certain amount of trial and error leads to the designs shown in figure 3.32, that is

Air	1.38	1.905	1.76	1.38	Glass
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	$0.5\lambda_0$	1.52

and

Air	1.38	2.13	1.9	1.38	Glass
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	$0.5\lambda_0$	1.52.

An alternative approach is to broaden the quarter-quarter design of figure 3.26 by inserting a half-wave layer between the two quarter-waves. In

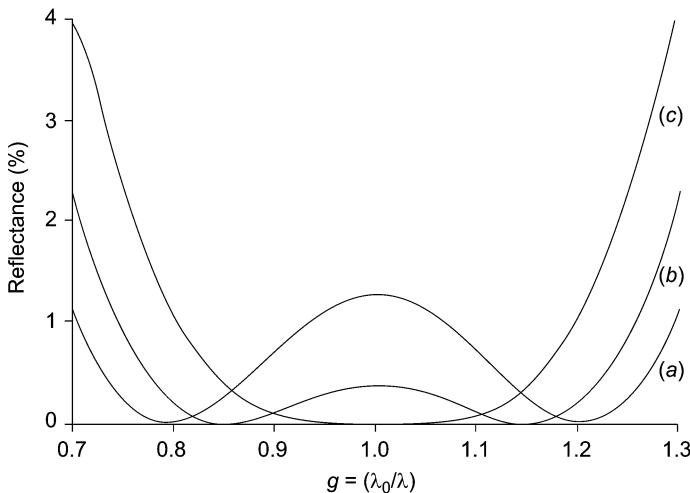


Figure 3.31. Progressive changes in an antireflection coating consisting of three quarter-wave layers. (a) The original coating:

Air	1.38	1.90	1.90	Glass
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	1.52.

The two 1.90 index layers combine to form a single half-wave layer. This is known as a W-coat because of the shape of the characteristic. (b) The index of the central layer is increased to 2.00. (c) The index of the central layer is increased further to 2.13.

order to achieve the broadening effect it must, of course, be of high index, so that the admittance plot will be of the form shown in figure 3.33. The coating is frequently referred to as the quarter-half-quarter coating. Coatings that fit into this general type date back to the 1940s and were described by Lockhart and King [12]. A systematic design technique explaining the functions of the various layers, however, was not available until the detailed study of Cox *et al* [13]. A certain amount of trial and error leads to the characteristics of figure 3.34. However, good results are obtained with values of the index of the half-wave layer in the range 2.0–2.4. Cox *et al* also investigated the effect of varying the indices of the quarter-wave layers and found that, for the best results on crown glass, the outermost layer index should be between 1.35 and 1.45, and the innermost layer index between 1.65 and 1.70. The outermost layer is the most critical in the design.

Figure 3.35 also comes from their paper and shows the measured reflectance of an experimental coating consisting of magnesium fluoride, index 1.38, zirconium oxide, index 2.1, and cerium fluoride, which was evaporated rather too slowly and had an index of 1.63, which accounts for the slight rise in the middle of the range. Otherwise, the coating is an excellent practical confirmation of the theory.

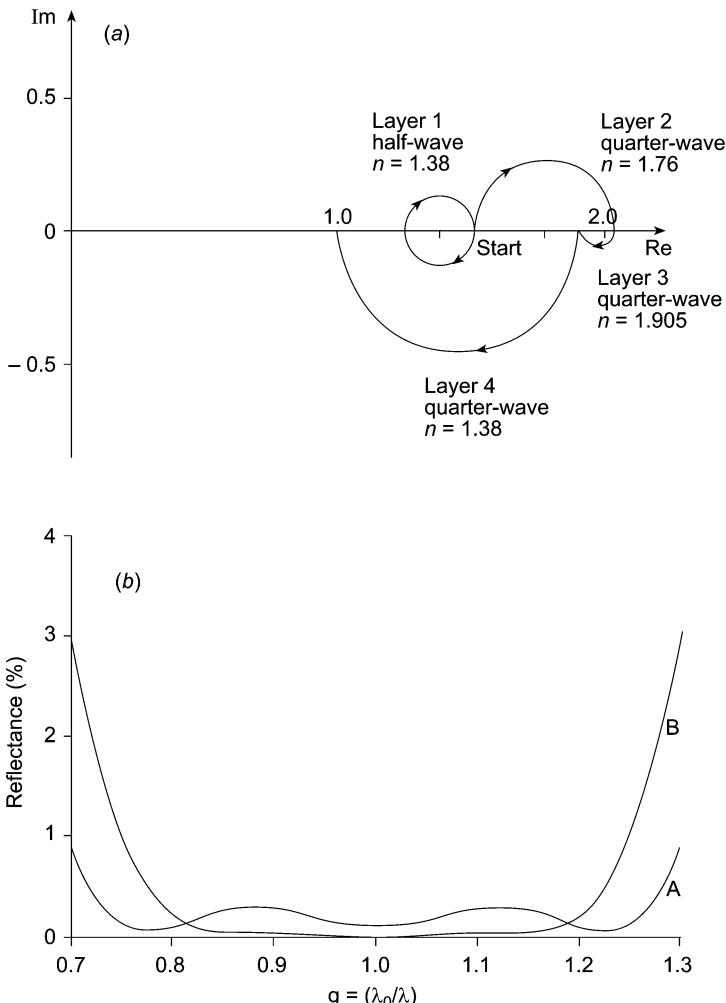


Figure 3.32. (a) The admittance locus of the coating:

Air	1.38	1.905	1.76	1.38	Glass
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	$0.5\lambda_0$	1.52.

(b) The characteristics of (A) the coating of figure 3.32(a) and (B) the coating (c) of figure 3.31 with a half-wave flattening layer of index 1.38 added next to the substrate.

The effect of variations in angle of incidence has also been examined. Cox *et al.*'s results for tilts up to 50° of a coating designed for normal incidence are shown in figure 3.36. The performance of the coating is excellent up to 20° but begins to fall off beyond 30° . The coatings can, of course, be designed for use

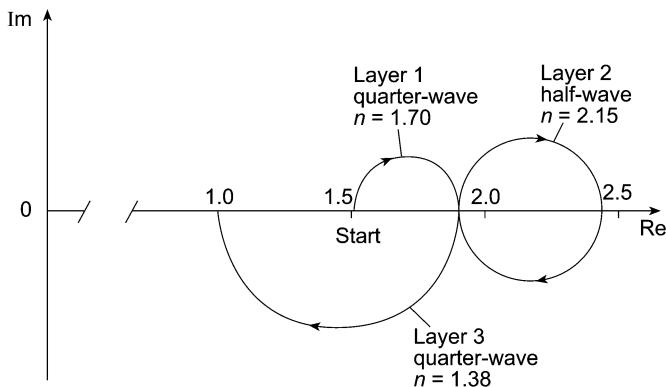


Figure 3.33. Admittance locus of the quarter–half–quarter coating:

Air	1.38	2.15	1.70	Glass
1.00	$0.25\lambda_0$	$0.5\lambda_0$	$0.25\lambda_0$	1.52.

The half-wave layer acts to flatten the performance of the two quarter-waves.

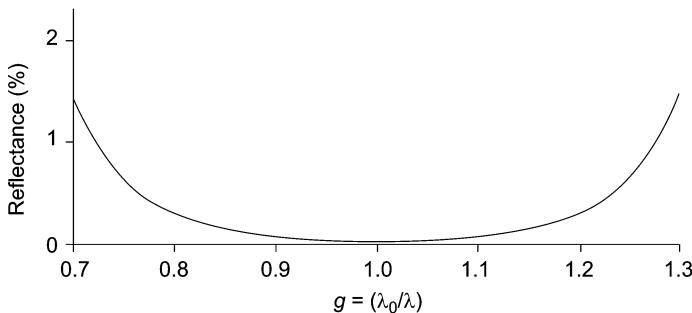


Figure 3.34. The calculated reflectance of the quarter–half–quarter coating shown in figure 3.33.

at angles of incidence other than normal, and Turbadar [14] has published a full account of a design for use at 45° . The particular design depends on whether light is s- or p-polarised and figure 3.37 shows sets of equireflectance contours for both designs.

The quarter–half–quarter coating is certainly the most significant of the early multilayer coatings for low-index glass and it has had considerable influence on the development of the field.

The success of the broadening effect of the half-wave layer on the quarter–quarter coating prompts us to consider inserting a similar half-wave in the two-layer coating of figure 3.25. In this case, there is an advantage in using a layer of

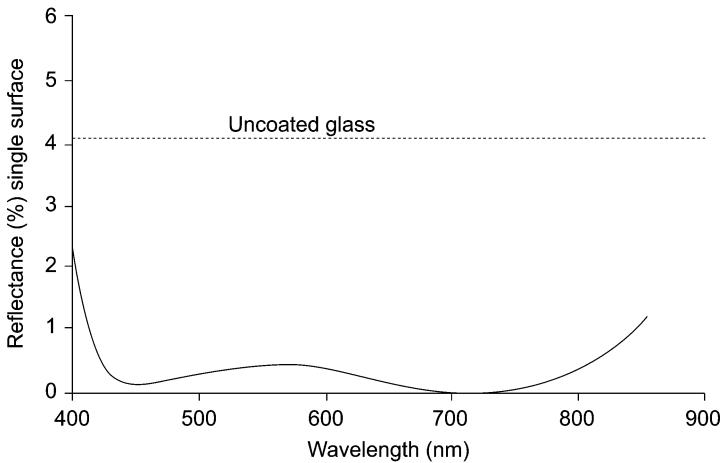


Figure 3.35. Measured reflectance of a quarter–half–quarter antireflection coating of $\text{MgF}_2 + \text{ZrO}_2 + \text{CeF}_3$ on crown glass. $\lambda_0 = 550$ nm. (After Cox *et al* [13].)

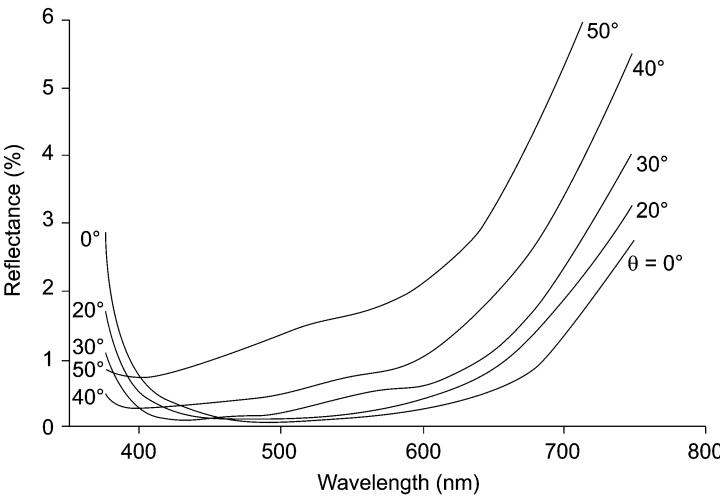


Figure 3.36. Calculated reflectance as a function of wavelength for quarter–half–quarter antireflection coatings on glass at various angles of incidence. $n_0 = 1.00$, $n_1 = 1.38$, $n_2 = 2.2$, $n_3 = 1.70$, $n_m = 1.51$. (After Cox *et al* [13].)

the same index as that next to the substrate. Here we cannot split the coating at the interface between the high- and the low-index layers, because the admittance plot would not show the correct broadening configuration. Instead, we must split the coating at the point where the low-index locus cuts the real axis so that the

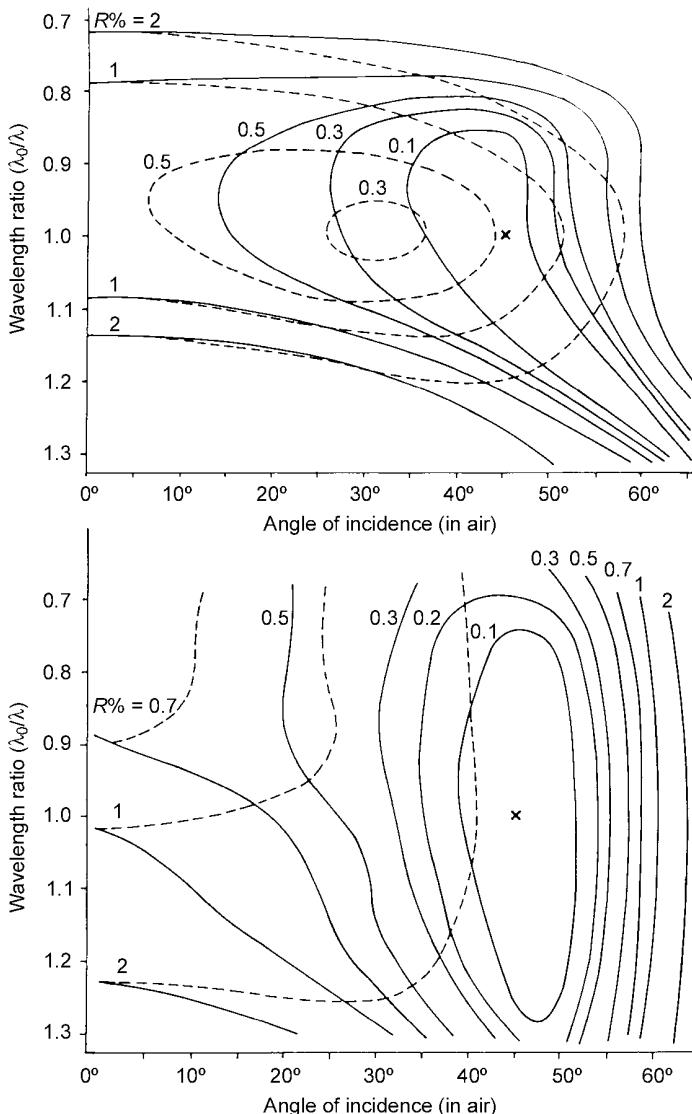


Figure 3.37. (a) Equireflectance contours for a quarter–half–quarter antireflection coating designed for use at 45° on crown glass. The indices are chosen for best performance with s-polarisation (TE). $n_0 = 1.00$, $n_1 = 1.35$, $n_2 = 2.45$, $n_3 = 1.70$, $n_m = 1.50$. Solid curves s-polarisation (TE); dashed curves p-polarisation (TM). (After Turbadar [14].) (b) Equireflectance contours for a quarter–half–quarter antireflection coating designed for use at 45° on crown glass. The indices are chosen for best performance with p-polarisation (TM). $n_0 = 1.00$, $n_1 = 1.40$, $n_2 = 1.75$, $n_3 = 1.58$, $n_m = 1.50$. Solid curves p-polarisation (TM); dashed curves s-polarisation (TE). (After Turbadar [14].)

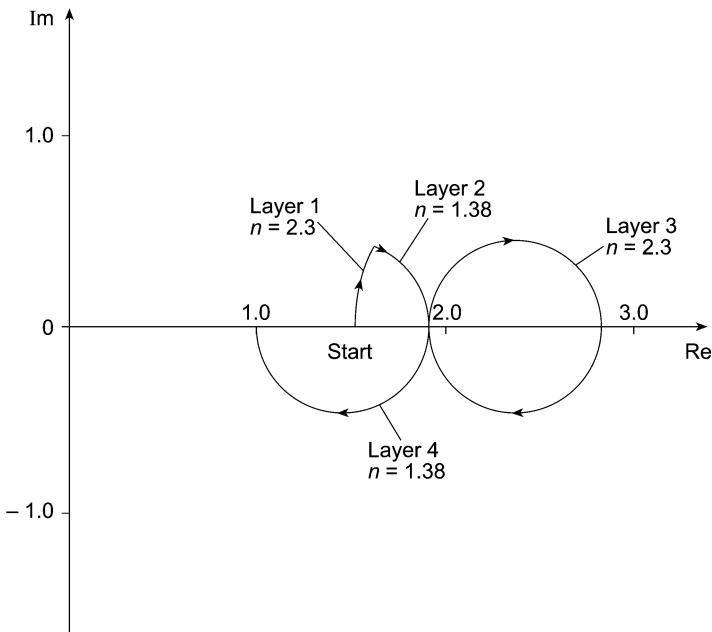


Figure 3.38. The two-layer coating of figure 3.25 with the low-index layer split where it intersects the real axis and a high-index flattening layer inserted.

plot appears as in figure 3.38. The design of the coating is then

Air	1.38	2.30	1.38	2.30	Glass
1.0	$0.25\lambda_0$	$0.5\lambda_0$	$0.0734\lambda_0$	$0.0522\lambda_0$	1.52

where this time we have used a value of 2.30 for the high index, and the performance is shown in figure 3.39. There is a considerable resemblance between this admittance plot and that of the quarter-half-quarter design. This design approach can be attributed originally to Frank Rock, who used the properties of reflection circles in deriving it, rather than admittance loci.

Vermeulen [15] arrived independently at an ultimately similar design in a completely different way. There is a difficulty in achieving the correct value for the intermediate index in the quarter-half-quarter design in practice and Vermeulen realised that the deposition of a low-index layer over a high-index layer of less than a quarter-wave would lead to a maximum turning value in reflectance rather lower than would have been achieved with a quarter-wave of high index on its own. He therefore designed a two-layer high-low combination to give an identical turning value to that which should be obtained with the 1.70 index layer of the quarter-half-quarter coating, and he discovered that good performance was maintained. The turning value in reflectance must, of course,

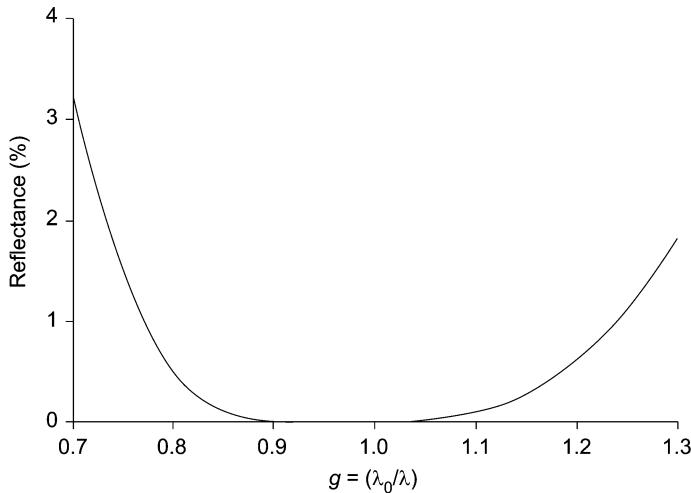


Figure 3.39. The performance of the coating of figure 3.38. Although arrived at by way of the admittance plot of figure 3.38, the design is virtually identical to one published by Vermeulen whose design technique was quite different (see text).

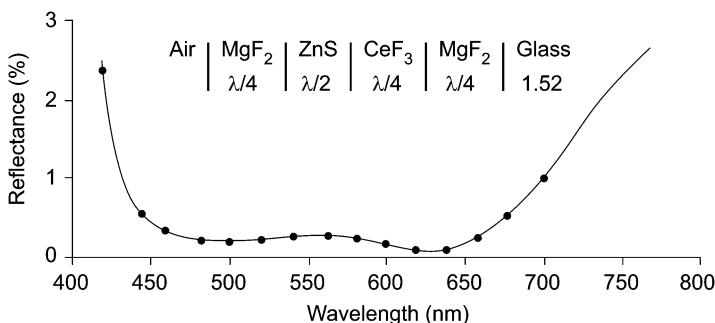


Figure 3.40. Measured reflectance of a four-layer antireflection coating on crown glass. The results are for a single surface. (After Shadbolt [16].)

correspond to the intersection of the locus with the real axis, and the rest follows. We shall return to this coating later.

The quarter-half-quarter coating can be further improved by replacing the layer of intermediate index by two quarter-wave layers. The layer next to the substrate should have an index lower than that of the substrate. A practical coating of this general type is shown in figure 3.40. Trial and error leads to a design

Air	1.38	2.05	1.60	1.45	Glass
1.0	$0.25\lambda_0$	$0.5\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	1.52

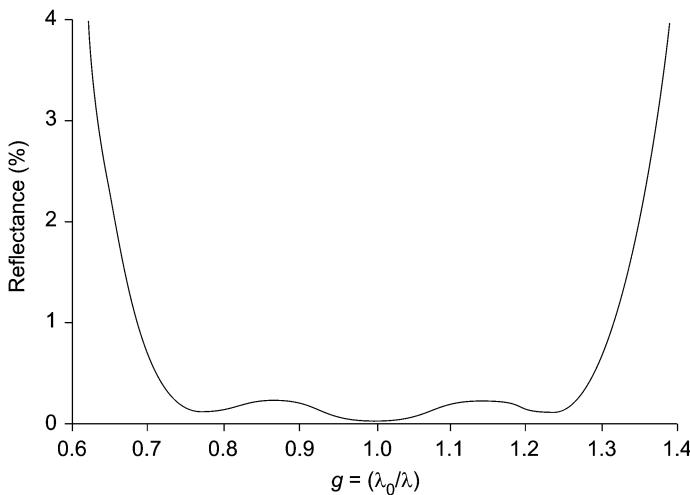


Figure 3.41. The performance of the four-layer coating of design:

Air	1.38	2.05	1.60	1.45	Glass
1.00	$0.25\lambda_0$	$0.5\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	1.52.

the theoretical performance of which is shown in figure 3.41. Similar designs with slightly different index values are given by Cox and Hass [17] and by Musset and Thelen [6]. Ward [18] has published a particularly useful version of this coating with indices chosen to match those of available materials rather than to achieve optimum performance. Examples of four-layer coatings for substrates of indices other than 1.52 are also given by Ward and by Musset and Thelen [6].

Yet a further four-layer design can be obtained by splitting the half-wave layer of the quarter–half–quarter coating into two quarter-waves and adjusting the indices to improve the performance. A five-layer design (see figure 3.42) derived in a similar way from the design of figure 3.41 is:

Air	1.38	2.13	2.13	1.38	2.30	Glass
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	1.52.

The possibilities are clearly enormous and problems are found much more in the construction of the coatings because not all the required indices are readily available. One solution is discussed in the next section.

A rather interesting design based on four layers of alternate high and low index has been published by C Reichert Optische Werke AG [19]. Full details of the design method are, unfortunately, not given. The thicknesses and materials are given in table 3.2. Note that the thicknesses are quoted as optical. The reflectance of this coating, figure 3.43, is slightly better than the unrefined performance of figure 3.30 but inferior to the refined curve.

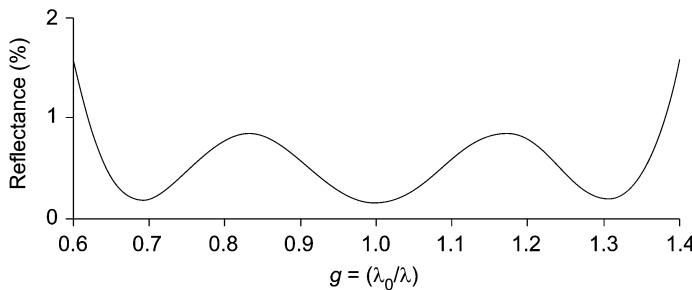


Figure 3.42. A five-layer design derived from figure 3.41 by replacing the half-wave layer by two quarter-wave layers and adjusting the values of the indices. Design:

Air	1.38	1.86	1.94	1.65	1.47	Glass
1.00	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	1.52.

Table 3.2.

Material	Index	Optical thickness (nm)
Air	1.00	Massive
MgF ₂	1.37	161
TiO ₂	2.28	78.5
MgF ₂	1.37	56.5
TiO ₂	2.28	54
Glass	1.52	Massive

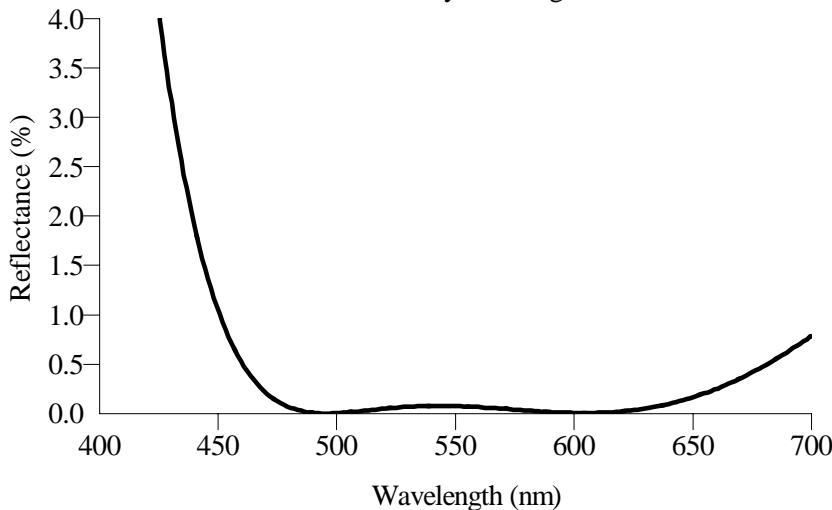
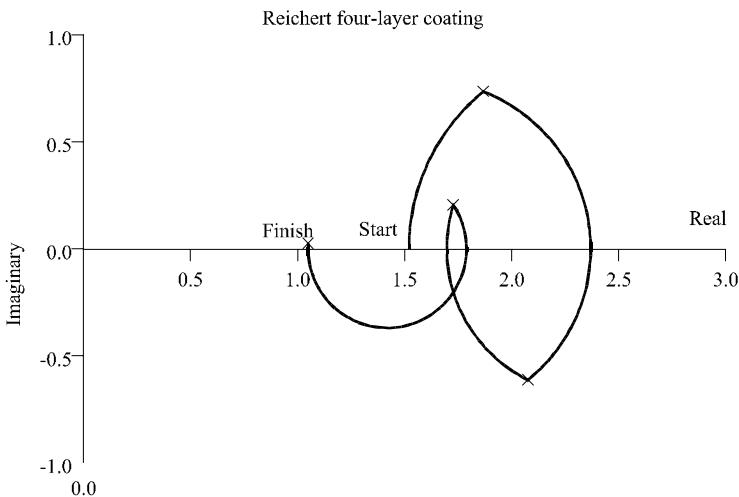
Although the Reichert design technique is not described, nevertheless it is a good exercise to attempt to understand how the coating functions. For this it is easiest if we simply draw an admittance diagram. Since the coating is clearly centred on 550 nm we draw the diagram for that wavelength.

The admittance diagram, figure 3.44, shows that the Reichert design can be considered as derived by applying two Vermeulen equivalents to the W-coat and its three-layer variations in figure 3.31. A particularly interesting feature of the Reichert coating is that it is quite thin compared with the W-coat from which it is derived. This double Vermeulen equivalent is a powerful replacement for a flattening half-wave in a design. We shall return to this structure later when we consider buffer layers.

3.3 Equivalent layers

There are great advantages in using a series of quarter-waves or multiples of quarter-waves in the first stages of the design of antireflection coatings because the

Reichert four-layer coating

**Figure 3.43.** The Reichert four-layer two-material antireflection coating.**Figure 3.44.** The admittance locus of the Reichert design at 550 nm.

characteristic curves of such coatings are symmetrical about $g = 1.0$. However, problems are presented in construction because the indices which are specified in this way do not often correspond exactly with indices which are readily available. Using mixtures of materials of higher and lower indices to produce

a layer of intermediate index is a technique which has been used successfully (see chapter 9), but a more straightforward method is to replace the layers by equivalent combinations involving only two materials, one of high index and one of low index. These two materials can be well-tried, stable materials, the characteristics of which have been established over many production runs in the plant that will be used for, and under the conditions that will apply to, the production of the coatings. To illustrate the method, we assume two materials of index 2.30 and 1.38, corresponding approximately to titanium dioxide and magnesium fluoride, respectively.

The first technique to mention is that of Vermeulen [15] which has already been referred to. It involves the replacing of a quarter-wave by a two-layer equivalent. The analysis is exactly that already given for the two-layer antireflection coating and it is assumed that the quarter-wave to be replaced has a locus which starts and terminates at predetermined points on the real axis. The replacement is, therefore, valid for the particular starting and terminating points used in its derivation only, and for that single wavelength for which the original layer is a quarter-wave. Under conditions which are increasingly remote from these ideal ones, the two-layer replacement becomes increasingly less satisfactory. It is advisable, when calculating the parameters of the layers, to sketch a rough admittance plot because otherwise there is a real danger of picking incorrect values of layer thickness. In the particular case we are considering, the starting admittance is 1.52 on the real axis and the terminating admittance is 1.9044, which will ensure that the outermost 1.38 index quarter-wave layer will terminate at the point 1.00 on the real axis. Clearly the high-index layer should be next to the substrate. The thicknesses are then, using equations (3.6) and selecting the appropriate pair of solutions, 0.05217 and 0.07339 full waves for the high- and low-index layers, respectively. We complete the design by adding a half-wave of index 2.30 and a quarter-wave of index 1.38. The characteristic curve of this coating is shown in figure 3.39, which, we recall, was arrived at in a completely different way.

As already mentioned, the four-layer Reichert coating, table 3.2, can be thought of as a Vermeulen equivalent of the coatings of figure 3.31. To obtain a replacement for a quarter-wave that does not depend on the starting point, we turn to a technique originated by Epstein [20] involving the symmetrical periods and the Herpin admittance mentioned briefly in chapter 2. We recall that any symmetrical combination of layers acts as a single layer with an equivalent phase thickness and equivalent optical admittance. In this particular application we consider combinations of the form ABA only. We choose for the indices of A and B those of the two materials from which the coating is to be constructed. Then for each quarter-wave layer of the coating we construct a three-layer symmetrical period which has an equivalent thickness of one quarter-wave and an equivalent admittance equal to that required from the original

To proceed further, we need expressions for the equivalent thickness and admittance of a symmetrical period. These are derived later in chapter 6. Since

the symmetrical period is of the form ABA, then

$$y_E = y_A$$

$$\times \left(\frac{\sin 2\delta_A \cos \delta_B + \frac{1}{2}[(y_B/y_A) + (y_A/y_B)] \cos 2\delta_A \sin \delta_B + \frac{1}{2}[(y_B/y_A) - (y_A/y_B)] \sin \delta_B}{\sin 2\delta_A \cos \delta_B + \frac{1}{2}[(y_B/y_A) + (y_A/y_B)] \cos 2\delta_A \sin \delta_B - \frac{1}{2}[(y_B/y_A) - (y_A/y_B)] \sin \delta_B} \right)^{1/2} \quad (3.22)$$

$$\cos \gamma = \cos 2\delta_A \cos \delta_B - \frac{1}{2}[(y_B/y_A) + (y_A/y_B)] \sin 2\delta_A \sin \delta_B, \quad (3.23)$$

where y_E is the equivalent optical admittance and γ is the equivalent phase thickness. The important feature of the symmetrical combination is that it behaves as a single layer of phase thickness γ and admittance y_E regardless of the starting point for the admittance locus.

In our particular case, the equivalent thickness of the combination should be a quarter-wave, that is

$$\begin{aligned} \cos \gamma &= \cos(\pi/2) = 0 \\ &= \cos 2\delta_A \cos \delta_B - \frac{1}{2}[(y_B/y_A) + (y_A/y_B)] \sin 2\delta_A \sin \delta_B \end{aligned}$$

which gives

$$\tan 2\delta_A \tan \delta_B = \frac{2y_A y_B}{y_A^2 + y_B^2}. \quad (3.24)$$

Substituting in equation (3.22) and manipulating the expression we have

$$y_E = y_A \left(\frac{1 + [(y_B^2 - y_A^2)/(y_B^2 + y_A^2)] \cos 2\delta_A}{1 - [(y_B^2 - y_A^2)/(y_B^2 + y_A^2)] \cos 2\delta_A} \right)^{1/2} \quad (3.25)$$

which yields

$$\cos 2\delta_A = \frac{(y_B^2 + y_A^2)(y_E^2 - y_A^2)}{(y_B^2 - y_A^2)(y_E^2 + y_A^2)}. \quad (3.26)$$

δ_B is given by equation (3.24), i.e.

$$\tan \delta_B = \frac{2y_A y_B}{y_A^2 + y_B^2} \cdot \frac{1}{\tan 2\delta_A} \quad (3.27)$$

and the optical thicknesses are then

$$\begin{aligned} \frac{n_A d_A}{\lambda_0} &= \frac{\delta_A}{2\pi} \text{ full waves at } \lambda_0 \\ \frac{n_B d_B}{\lambda_0} &= \frac{\delta_B}{2\pi} \text{ full waves at } \lambda_0. \end{aligned} \quad (3.28)$$

If an equivalent combination for a half-wave layer is required, then it is considered as two quarter-waves in series.

As an example of the application of this technique we take the four-layer coating of figure 3.32:

Air	1.38	2.13	1.9	1.38	Glass
	1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$

The layers which must be replaced are the quarter-waves with indices 2.13 and 1.90. There are two possible combinations, *HLH* or *LHL*, for each of these layers.

$$\begin{array}{c} 2.13 \\ 0.25\lambda_0 \end{array} \rightarrow \left\{ \begin{array}{c|c|c} 1.38 & 2.30 & 1.38 \\ 0.04128\lambda_0 & 0.15861\lambda_0 & 0.04128\lambda_0 \\ 2.30 & 1.38 & 2.30 \\ 0.11198\lambda_0 & 0.02302\lambda_0 & 0.11198\lambda_0 \end{array} \right. \\ \begin{array}{c} 1.90 \\ 0.25\lambda_0 \end{array} \rightarrow \left\{ \begin{array}{c|c|c} 1.38 & 2.30 & 1.38 \\ 0.06793\lambda_0 & 0.10438\lambda_0 & 0.06793\lambda_0 \\ 2.30 & 1.38 & 2.30 \\ 0.09216\lambda_0 & 0.05868\lambda_0 & 0.09216\lambda_0 \end{array} \right. \end{array}$$

As an indication of the closeness of fit between the symmetrical periods and the layers they replace, the variation, with g , of equivalent admittance and equivalent optical thickness is plotted in figure 3.45.

We can now replace the layers in the actual design of the antireflection coating. There are two possible replacements for each of the relevant layers, but where *HLH* and *LHL* combinations are mixed, there is a tendency towards an excessive number of layers in the final design, and so we consider two possibilities only, one based on *HLH* periods and one on *LHL*. These are shown in table 3.3.

The spectral characteristics of these coatings along with the original design are shown in figure 3.46. The replacements have a slightly inferior performance due to the effective dispersion that can be seen in figure 3.45. The process of design need not stop at this point, however, because the designs are excellent starting points for refinement. Figure 3.47 shows the performance of a refined version of one of the coatings. In practice, the refinement will include an allowance for the dispersion of the indices of the materials and there will be a certain amount of adjustment of the coating during the production trials.

If performance over a much wider region is required, then the apparent dispersion of the equivalent periods may become a problem. This dispersion can be reduced by using equivalent periods of 1/8-wave thickness instead of a quarter-wave. Each quarter-wave in the original design is then replaced by two periods in series. This adds considerably to the number of layers and the solution of the appropriate equations is no longer simple.

3.4 Antireflection coatings for two zeros

There are occasional applications where antireflection coatings are required which have zeros at certain well-defined wavelengths rather than over a wide spectral

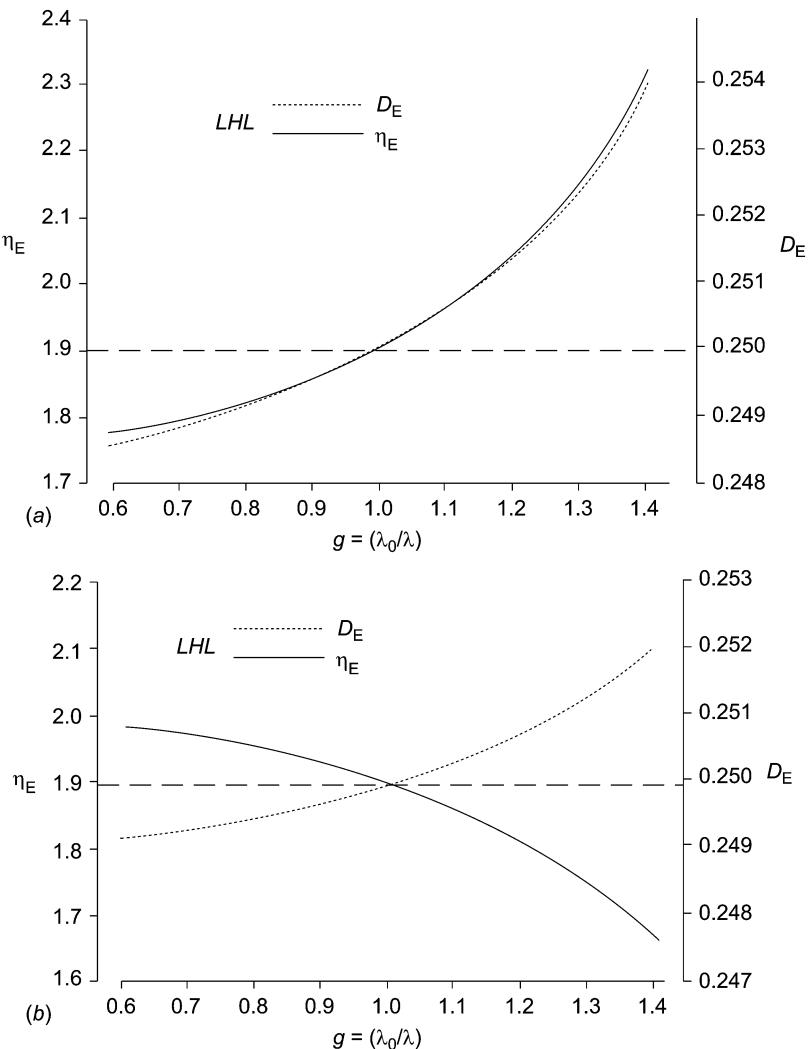


Figure 3.45. The equivalent admittances and optical thickness as a function of g ($= \lambda_0/\lambda$) of symmetrical period replacements for a single quarter-wave of index 1.90. The indices used in the symmetrical replacement are 2.30 for the high index and 1.38 for the low index. (a) LHL combination. (b) HLH combination. For a perfect match D_E and η_E should both be constant at $0.25\lambda_0$ and 1.9 respectively, whatever the value of g .

region. One of the most frequent of these applications is frequency doubling, where antireflection is required at two wavelengths, one of which is twice the other.

Table 3.3.

Layer number	Design based on <i>LHL</i> periods		Design based on <i>HLH</i> periods	
	Index	Thickness	Index	Thickness
0	1.0	Incident medium	1.0	Incident medium
1	1.38	$0.291\ 28\lambda_0$	1.38	$0.25\lambda_0$
2	2.30	$0.158\ 61\lambda_0$	2.30	$0.111\ 98\lambda_0$
3	1.38	$0.109\ 21\lambda_0$	1.38	$0.023\ 02\lambda_0$
4	2.30	$0.104\ 38\lambda_0$	2.30	$0.204\ 14\lambda_0$
5	1.38	$0.567\ 93\lambda_0$	1.38	$0.058\ 68\lambda_0$
6	1.52	Substrate	2.30	$0.092\ 16\lambda_0$
7			1.38	$0.5\lambda_0$
8			1.52	Substrate

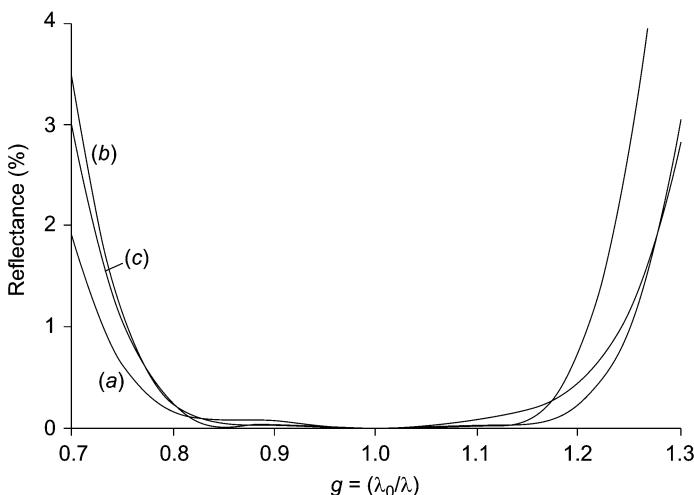


Figure 3.46. The performance of the designs of table 3.3. (a) Five-layer design based on *LHL* periods. (b) Seven-layer design based on *HLH* periods. (c) The original four-layer design from which (a) and (b) were derived.

The simplest coating that will satisfy this requirement is the quarter-quarter that has already been considered. We recall that the coating has two zeros at $\lambda = 3\lambda_0/4$ and $\lambda = 3\lambda_0/2$, just what is required. The conditions are

$$n_1 = (n_0^2 n_m)^{1/3} \quad (3.29)$$

$$n_2 = (n_0 n_m^2)^{1/3}.$$

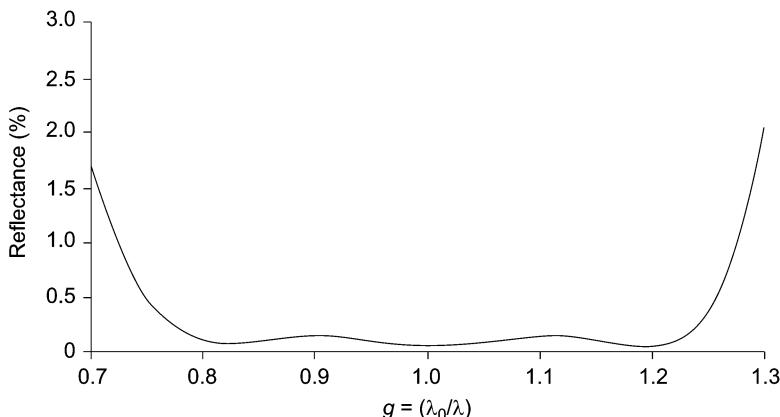


Figure 3.47. Refined version of the five-layer design of figure 3.46 and table 3.3. Design:

Air	1.38	2.30	1.38	2.30	1.38	Glass
1.00	$0.2973\lambda_0$	$0.1252\lambda_0$	$0.1244\lambda_0$	$0.0874\lambda_0$	$0.5597\lambda_0$	1.52.

The principal problem with this coating is once again the low-index substrate. With an index of 1.38 as the lowest value for n , the lowest value of substrate index that can be accommodated, from equation (3.1), is $1.38^3 = 2.63$. Thus the coating is suitable only for high-index substrates.

A common material that requires antireflection coatings at λ and 2λ is lithium niobate, which has an index of around 2.25. The quarter-quarter coating should have indices of 1.310 and 1.717. Indices of 1.38 and 1.717 give a reflection loss of 0.2%, which will probably be adequate for many applications, and indeed similar performance is obtained with any index between 1.7 and 1.8 for the high-index layer.

Should this performance be inadequate, then an additional layer can be added. Provided we keep to quarter-waves and multiples of quarter-waves, we retain the symmetry about $g = 1$ and therefore have to consider the performance at $g = 2/3$ only, since that at $g = 3/4$ will be automatically equivalent. From the point of view of the vector diagram, the problem with the quarter-quarter coating is ρ_a , the amplitude reflection coefficient from the first interface, which is too large. The vectors are inclined at 120° to each other and for zero reflectance they should be of equal length so that they form an equilateral triangle. If an extra quarter-wave n_3 is added, there will be four vectors and the fourth, ρ_d , will be along the same direction as ρ_a . If ρ_d is made to be of opposite sense to ρ_a , that is if $n_3 > n_m$, then it is possible to reduce the resultant of the two vectors to the same length as the other two. This can be achieved by the design

Air	1.38	1.808	2.368	Lithium niobate
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	2.25.

We can take 2.35, the index of zinc sulphide, for n_3 , and then any index in the range 1.75–1.85 for n_2 , to keep the minimum reflectance at $g = 2/3$ to below 0.1%.

There are many other possible arrangements. A coating with the first layer a half-wave, instead of a quarter-wave, can give a similar improvement, this time through a combination with ρ_c which means that $n_2 > n_3$. Here the ideal design is

Air	1.38	1.81	1.72	Lithium niobate
1.0	$0.5\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	2.25

and once again there is reasonable flexibility in the values of n_2 and n_3 if the aim is simply a reflectance of less than 0.1%. It is interesting to note the similarity between this coating and the quarter-quarter. The quarter-quarter has another zero at $g = 8/3$. If the inner quarter-waves in the above design were merged into a single half-wave of index around 1.75, then the coating would be identical with the quarter-quarter used at $g = 4/3$ and $g = 8/3$. Figure 3.48 shows the performance of these coatings.

This idea of using the fourth vector to trim the length of one of the other three so that a low reflectance is obtained can be extended to low-index substrates. The coating now, of course, departs considerably from the original quarter-quarter coating. A quarter-quarter-quarter design based on this approach is

Air	1.38	1.808	2.368	Glass
1.0	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	2.25

and its performance is shown in figure 3.49 where the monitoring wavelength has been assumed to be 707 nm and the two zeros are situated at 530 nm and 1.06 μm .

The method can be extended to four and even more quarter-waves, although the derivation of the final designs is very much more of a trial-and-error process because of the rather cumbersome expressions that cannot be reduced to explicit formulae for the various indices. Indeed, there are now too many parameters for there to be just one solution and the surplus can be used in an optimising process for broadening the reflectance minima. A number of interesting designs is given by Baumeister [21].

Mouchart [22] has also considered the derivation of antireflection coatings intended to eliminate reflection at two wavelengths. In coatings where all layers have thicknesses that are specified in advance to be multiples of a quarter-wave at $g = 1$, it is possible arbitrarily to choose the indices of all the layers except the final two, which can then be calculated from the values given to the others. The calculation involves the solution of an eighth-order equation that can be set up using expressions derived by Mouchart. The values of $\partial^2 R / \partial \lambda^2$ at the antireflection wavelength, which is inversely related to the bandwidth of the coating, can be used to assist in choosing the more promising designs from the enormous number that can be produced. Mouchart considers three-layer coatings of this type in some detail.

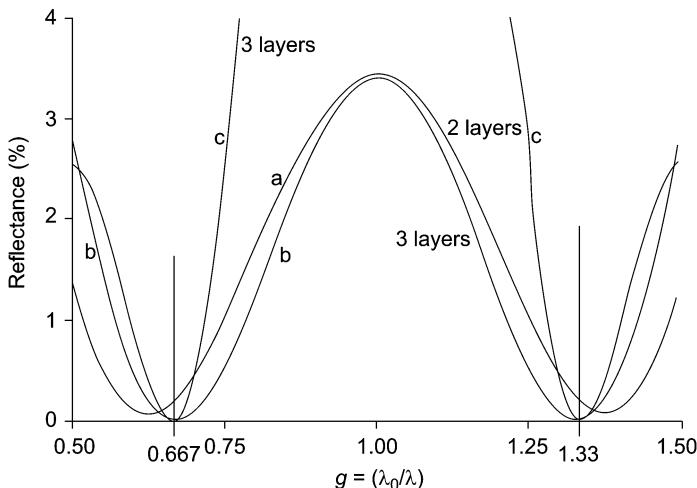


Figure 3.48. The performance of various two-zero 2:1 antireflection coatings on a high-index substrate such as lithium niobate with $n = 2.25$. The ideal positions for the two zeros are $g = 0.667$ and $g = 1.333$.

(a)	Air	1.38	1.72	Lithium niobate
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	2.25
(b)	Air	1.38	1.808	2.368
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$
				2.25
(c)	Air	1.38	1.81	1.72
	1.00	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$
				2.25.

3.5 Antireflection coatings for the visible and the infrared

There are frequent requirements for coatings that span the visible region and also reduce the reflectance at an infrared wavelength corresponding to a laser line. Such coatings are required in instruments where visual information and laser light share common elements, such as surgical instruments, surveying devices and the like. There are very many designs for such coatings and manufacturers seldom publish them. Design is largely a process of trial and error, and frequently the final operation is to replace the unobtainable or difficult indices by symmetrical combinations of better behaved materials and to refine the design so obtained to take account of the dispersion of the optical constants of real materials and to compensate for the apparent dispersion that occurs in connection with the symmetrical periods. In this section we consider the fundamental design process only, neglecting dispersion and in most cases retaining the ideal values of the index. We assume that the substrate is always glass of index 1.52 and that, as usual, the incident medium is air of index 1.0.

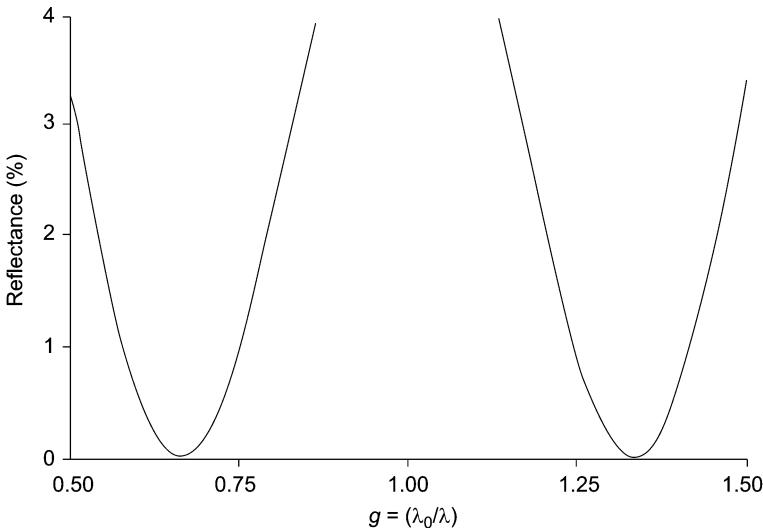


Figure 3.49. A three-layer two-zero 2:1 antireflection coating for a low-index substrate. Design ($\lambda_0 = 707$ nm):

Air	1.38	1.585	1.82	Glass
1.00	$0.25\lambda_0$	$0.25\lambda_0$	$0.25\lambda_0$	1.52.

The simplest type of coating that has low reflectance in the visible region and at a wavelength in the near infrared is a single layer of low- index material of thickness three quarter-waves. This has low reflectance at both λ_0 and $3\lambda_0$. Unfortunately, the lowest index, of 1.38, corresponding to magnesium fluoride, gives a residual reflectance of 1.25% at the minima and the performance in the visible region is rather narrower than that for the single quarter-wave coating, since the layer is three times thicker. The magnesium fluoride layer could be considered as an outer quarter-wave over an inner half-wave and a high-index half-wave flattening layer, of index 1.8, could be introduced between them giving the design:

$$\text{Air} | LHHLL | \text{Glass}.$$

Unfortunately, the half-wave layer, while it flattens the performance in the visible region, destroys the performance in the infrared at $3\lambda_0$, where it is two-thirds of a quarter-wave thick. The solution is to make the layer three half-waves thick in the visible, so that it is still a half-wave, and therefore an absentee, at $3\lambda_0$. The design then becomes:

$$\text{Air} | L6H2L | \text{Glass}$$

and the performance is shown in figure 3.50, where the reference wavelength is 510 nm. The performance in the visible region is indeed flattened in the normal

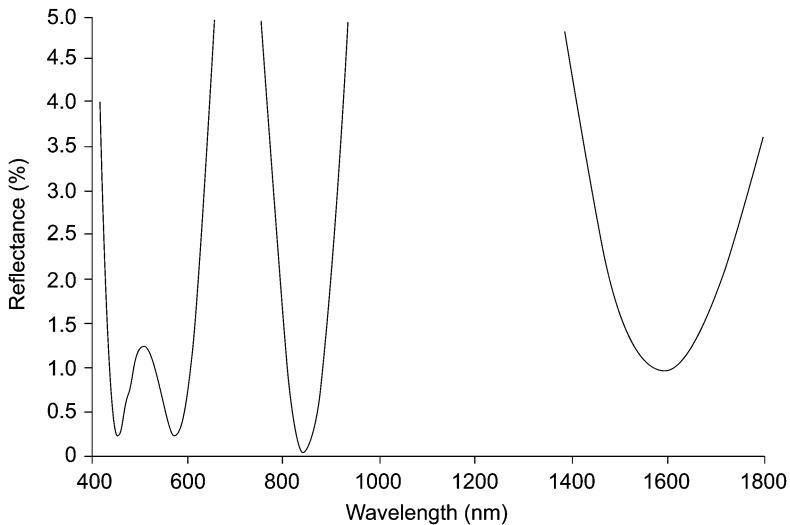


Figure 3.50. The performance of the coating:

Air (1.0) | $L6H2L$ | Glass (1.52)

with L a quarter-wave of index 1.38 and H of 1.8. λ_0 is 510 nm.

way, although, because the flattening layer is three times thicker than normal, the characteristic rises sharply in the blue and red regions. The minimum in the infrared around $1.53 \mu\text{m}$ is still present, although slightly skewed because of the half-wave layer. However, perhaps the most surprising feature is the appearance of a third and very deep minimum at 840 nm. We use the admittance diagram to help in understanding the origin of this dip.

Figure 3.51 shows the admittance diagram for the coating at the wavelength 840 nm. Layer 2, the 1.8 index layer, is almost two half-waves thick at this wavelength and so describes almost two complete revolutions, linking the ends of the loci of the two 1.38 index layers in such a way that almost zero reflectance is obtained. The loci of the two low-index layers are not very sensitive to changes in wavelength and therefore the position of the dip is fixed almost entirely by the high-index layer. Changes in its thickness will change the position of the dip. Making it thinner, 1.0 full waves instead of 1.5, for example, will move the dip to a longer wavelength. The performance characteristic of a coating of design

Air | $L4H2L$ | Glass

is shown in figure 3.52. The dip is now fairly near the desired wavelength of $1.06 \mu\text{m}$.

A coating that gives good performance over the visible region but has high

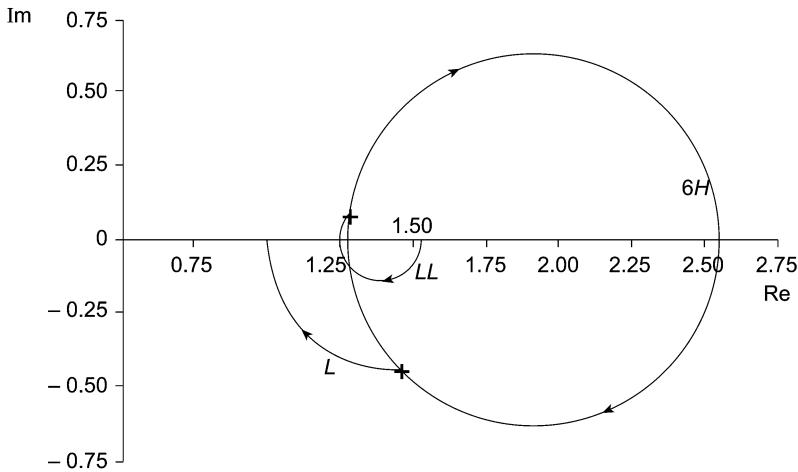


Figure 3.51. The admittance diagram for the coating of figure 3.50 at 840 nm, corresponding to the unexpected sharp zero, explains the occurrence of the dip.

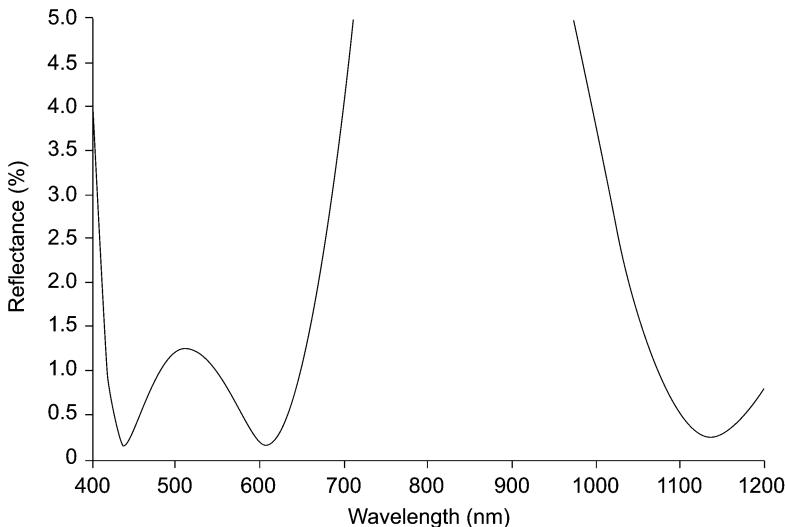


Figure 3.52. The performance of the coating:

Air (1.0) | $L4H2L$ | Glass (1.52)

with L a quarter-wave of index 1.38, H of 1.8 and reference wavelength, λ_0 , 510 nm. Note that the dip has moved to a longer wavelength than in figure 3.50.

reflectance at $1.06 \mu\text{m}$ is the quarter–half–quarter coating. The admittance diagram at λ_0 for such a coating is shown in figure 3.33. The locus intersects or crosses the real axis at the points 1.9 and 2.45. It is possible to insert layers of index 1.9 or 2.45, respectively, at these points in the design without any effect on the performance at λ_0 at all. The loci of these layers, whatever their thicknesses, would simply be points. Such layers are known as ‘buffer layers’ and were devised by Mouchart [23]. At the reference wavelength they exert no influence whatsoever but at other wavelengths, where the starting points of their loci move away from their reference wavelength positions, the loci appear in the normal way and can have important effects on performance. They are similar in some respects to half-wave layers that, by virtue of their precise thickness, are absentees at λ_0 but which have considerable influence on other wavelengths. The index can be chosen to sharpen or flatten a characteristic. The buffer layer has a precise value of index, but can have any thickness, which can be chosen to adjust performance at wavelengths other than λ_0 . Here we attempt to use buffer layers to alter the performance at $1.06 \mu\text{m}$. One buffer layer is not sufficient and we need to insert the two possible 1.9 index layers so that the design becomes:

$$\text{Air} \mid LB'HHB''N \mid \text{Glass}$$

where $y_L = 1.38$, $y_H = 2.15$ and $y_N = 1.70$. B' and B'' are buffer layers of admittance 1.9. Trial and error establishes thicknesses for B' of $0.342\lambda_0$, and for B'' of $0.084\lambda_0$. However, although the reflectance at $1.06 \mu\text{m}$ is reduced considerably, the buffer layers do distort the performance characteristic somewhat in the visible region (figure 3.53) and only by refining the design is a completely satisfactory performance obtained. The final design, also illustrated in figure 3.53, is:

$$\begin{array}{c|c|c|c|c|c|c|c} \text{Air} & 1.38 & 1.90 & 2.15 & 1.90 & 1.70 \\ \hline 1.00 & 0.2667\lambda_0 & 0.3085\lambda_0 & 0.5395\lambda_0 & 0.1316\lambda_0 & 0.1796\lambda_0 \end{array} \mid \text{Glass.}$$

Many of the designs currently used for the visible and $1.06 \mu\text{m}$ involve just two materials of high and low index. Designs of this type can be arrived at in a number of ways. The arrangements above that use ideal layers can be replaced by symmetrical periods in the way already discussed. This type of design is seldom immediately acceptable because the very wide wavelength range makes it difficult to match exactly the layers with symmetrical periods and they are therefore usually refined by computer.

Figure 3.54 shows the performance of a six-layer design arrived at by computer synthesis:

$$\begin{array}{c|c|c|c|c|c|c|c} \text{Air} & 1.38 & 2.25 & 1.38 & 2.25 & 1.38 & 2.25 \\ \hline 1.00 & 0.3003\lambda_0 & 0.1281\lambda_0 & 0.0657\lambda_0 & 0.6789\lambda_0 & 0.0718\lambda_0 & 0.0840\lambda_0 \end{array} \mid \text{Glass.}$$

Buffer layers are very useful in such coatings. Half-wave absentee layers correct performance rapidly as the wavelength moves from that for which they

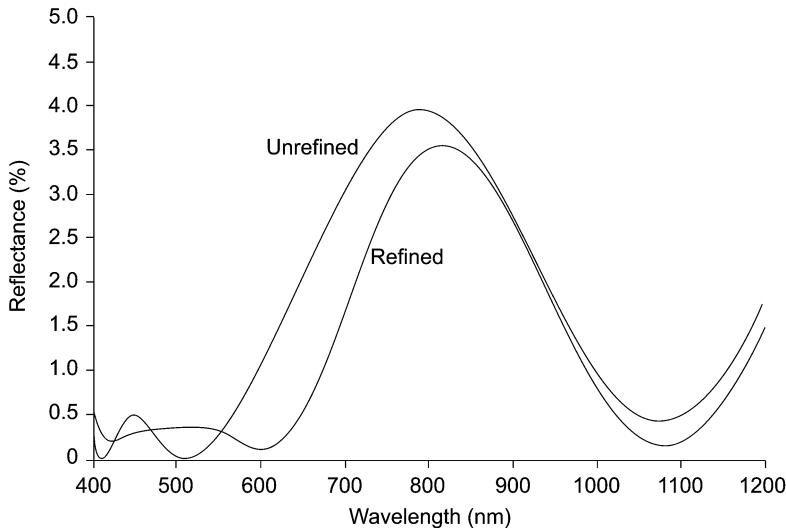


Figure 3.53. The performance of the design:

$$\text{Air (1.0)} \mid LB'HHB''M \mid \text{Glass (1.52)}$$

with L , H , M quarter-waves of indices 1.38, 2.15 and 1.70 respectively. B' and B'' are buffer layers of index 1.9 (see text) and thicknesses $0.342\lambda_0$ and $0.084\lambda_0$, respectively. λ_0 is 510 nm. The design has also been refined to yield the second performance curve. The refined design is given in the text.

are half-waves. Buffer layers react more slowly and therefore are very helpful when reflectance must remain low over a wide spectral region. The difficulty with buffer layers is that their refractive index is fixed by the axis crossings of the admittance locus of the coating in which they are to be inserted. We normally have a limited set of indices corresponding to the particular materials we are using and, in order to employ such layers as buffers, we must engineer an axis crossing at the appropriate value of admittance. The double Vermeulen structure makes this possible. In figure 3.44, the axis crossing on the extreme right can be moved simply by adjusting the thicknesses of the layers making up the structure. It is straightforward to arrange that the axis crossing should actually coincide with the index of the high-index layer already used in the design. This has been achieved with the first of the designs in table 3.4. Note that the thicknesses are optical so that they can be directly compared with those in table 3.2.

Figure 3.55 shows the admittance locus of the adjusted coating. The axis crossing has been arranged and the final three layers of the design have been adjusted to give good performance over the visible region. The performance of the coating is shown in grey in figure 3.56. The design is given in the first design

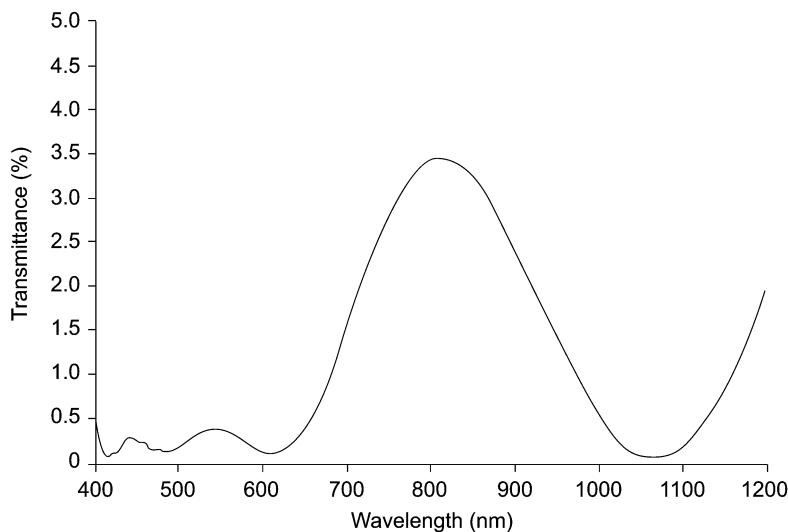


Figure 3.54. The performance of a six-layer design of antireflection coating for the visible region and $1.06 \mu\text{m}$, arrived at purely by computer synthesis. The reference wavelength is 510 nm and the design is given in the text.

Table 3.4.

Material	Index	Starting design Optical thickness (nm)	With buffer Optical thickness (nm)	With buffer and absentee Optical thickness (nm)
Air	1.00	Massive	Massive	Massive
MgF_2	1.37	154.47	154.47	140.80
TiO_2	2.28	57.96	57.96	50.70
MgF_2	1.37	22.66	22.66	17.46
TiO_2	2.28	—	247.50	240.84
MgF_2	1.37	35.06	35.06	44.31
TiO_2	2.28	49.23	49.23	39.99
MgF_2	1.37	—	—	294.54
Glass	1.52	Massive	Massive	Massive

column of table 3.4. Then the buffer layer of TiO_2 is added and the appearance of the admittance locus does not change with buffer layer thickness. Adjustment of the buffer layer by trial and error gives the improvement shown in figure 3.56.

Addition of a half-wave layer of low index between the coating and the glass substrate followed by refinement of all layers yields the performance shown in

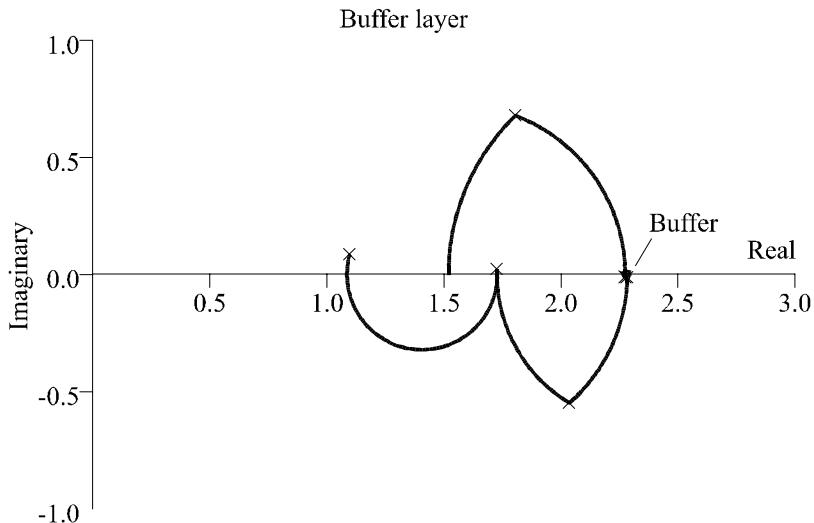


Figure 3.55. The admittance locus of the adjusted coating showing the axis crossing at 2.28. A buffer layer has been inserted there.

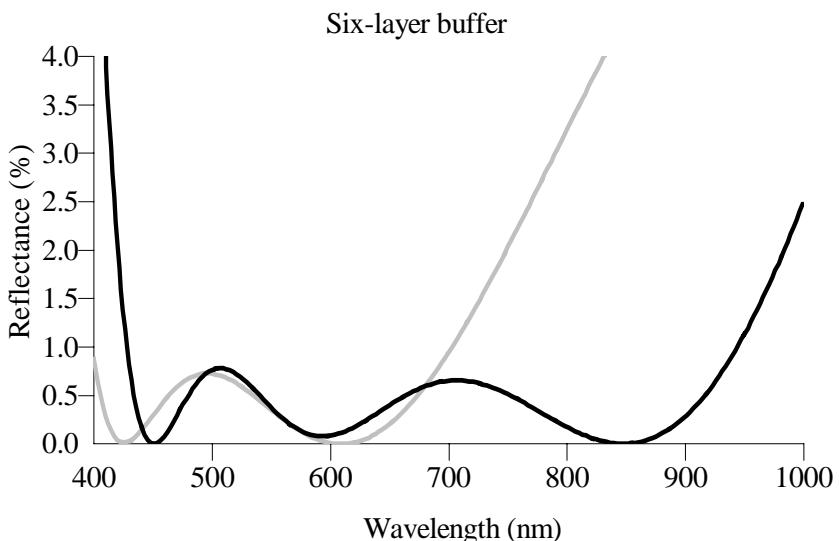


Figure 3.56. The starting four-layer coating performance is shown in grey. The addition of the buffer layer makes the coating into a six-layer system. Adjustment of the buffer layer thickness until just less than a half-wave gives the performance shown by the black line. The designs are given in table 3.4.

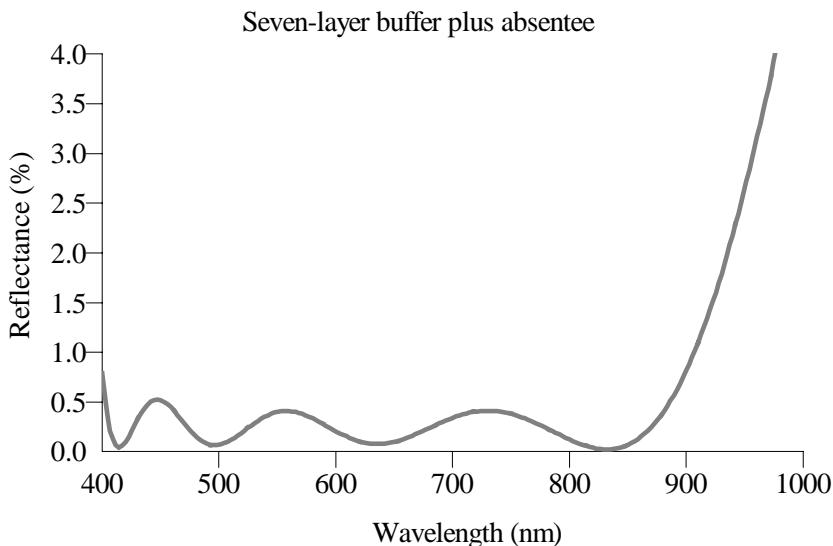


Figure 3.57. Performance of the seven-layer coating contained in table 3.4.

figure 3.57. This is as good a performance as we are likely to get with seven layers of the given indices. Significant improvement in performance demands more layers.

The major determinant of antireflection-coating performance for low-index substrates is the lowest index of refraction of the design materials. Magnesium fluoride is the usual choice but, unfortunately, it is not ideal. It suffers from high tensile stress and for reasonable durability must be deposited on a heated substrate. Silicon dioxide is much tougher and more stable and would be preferred over magnesium fluoride were it not for the fact that the refractive index is rather higher at around 1.45 compared with magnesium fluoride's 1.38. In multilayer coatings therefore it is quite common practice to use silicon dioxide as the low-index material through the coating but to continue to use magnesium fluoride as the outermost layer. The layer next to the air is critical. The layers distributed within the coating are less so.

3.6 Inhomogeneous layers

Inhomogeneous layers are ones in which the refractive index varies through the thickness of the layer. As we shall see in chapter 9, many of the thin-film materials which are commonly used give films that are inhomogeneous. This inhomogeneity is often quite small and the layers can safely be treated as if they were homogeneous in all but the most precise and exacting coatings. There is, however, a number of films which show sufficient inhomogeneity to affect the

performance of an antireflection coating perceptibly. If such a layer is used instead of a homogeneous one in a well-corrected antireflection coating then a reduction in performance is the normal result. Provided the inhomogeneity is not large, an adjustment of the indices of the other layers is usually sufficient correction and, as Ogura [24] has pointed out, an index that decreases slightly with thickness associated with the high-index layer in the quarter-half-quarter coating can actually broaden the characteristic. Zirconium oxide is a much used material which exhibits an index which increases with film thickness when deposited at room temperature, but decreases with thickness when deposited at substrate temperatures above 200 °C. Vermeulen [25] has considered the effect of the inhomogeneity of zirconium oxide on the quarter-half-quarter coating and has shown how it is possible to correct for the inhomogeneity by varying the index of the intermediate-index layer which, for virtually complete compensation, should be of the two-layer composite type [15] already referred to in this chapter. This type of inhomogeneity is one which is intrinsic and relatively small. By arranging for the evaporation of mixtures of composition varying with film thickness it is possible to produce layers which show an enormous degree of inhomogeneity and which permit the construction of entirely new types of antireflection coating.

Accurate calculation techniques for such layers are reviewed by Jacobsson [26] and by Knittl [27]. The simplest method involves the splitting of the inhomogeneous layer into a very large number of thin sublayers. Each sublayer is then replaced by a homogeneous layer of the same thickness and mean refractive index so that the smoothly varying index of the inhomogeneous layer is represented by a series of small steps. Computation can then be carried out as for a multilayer of homogeneous layers. There is no difficulty, with modern computers, in accommodating very large numbers of sub-layers so that, although an approximation, the method can be made to yield results identical for all practical purposes with those which would have been obtained by exact calculation (in cases where exact calculation techniques exist).

For our purposes, we can approach the theory of such coatings from the starting point of the multilayer antireflection coating for high-index substrates. As more and more layers are added to the coating, the performance, both from the bandwidth and the maximum reflectance in the low-reflectance region, steadily improves. In the limit, there will be an infinite number of layers with infinitesimal steps in optical admittance from one layer to the next. If, as layers are added, the total optical thickness of the multilayer is kept constant, the thickness of the individual layers will tend to zero and the multilayers will become indistinguishable from a single layer of identical optical thickness, but with optical admittance varying smoothly from that of the substrate to that of the incident medium.

If there are n layers in the multilayer, then the total optical thickness of the coating will be $n\lambda_0/4$ which may be denoted by T . There will be n zeros of

reflectance extending from a shortwave limit

$$\lambda_S = \left(\frac{(n+1)}{n} \right) \frac{\lambda_0}{2}$$

to a longwave limit

$$\lambda_L = [(n+1)] \frac{\lambda_0}{2}.$$

In terms of T , the total optical thickness, these limits are

$$\begin{aligned}\lambda_S &= \left(\frac{2(n+1)}{n^2} \right) T \\ \lambda_L &= \left(\frac{2(n+1)}{n} \right) T.\end{aligned}$$

At wavelengths of $2\lambda_L$ or longer, the arrows in the vector diagram are confined to the third and fourth quadrant so that the antireflection coating is no longer effective.

If now n tends to infinity but T remains finite, the multilayer tends to a single inhomogeneous layer, λ_S tends to zero, and λ_L tends to $2T$. For all wavelengths between these limits the reflectance of the assembly is zero. Thus the inhomogeneous film with smoothly varying refractive index is a perfect antireflection coating for all wavelengths shorter than twice the optical thickness of the film. At wavelengths longer than this limit the performance falls off, and at the wavelength given by four times the optical thickness of the film, the coating is no longer effective.

Of course, in practice there is no useful thin-film material with refractive index as low as unity and any inhomogeneous thin film must terminate with an index of around 1.35, say, which, in the infrared, is the index of magnesium fluoride. The reflectance of such a coated component will be equal to that of a plate of magnesium fluoride, 2.2% per surface.

Jacobsson and Martensson have actually produced an inhomogeneous antireflection coating of this type on a germanium plate [28]. The films were manufactured by the simultaneous evaporation of germanium and magnesium fluoride, the relative proportions of which were varied throughout the deposition to give a smooth transition between the indices of the two materials. An example of the performance attained is shown in figure 3.58. For this particular coating the physical thickness is quoted as $1.2 \mu\text{m}$. To find the optical thickness we assume that the variation of refractive index with physical thickness is linear (mainly because any other assumed law of variation would lead to very difficult calculations, although possibly more accurate results). The optical thickness is then given by the physical thickness times the mean of the two terminal indices. For this present film, starting with an index of 4.0 and finishing with 1.35, the mean is 2.68 and the optical thickness, therefore, $2.68 \times 1.2 \mu\text{m}$, i.e. $3.2 \mu\text{m}$.

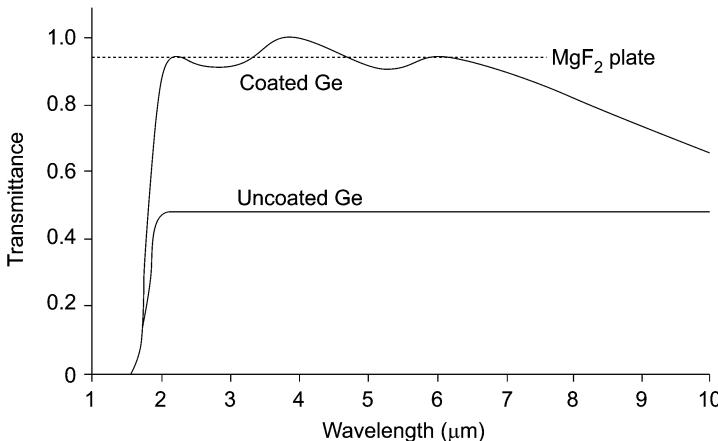


Figure 3.58. Measured transmittance of a germanium plate coated on both sides with an inhomogeneous Ge–MgF₂ film with geometrical thickness 1.2 μm. (After Jacobsson and Martensson.)

This implies that the coating should give excellent antireflection for wavelengths out to 6.4 μm, after which it should show a gradually reducing transmission until a wavelength of $4 \times 3.2 \mu\text{m}$, i.e. 12.8 μm. The curve of the coated component in figure 3.58 shows that this is indeed the case.

Berning [29] has suggested the use of the Herpin index concept for the design of antireflection coatings which are composed of homogeneous layers of two materials, one of high index and the other of low index, which are step approximations to the inhomogeneous layer and which, because they involve homogeneous layers of well-understood and stable materials, might be easier to manufacture than the ideal inhomogeneous layers. He has suggested designs for the antireflection coating of germanium consisting of up to 39 alternate layers of germanium and magnesium fluoride equivalent to 20 quarter-waves of gradually decreasing index.

As with coatings consisting of homogeneous layers, the most serious limitation is the lack of low-index materials. A single inhomogeneous layer to match a substrate to air must terminate at an index of around 1.38, which means that the best that can be done with such a layer is a residual reflectance of 2.5%. This limits their direct use to high-index substrates. For low-index substrates it is likely that their role will remain in the improvement of the performance of designs incorporating homogeneous materials.

3.7 Further information

It has not been possible in a single chapter in this book to cover completely the field of antireflection coatings. Further information will be found in Cox and Hass [17] and Musset and Thelen [6]. There is also a very useful account of antireflection coatings in Knittl [27] which contains some alternative techniques.

References

- [1] Cox J T and Hass G 1958 Antireflection coatings for germanium and silicon in the infrared *J. Opt. Soc. Am.* **48** 677–80
- [2] Catalan L A 1962 Some computed optical properties of antireflection coatings *J. Opt. Soc. Am.* **52** 437–40
- [3] Schuster K 1949 Anwendung der Vierpoltheorie auf die Probleme der optischen Reflexionsminderung, Reflexionsverstärkung, und der Interferenzfilter *Ann. Phys.* **4** 352–6
- [4] Cox J T 1961 Special type of double-layer antireflection coefficient for infrared optical materials with high refractive index *J. Opt. Soc. Am.* **51** 1406–8
- [5] Cox J T, Hass G and Jacobus G F 1961 Infrared filters of antireflected Si, Ge, InAs and InSb *J. Opt. Soc. Am.* **51** 714–18
- [6] Musset A and Thelen 1966 Multilayer antireflection coatings *Progress in Optics* ed E Wolf (Amsterdam: North Holland) pp 201–37
- [7] Thelen A 1969 Design of multilayer interference filters *Physics of Thin Films* ed G Hass and R E Thun (New York: Academic) pp 47–86
- [8] Young L 1961 Synthesis of multiple antireflection films over a prescribed frequency band *J. Opt. Soc. Am.* **51** 967–74
- [9] Turbadar T 1964 Equireflectance contours of double layer antireflection coatings *Opt. Acta* **11** 159–70
- [10] Thetford A 1969 A method of designing three-layer antireflection coatings *Opt. Acta* **16** 37–44
- [11] Thetford A 1968 *Four-Layer Coating Design* Private communication (University of Reading)
- [12] Lockhart L B and King P 1947 Three-layered reflection-reducing coatings *J. Opt. Soc. Am.* **37** 689–94
- [13] Cox J T, Hass G and Thelen A 1962 Triple-layer antireflection coating on glass for the visible and near infrared *J. Opt. Soc. Am.* **52** 965–9
- [14] Turbadar T 1964 Equireflectance contours of triple-layer antireflection coatings *Opt. Acta* **11** 195–205
- [15] Vermeulen A J 1971 Some phenomena connected with the optical monitoring of thin-film deposition and their application to optical coatings *Opt. Acta* **18** 531–8
- [16] Shadbolt M J 1967 *Measured Results of Four-Layer Antireflection Coating Deposition* Private communication (Sira Institute, Chislehurst, Kent)
- [17] Cox J T and Hass G 1964 Antireflection coatings *Physics of Thin Films* ed G Hass and R E Thun (New York: Academic) pp 239–304
- [18] Ward J 1972 Towards invisible glass *Vacuum* **22** 369–75
- [19] C Reichert Optische Werke AG 1962 *Improvements in or Relating to Optical Components Having Reflection-Reducing Coatings* UK Patent 991 635

- [20] Epstein L I 1952 The design of optical filters *J. Opt. Soc. Am.* **42** 806–10
- [21] Baumeister P W, Moore R and Walsh K 1977 Application of linear programming to antireflection coating design *J. Opt. Soc. Am.* **67** 1039–45
- [22] Mouchart J 1978 Thin film optical coatings. 6: Design method for two given wavelength antireflection coatings *Appl. Opt.* **17** 1458–65
- [23] Mouchart J 1978 Thin film optical coatings. 5: Buffer layer theory *Appl. Opt.* **17** 72–5
- [24] Ogura S 1975 Some features of the behaviour of optical thin films *PhD Thesis* (Newcastle upon Tyne Polytechnic)
- [25] Vermeulen A J 1976 Influence of inhomogeneous refractive indices in multilayer anti-reflection coatings *Opt. Acta* **23** 71–9
- [26] Jacobsson R 1975 Inhomogeneous and coevaporated homogeneous films for optical applications *Phys. Thin Films* **8** 51–98
- [27] Knittl Z 1976 *Optics of Thin Films* (London: Wiley)
- [28] Jacobsson R and Martensson J O 1966 Evaporated inhomogeneous thin films *Appl. Opt.* **5** 29–34
- [29] Berning P H 1962 Use of equivalent films in the design of infrared multilayer antireflection coatings *J. Opt. Soc. Am.* **52** 431–6

Chapter 4

Neutral mirrors and beam splitters

4.1 High-reflectance mirror coatings

Almost as important as the transmitting optical components of the previous chapter are those whose function is to reflect a major portion of the incident light. In the vast majority of cases the sole requirement is that the specular reflectance should be as high as conveniently possible, although, as we shall see, there are specialised applications where not only should the reflectance be high, but also the absorption should be extremely low. For mirrors in optical instruments, simple metallic layers usually give adequate performance and these will be examined first. For some applications where the reflectance must be higher than can be achieved with simple metallic layers, their reflectance can be boosted by the addition of extra dielectric layers. Multilayer all-dielectric reflectors, which combine maximum reflectance with minimum absorption, and which transmit the energy which they do not reflect, are reserved for the next chapter.

4.1.1 Metallic layers

The performance of the commonest metals used as reflecting coatings is shown [1] in figure 4.1.

Aluminium is easy to evaporate and has good ultraviolet, visible and infrared reflectance, together with the additional advantage of adhering strongly to most substances, including plastics. As a result it is the most frequently used film material for the production of reflecting coatings. The reflectance of an aluminium coating does drop gradually in use, although the thin oxide layer, which always forms on the surface very quickly after coating, helps to protect it from further corrosion. In use, especially if the mirror is at all exposed, dust and dirt invariably collect on the surface and cause a fall in reflectance. The performance of most instruments is not seriously affected by a slight drop in reflectance, but in some cases where it is important to collect the maximum amount of light, as it is difficult to clean the coatings without damaging them, the components

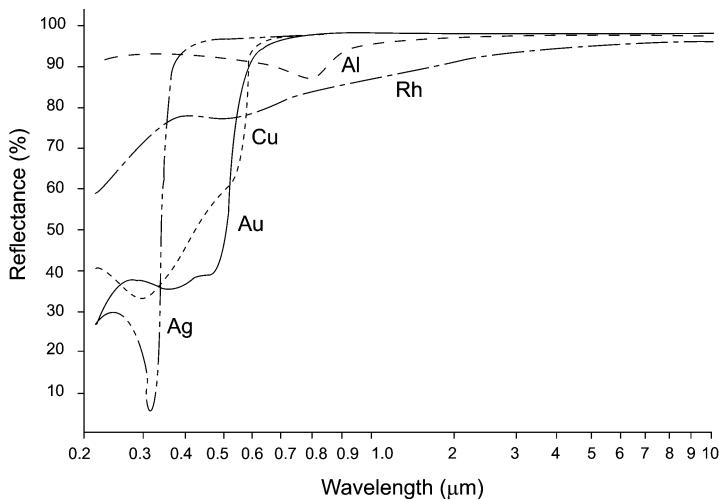


Figure 4.1. Reflectance of freshly deposited films of aluminium, copper, gold, rhodium and silver as a function of wavelength from 0.2–10 μm (After Hass [1].)

are recoated periodically. This applies particularly to the mirrors of large astronomical reflecting telescopes. The primary mirrors of these are recoated with aluminium usually around once a year in coating plants which are installed in the observatories for this purpose. Because the primaries are very large and heavy (for example, the 98-inch primary of the Isaac Newton Memorial Telescope of the Royal Greenwich Observatory weighs some 9000 lb), it is not usual to rotate them during coating and the uniformity of coating is achieved through the use of multiple sources.

Silver was once the most popular material of all. It does tarnish when exposed to the atmosphere, owing mainly to the formation of silver sulphide, but the initial high reflectance and the extreme ease of evaporation still make it a common choice for components used only for a short period of time. Silver is also often used where it is necessary to coat temporarily a component, such as an interferometer plate, for a test of flatness.

Gold is probably the best material for infrared reflecting coatings. Its reflectance drops off rapidly in the visible region and it is really useful only beyond 700 nm. On glass, gold tends to form rather soft, easily damaged films, but it adheres strongly to a film of chromium or Nichrome, and this is often used as an underlayer between the gold and the glass substrate.

The reflectance of rhodium and platinum is much less than that of the other metals mentioned and these metals are used only where stable films very resistant to corrosion are required. Both materials adhere very strongly to glass.

4.1.2 Protection of metal films

Most metal films are rather softer than hard dielectric films and can be scratched easily. Unprotected evaporated aluminium layers, for example, can be badly damaged if wiped with a cloth, while gold and silver films are even softer. This is a serious disadvantage, especially when periodic cleaning of the mirrors is necessary. One solution, as we have seen, is periodic recoating. An alternative, which improves the ruggedness of the coatings and also protects them from atmospheric corrosion, is overcoating with an additional dielectric layer. The behaviour of a single dielectric layer on a metal is a useful illustration of the calculation techniques of chapter 2. We shall also require some related results later and so it is useful to spend a little time on the problem.

First of all, the admittance diagram (figure 4.2) gives us a qualitative picture of the behaviour of the system as the dielectric layer is added. The metal layer will normally be thick enough for the optical admittance at its front surface to be simply that of the metal, the substrate optical constants having no effect. The optical admittance of the metal will always be in the fourth quadrant and so, as a dielectric layer is added, the reflectance must fall until the locus of the admittance of the assembly crosses the real axis. (The reflectance associated with the locus of a dielectric layer of index higher than the incident medium always falls as the locus is traced out in the fourth quadrant and always rises in the first—figure 2.11(a).) This minimum of reflectance will occur at a dielectric layer thickness of less than a quarter-wave. For layer thicknesses of up to twice this figure, therefore, the reflectance of the protected metal film will be reduced. The reduction in reflectance depends very much on the particular metal and the index of the dielectric film.

We can mark the position of the quarter-wave dielectric layer thickness by a simple construction. We draw the line from the origin to the starting point of the dielectric locus, that is the metal admittance $(\alpha, -\beta)$ which lies in the fourth quadrant. This line makes an angle θ with the real axis. Then, also through the origin, we draw a line in the first quadrant making the same angle θ with the real axis. This cuts the dielectric locus in two points. One is the point (α, β) , the image of the starting point in the real axis, and at this point the reflectance of the assembly is identical to that of the uncoated metal. The second point of intersection is

$$\left(\frac{\eta_f^2 \alpha}{(\alpha^2 + \beta^2)}, \frac{\eta_f^2 \beta}{(\alpha^2 + \beta^2)} \right) \quad \text{i.e.} \quad \frac{\eta_f^2}{\alpha - i\beta}$$

and at this point the layer is one quarter-wave thick.

We can derive straightforward analytical expressions for the various parameters, and, in particular, the points of intersection of the locus with the real axis, which we know correspond to the points of maximum and minimum reflectance.

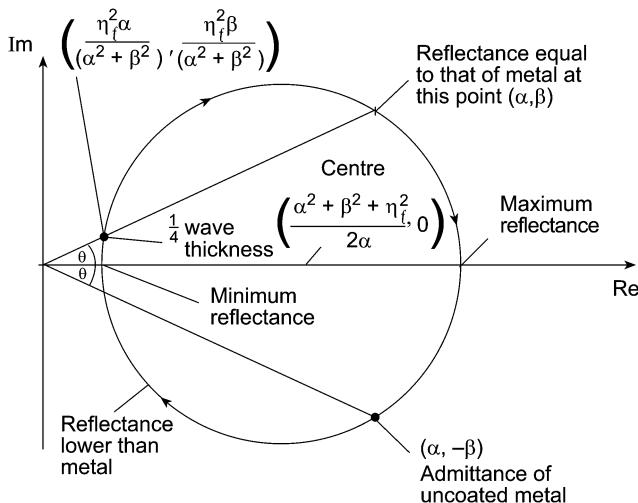


Figure 4.2. Admittance diagram of a dielectric layer deposited over a metal. The metal admittance would usually be much closer to the imaginary axis but has been moved for greater clarity in the diagram. The dielectric locus starts at the admittance of the uncoated metal. The construction to find the quarter-wave point is explained in the text, as are the other parameters.

The characteristic matrix is given by

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_f & i(\sin \delta_f / \eta_f) \\ i\eta_f \sin \delta_f & \cos \delta_f \end{bmatrix} \begin{bmatrix} \alpha - i & 1 \\ \beta & \end{bmatrix} \quad (4.1)$$

where $\alpha - i\beta$ is the characteristic admittance of the metal, i.e. $\mathcal{Y}(n_m - ik_m)$ at normal incidence, $\delta_f = 2\pi n_f d_f \cos \theta_f / \lambda$, and η_f is the characteristic admittance of the film material. Then

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_f + (\beta \sin \delta_f) / \eta_f + i(\alpha \sin \delta_f) / \eta_f \\ \alpha \cos \delta_f + i(\eta_f \sin \delta_f - \beta \cos \delta_f) \end{bmatrix}.$$

Now, at the points of intersection of the locus with the real axis, we must have that the admittance, which we can denote by μ , is real. But

$$\mu = C/B$$

and, equating real and imaginary parts,

$$\alpha \cos \delta_f = \mu [\cos \delta_f + (\beta \sin \delta_f) / \eta_f] \quad (4.2)$$

$$\eta_f \sin \delta_f - \beta \cos \delta_f = \mu (\alpha \sin \delta_f) / \eta_f. \quad (4.3)$$

Hence, first eliminating μ ,

$$(\alpha \cos \delta_f)(\alpha \sin \delta_f) / \eta_f = (\eta_f \sin \delta_f - \beta \cos \delta_f)[\cos \delta_f + (\beta \sin \delta_f) / \eta_f]$$

i.e.

$$[(\alpha^2 + \beta^2 - \eta_f^2)/(2\eta_f)] \sin(2\delta_f) = -\beta \cos(2\delta_f).$$

Thus

$$\tan(2\delta_f) = 2\beta\eta_f/(\eta_f^2 - \alpha^2 - \beta^2)$$

so that

$$\delta_f = \frac{1}{2} \tan^{-1}[2\beta\eta_f/(\eta_f^2 - \alpha^2 - \beta^2)] + \frac{m\pi}{2} \quad m = 0, 1, 2, 3 \dots \quad (4.4)$$

or, in full waves,

$$D_f/\lambda_0 = (1/4\pi) \tan^{-1}[2\beta\eta_f/(\eta_f^2 - \alpha^2 - \beta^2)] + m/4 \quad (4.5)$$

where the arctangent is to be taken in either the first or second quadrant so that δ_f for $m = 0$ is positive and represents the first intersection with the real axis where the film is less than, or at the very most, equal to a quarter-wave. A similar result has been derived by Park [2] using a slightly different technique.

The value of μ can be found by rearranging equations (4.2) and (4.3) slightly:

$$\begin{aligned} (\mu - \alpha) \cos \delta_f + (\beta\mu/\eta_f) \sin \delta_f &= 0 \\ \beta \cos \delta_f + [(\mu\alpha/\eta_f) - \eta_f] \sin \delta_f &= 0 \end{aligned}$$

and, eliminating δ_f ,

$$(\mu - \alpha)[(\mu\alpha/\eta_f) - \eta_f] - \beta(\beta\mu/\eta_f) = 0.$$

The two solutions are

$$\mu = [(\alpha^2 + \beta^2 + \eta_f^2)/2\alpha] \pm \{[(\alpha^2 + \beta^2 + \eta_f^2)/4\alpha^2] - \eta_f^2\}^{1/2}$$

but this is not the best form for calculation. We know that the two solutions μ_1 and μ_2 are related by $\mu_1\mu_2 = \eta_f^2$ and so we write

$$\mu_1 = 2\alpha\eta_f^2/[(\alpha^2 + \beta^2 + \eta_f^2) + [(\alpha^2 + \beta^2 + \eta_f^2)^2 - 4\alpha^2\eta_f^2]^{1/2}] \quad (4.6)$$

$$\mu_2 = [(\alpha^2 + \beta^2 + \eta_f^2)/2\alpha] + \{[(\alpha^2 + \beta^2 + \eta_f^2)/4\alpha^2] - \eta_f^2\}^{1/2} \quad (4.7)$$

and the value which corresponds to the first intersection ($m = 0$ in equation (4.4)) is

$$\mu_1 = 2\alpha\eta_f^2/[(\alpha^2 + \beta^2 + \eta_f^2) + [(\alpha^2 + \beta^2 + \eta_f^2)^2 - 4\alpha^2\eta_f^2]^{1/2}]. \quad (4.6)$$

Often

$$(\alpha^2 + \beta^2 + \eta_f^2)^2 \gg 4\alpha^2\eta_f^2$$

Table 4.1.

Aluminium (0.82 – i5.99)	R_{uncoated} (%)	R_{min} (%)	D_{min} (Full waves)	R_{max} (%)	D_{max} (Full waves)
Quartz (1.45)	91.63	83.64	0.2128	91.86	0.4628
CeO ₂ (2.30)	91.63	65.90	0.1925	92.44	0.4425

and in that case

$$\mu_1 = \alpha\eta_f^2/(\alpha^2 + \beta^2 + \eta_f^2) \quad (4.8)$$

$$\mu_2 = (\alpha^2 + \beta^2 + \eta_f^2)/\alpha. \quad (4.9)$$

The limits of reflectance are given by

$$R_{\text{minimum}} = [(\eta_0 - \mu_1)/(\eta_0 + \mu_1)]^2 \quad (4.10)$$

$$R_{\text{maximum}} = [(\eta_0 - \mu_2)/(\eta_0 + \mu_2)]^2. \quad (4.11)$$

The higher the index of the dielectric film, the greater is the fall in reflectance at the minimum. The reflectance rises above that of the bare metal at the maximum, but, for the metals commonly used as reflectors, the increase is not great, and so the lower-index films are to be preferred as protecting layers. As an example, we can consider aluminium, which has a refractive index of 0.82 – i5.99 at 546 nm [3], with protecting layers of quartz of index 1.45 or a high-index layer, 2.3, such as cerium oxide. The results in table 4.1 were calculated from equations (4.5)–(4.7), (4.10) and (4.11). Clearly, if high-index films are used for protecting metal layers, then the monitoring of layer thickness must be accurate, otherwise there is a risk of a sharp drop in reflectance.

Aluminium is probably the commonest mirror coating material for the visible region, and, in addition to the quartz and cerium oxide mentioned above, there is a large number of materials which can be used for protecting it. Silicon oxide, SiO, for example, is also a very effective protecting material, but it has strong absorption at the blue end of the spectrum, where it causes the reflectance of the composite coating to be rather low. Another useful coating is sapphire Al₂O₃. This can be vacuum deposited, or the aluminium at the surface of the coating can be anodised by an electrolytic technique [1], forming a very hard layer of aluminium oxide. Gold and silver are more difficult to protect because of the difficulty of getting films to stick to them. However, it has been found that aluminium oxide sticks very well to silver [4, 5]. Aluminium oxide does not appear to be a very effective barrier against moisture and so it has been used principally as a bonding layer between the silver and a layer of silicon oxide which affords good moisture resistance and which, although it adheres only weakly to silver, adheres strongly to the aluminium oxide. Further details of the coating are

given by Hass and his colleagues [4]. To reduce the absorption at the blue end of the spectrum, the silicon oxide should be reactively deposited (see chapter 9) when the actual oxide which is produced lies between SiO and SiO₂. With such a coating it is possible to achieve a reflectance greater than 95% over the visible and infrared from 0.45–20 μm.

Aluminium oxide and silicon oxide are absorbing at wavelengths longer than 8 μm and it has been discovered by Pellicori [6] and confirmed theoretically by Cox *et al* [5] that reflectors protected by these materials exhibit a sharp dip in reflectance at high angles of incidence, that is, 45° and above. The dip can be avoided by the use of a protecting material which does not absorb in this region. Magnesium fluoride is such a material, but it must be deposited on a hot substrate (temperatures in excess of 200 °C) if it is to be robust. The metals have their best performance if deposited at room temperature and thus the substrates should only be heated after they have been coated with the metal.

4.1.3 Overall system performance, boosted reflectance

In optical instruments of any degree of complexity there will be a number of reflecting components in series, and the overall transmission of the system will be given by the product of the reflectances of the various elements. Figure 4.3 gives the overall transmission of any system with a number of components in series, with identical values of reflectance. It is obvious from the diagram that even with the best metal coatings, the performance with ten elements, say, is low. If the instrument is to be used over a wide range there is little that can be done to alleviate the situation. Most spectrometers, for instance, have ten or more reflections with a consequent severe drop in transmission, but are required to work over a wide region—possibly as much as a 25:1 variation in wavelength. The spectrometer designer normally just accepts this loss and designs the rest of the instrument accordingly.

In cases where the wavelength range is rather more limited, say, to the visible region or to a single wavelength, it is possible to increase the reflectance of a simple metal layer by boosting it with extra dielectric layers.

The characteristic admittance of a metal can be written $n - ik$ and the reflectance in air at normal incidence is

$$R = \left| \frac{1 - (n - ik)}{1 + (n - ik)} \right|^2 = \frac{(1 - n)^2 + k^2}{(1 + n)^2 + k^2} = \frac{1 - [2n/(1 + n^2 + k^2)]}{1 + [2n/(1 + n^2 + k^2)]}. \quad (4.12)$$

On p 53 it was shown that the optical admittance of an assembly Y becomes n^2/Y when a quarter-wave optical thickness of index n , that is admittance in free space units, is added.

If the metal is overcoated with two quarter-waves of material of indices n_1 and n_2 , n_2 being next to the metal, then the optical admittance at normal incidence

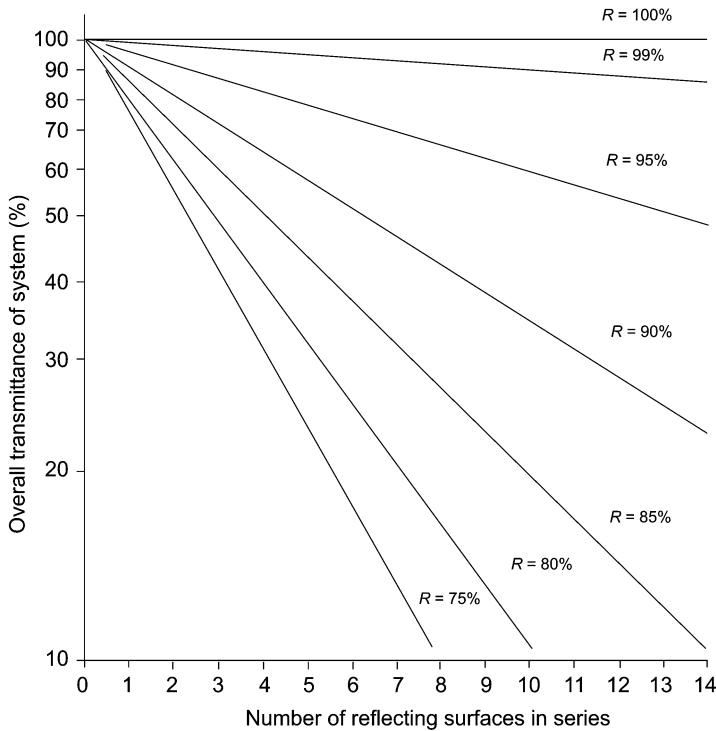


Figure 4.3. Overall transmittance of an optical system which has a number of reflecting elements in series.

is

$$\left(\frac{n_1}{n_2} \right)^2 (n - ik)$$

and the reflectance in air, also at normal incidence,

$$R = \left| \frac{1 - (n_1/n_2)^2(n - ik)}{1 + (n_1/n_2)^2(n - ik)} \right|^2$$

i.e.

$$\begin{aligned}
 R &= \frac{[1 - (n_1/n_2)^2 n]^2 + (n_1/n_2)^4 k^2}{[1 + (n_1/n_2)^2 n]^2 + (n_1/n_2)^4 k^2} \\
 &= \frac{1 - [2(n_1/n_2)^2 n]/[1 + (n_1/n_2)^4(n^2 + k^2)]}{1 + [2(n_1/n_2)^2 n]/[1 + (n_1/n_2)^4(n^2 + k^2)]}. \tag{4.13}
 \end{aligned}$$

This will be greater than the reflectance of the bare metal, given by

equation (4.12), if

$$\frac{2(n_1/n_2)^2 n}{1 + (n_1/n_2)^4(n^2 + k^2)} < \frac{2n}{1 + n^2 + k^2} \quad (4.14)$$

which is satisfied by either

$$\begin{aligned} \left(\frac{n_1}{n_2}\right)^2 &> 1 \\ \text{or} \\ \left(\frac{n_1}{n_2}\right)^2 &< \frac{1}{n^2 + k^2} \end{aligned} \quad (4.15)$$

assuming that $n^2 + k^2 \geq 1$.

The first solution is of greater practical value than the second, which can be ignored. This shows that the reflectance of any metal can be boosted by a pair of quarter-wave layers for which $(n_1/n_2) > 1$, n_1 being on the outside and n_2 next to the metal. The higher this ratio, the greater the increase in reflectance. As an example, consider aluminium at 550 nm with $n - ik = 0.92 - i5.99$. From equation (4.12), the untreated reflectance of this is approximately 91.6%.

If the aluminium is covered by two quarter-waves consisting of magnesium fluoride of index 1.38, next to the aluminium, followed by zinc sulphide of index 2.35, then $(n_1/n_2)^2 = 2.9$ and, from equation (4.13), the reflectance jumps to 96.9%.

An approximate result can be obtained very quickly using $A = (1 - R)$. When the two layers are added, A is reduced roughly to $A/(n_1/n_2)^2$. Inserting the above figures, for aluminium, A is 8.4% initially, and on addition of the layers drops to 2.9%, corresponding to a boosted reflectance of 97.1% (instead of the more accurate figure of 96.9%).

A second similar pair of dielectric layers will boost the reflectance even higher—to approximately 99%, and greater numbers of quarter-wave pairs may be used to give an even higher reflectance.

Unfortunately, the region over which the reflectance is boosted is limited. Outside this zone the reflectance is less than it would be for the bare metal. Jenkins [7] has measured the reflectance of an aluminium layer overcoated with six quarter-wave layers of cryolite, of index 1.35, and zinc sulphide of index 2.35. With layers monitored at 550 nm, the reflectance of the boosted aluminium was greater than 95% over a region 280 nm wide, and greater than 99% over the major part.

More robust coatings can be obtained using magnesium fluoride, silicon dioxide or aluminium oxide as the low-index layers, and cerium oxide or titanium oxide as the high-index layers. To attain maximum toughness, the dielectric layers should be deposited on a hot substrate. Aluminium, however, if deposited hot, tends to scatter badly and so the substrates should be heated only after deposition

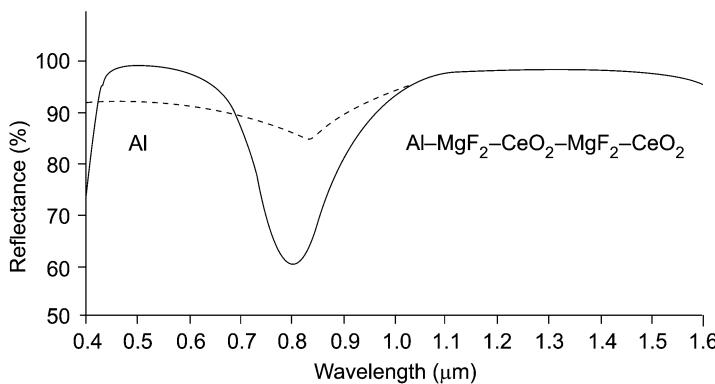


Figure 4.4. Reflectance of evaporated aluminium with (solid curve) and without (dashed curve) two reflectance-increasing film pairs of MgF_2 and CeO_2 as a function of wavelength from 0.4–1.6 μm . (After Hass [1].)

of the aluminium is complete. Figure 4.4 shows the reflectance of aluminium boosted by four quarter-wave layers, which enhanced the reflectance over the visible region.

We have already considered more exactly the behaviour of a single dielectric layer on a metal, and have shown, as did Park [2], that the thickness of the dielectric layer for minimum reflectance should be

$$D = \{\tan^{-1}[2\beta\eta_f/(\eta_f^2 - \alpha^2 - \beta^2)]\}[\lambda_0/(4\pi)]$$

where $(\alpha - i\beta)$ is the admittance of the metal and the angle is in the first or second quadrant. This is the thickness which the low-index layer next to the metal should have if the maximum possible increase in reflectance is to be achieved. A moment's consideration of the admittance diagram will show that this is indeed the case. Layers other than that next to the metal will, of course, retain their quarter-wave thicknesses.

4.1.4 Reflecting coatings for the ultraviolet

The production of high-reflectance coatings for the ultraviolet is a much more exacting task than for the visible and infrared. A very full review of the topic is given by Madden [8], supplemented in great detail by a later account by Hass and Hunter [9]. The following is a very brief summary.

The most suitable material known for the production of reflecting coatings for the ultraviolet out to around 100 nm is aluminium. To achieve the best results, the aluminium should be evaporated at a very high rate, 40 nm s^{-1} or more if possible, on to a cold substrate, the temperature of which should not be permitted to exceed 50°C , and at pressures of 10^{-6} torr or lower. The aluminium should be

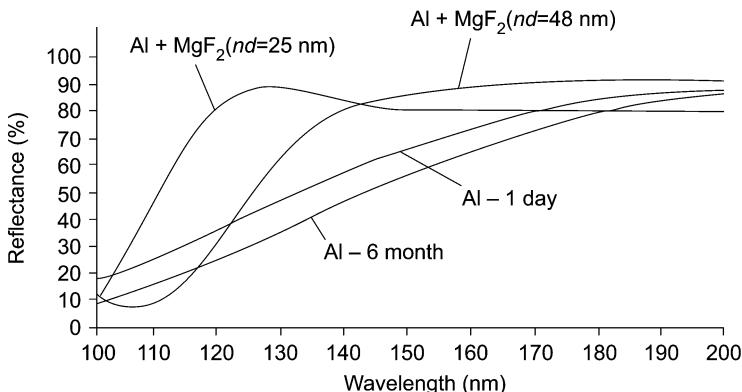


Figure 4.5. Reflectance of evaporated aluminium from 100–200 nm with and without protective layers of MgF₂ of two different thicknesses. (After Canfield *et al* [11].)

of the purest grade. Hass and Tousey [10] have quoted results which show that there is a significant improvement (as high as 10% at 150 nm) in the ultraviolet reflectance of aluminium films if 99.99% pure aluminium is used in preference to 99.5% pure. Aluminium should, in theory, have a much higher reflectance than is usually achieved in practice, particularly at the shortwave end of the range. This has been found to be due to the formation of a thin oxide layer on the surface, and as we have already shown, such a layer must, unless it is very thick, lead to a reduction in reflectance. This oxidation takes place even at partial pressures of oxygen below 10^{-6} torr. Unprotected aluminium films, therefore, inevitably show a rapid fall in reflectance with time when exposed to the atmosphere. The reflectance stabilises when the layer is of sufficient thickness to inhibit further oxidation, but this occurs only when the reflectance at short wavelengths has fallen catastrophically.

Attempts have been made to find suitable protecting material for aluminium to prevent oxidation, and very promising results have been obtained with magnesium fluoride (very robust coatings) and lithium fluoride (less robust), which in crystal form are very useful window materials for the ultraviolet. Figures 4.5 and 4.6 show the effect of an extra protecting layer of magnesium fluoride [11] or lithium fluoride [12] on the reflectance of aluminium. The increase in reflectance is partly due to the lack of oxide layer, but also to interference effects.

It is necessary to evaporate the protecting layer immediately after the aluminium in order that the minimum amount of oxidation should be allowed to take place. This is usually achieved by running the two sources simultaneously and arranging for the shutter which covers the aluminium source at the end of the deposition of the aluminium layer to uncover at the same time the magnesium or lithium fluoride source. The use of magnesium fluoride overcoated aluminium as

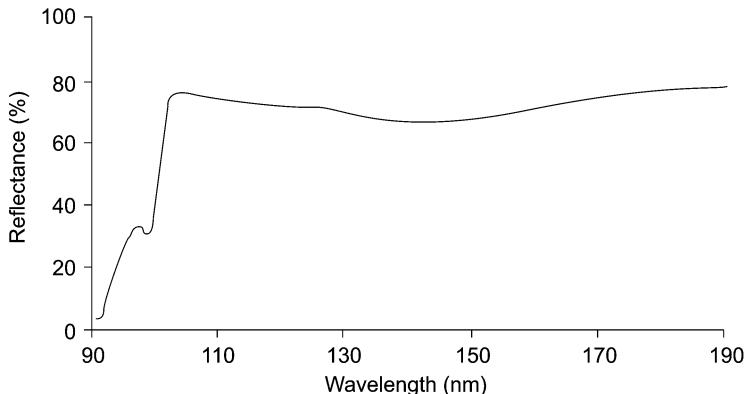


Figure 4.6. Reflectance of an evaporated aluminium film with a 14-nm thick LiF overcoating in the region of 90–190 nm. Measurements were begun 10 minutes after the evaporation was completed. (After Cox *et al* [12].)

a reflecting coating for the ultraviolet is now becoming standard practice.

The aluminium and magnesium fluoride coating is examined in some detail by Canfield *et al* [11]. Amongst other results they show that provided the magnesium fluoride is thicker than 10 nm the coatings will withstand, without deterioration, exposure to ultraviolet radiation and to electrons (up to 10^{16} , 1 MeV electrons/cm²) and protons (up to 10^{12} , 5 MeV protons/cm²).

4.2 Neutral beam splitters

A device which divides a beam of light into two parts is known as a beam splitter. The functional part of a beam splitter generally consists of a plane surface coated to have a specified reflectance and transmittance over a certain wavelength range. The incident light is split into a transmitted and a reflected portion at the surface, which is usually tilted so that the incident and reflected beams are separated. The ideal values of reflectance and transmittance may vary from one application to another. The beam splitters considered in this section are known as neutral beam splitters, because reflectance and transmittance should ideally be constant over the wavelength range concerned.

Neutral beam splitters are usually specified by the ideal values of transmittance and reflectance expressed as a percentage and written T/R . 50/50 beam splitters are probably the most common.

4.2.1 Beam splitters using metallic layers

Apart from a single uncoated surface, which is sometimes used, the simplest type of beam splitter consists of a metal layer deposited on a glass plate. Silver,

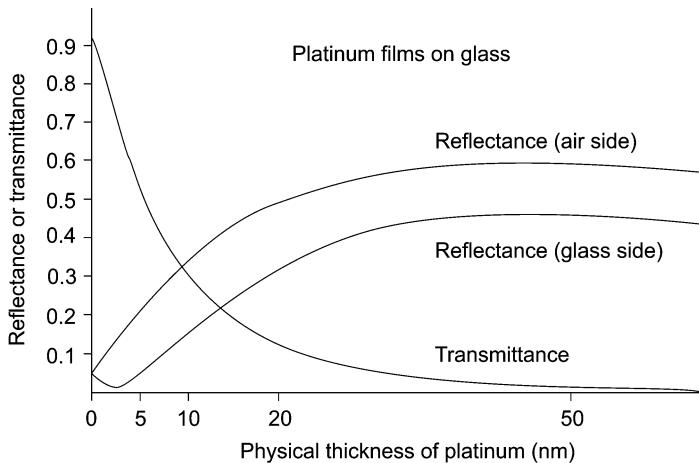


Figure 4.7. Reflectance and transmittance curves for a platinum film on glass, calculated from the optical constants on the bulk metal. (After Heavens [13].)

which has least absorption of all the common metals used in the visible region, is traditionally the most popular material for this. 50/50 beam splitters are frequently referred to as being ‘half-silvered’, although commercial beam splitters nowadays are usually constructed from metals such as chromium which are less prone to damage by abrasion and corrosion.

All metallic beam splitters suffer from absorption. The transmission of a metal film is the same, regardless of the direction in which it is measured. This is not so for reflectance, and that measured at the air side is slightly higher than that measured at the glass side. This effect does not appear with a transparent film. Since $T + A + R = 1$, the reduction in reflectance at the substrate side means that the absorption from that side must always be higher. Figure 4.7 shows curves for platinum demonstrating this behaviour [13]. Because of this difference in reflection, metallic beam splitters should always be used in the manner shown in figure 4.8 if the highest efficiency is to be achieved.

It is possible to decrease the absorption in metallic beam splitters by adding an extra dielectric layer. The method has been applied to chromium films by Pohlack [14] and figure 4.9 gives some of the measurements made.

The first pair of results is for a simple chromium film on glass of index 1.52 measured both from the air side and the glass side. The second pair of results shows how the absorption in the chromium can be reduced by the presence of a quarter-wave layer of high refractive index material (zinc sulphide of index approximately 2.4 in this case) between the metal and the glass. This layer forms an antireflection coating on the rear surface of the metal, and the effect can be seen particularly strongly in the results for reflectance and transmission from the glass side. There, the transmission remains exactly as before, but the reflectance

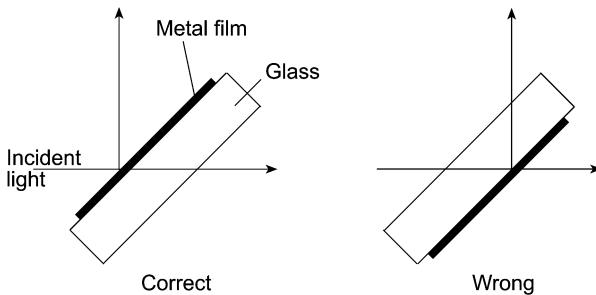


Figure 4.8. Correct use of a metallic beam splitter.

	Arrangement						
R	0.28	0.09	0.47	0.03	0.17	0.33	0.02
T	0.32	0.32	0.32	0.32	0.34	0.42	0.42
A	0.40	0.59	0.21	0.65	0.49	0.25	0.56

Air Chromium Zinc sulphide Glass
 $n = 1.0$ $N = 2 - i_3$ $n = 2.4$ $n = 1.52$

<span style="display: inline-block; width: 15px; height:

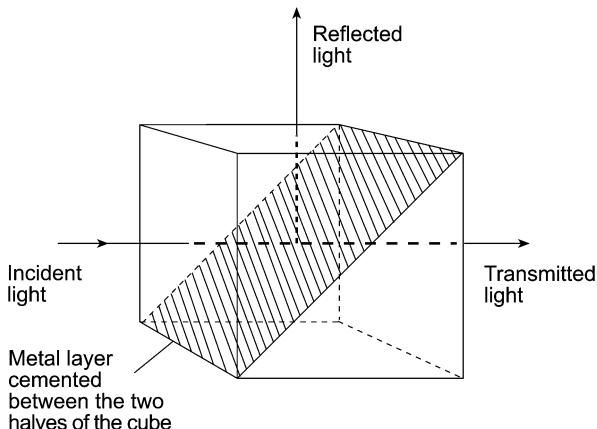


Figure 4.10. A cube beam splitter.

One complication found with beam splitters is a difference in the values of reflectance for the two planes of polarisation when the beam splitter is tilted. The TE (or s-) reflectance is higher than the TM (or p-) reflectance. In calculating the efficiency of a beam splitter this must be taken into account. Anders [16] describes a method for calculating efficiency and stray light performance.

It is not always possible to use the flat plate beam splitter in some optical systems. Reflections from the rear surface can be a problem in spite of the antireflection layer behind the metal film, and in applications where the light passing through the plate is not collimated, aberrations are introduced. To overcome these difficulties a beam-splitting cube, as shown in figure 4.10, can be used, although the absorption in the metal is greater in this configuration because both surfaces, instead of just one, are now in contact with a medium whose index is greater than unity. Since the cemented assembly protects the metal layers the choice of materials is wide. Silver is probably most frequently used, although chromium, aluminium and gold are also popular.

Chromium gives almost neutral beam splitting over the visible region, with an absorption of approximately 0.55 for both planes of polarisation, the TE reflectance being approximately 0.30 and the TM 0.15. Silver varies more with wavelength, the reflectance falling towards the blue end of the spectrum, but the absorption is rather less than for chromium, around 0.15 at 550 nm, with TE reflectance 0.50 and TM 0.30. Curves of the performance of several different metallic beam splitters are given by Anders [16].

4.2.2 Beam splitters using dielectric layers

There are many optical instruments where the light undergoes a transmission followed by a reflection, or vice versa, both at the same, or at the same type

of, beam splitter. In two-beam interferometers, for example, the beams are first of all separated by one pass through a beam splitter and then combined again by a further pass either through the same beam splitter, as in the Michelson interferometer, or through a second beam splitter, as in the Mach-Zehnder interferometer. The effective transmittance of the instrument is given by the product of the transmission and the reflectance of the beam splitter, taking into account the particular polarisation involved. For a perfect beam splitter, TR would be 0.25; for most metallic beam splitters it is around 0.08 or 0.10. The absorption in the film is the primary source of loss.

A beam splitter of improved performance, as far as the TR product is concerned, can be obtained by replacing the metallic layer with a transparent high-index quarter-wave. At normal incidence the reflectance of a quarter-wave is given by

$$R = \left(\frac{1 - n_1^2/n_2}{1 + n_1^2/n_2} \right)^2.$$

At 45° angle of incidence in air the position of the peak is shifted to a shorter wavelength, and the appropriate optical admittances must be used in calculating peak reflectance.

$$R = \left(\frac{\eta_0 - (\eta_1^2/\eta_2)}{\eta_0 + (\eta_1^2/\eta_2)} \right)^2$$

and since η varies with the plane of polarisation, R will have two values, R_{TE} and R_{TM} .

Figure 4.11 shows the peak reflectance of a quarter-wave of index between 1.0 and 3.0 on glass of index 1.52 for both 45° incidence and normal incidence. At 45° , the peak reflectance for unpolarised light, $\frac{1}{2}(R_{TE} + R_{TM})$, is within 1.5% of the peak value for normal incidence.

Zinc sulphide, with index 2.35, is a popular material for beam splitters. At 45° we have

$$(TR)_{TE} = (0.46 \times 0.54) = 0.248$$

$$(TR)_{TM} = (0.185 \times 0.815) = 0.151$$

and

$$(TR)_{\text{unpolarised}} = \frac{1}{2}(0.248 + 0.151) = 0.200.$$

$(TR)_{\text{unpolarised}}$ cannot be calculated using $T_{\text{mean}}R_{\text{mean}}$ ($= 0.219$) because the light, after having undergone one reflection or transmission, is then partly polarised.)

If a more robust film is required, cerium oxide, with an index approximately 2.25, is a good choice. Here

$$(TR)_{TE} = (0.423 \times 0.577) = 0.244$$

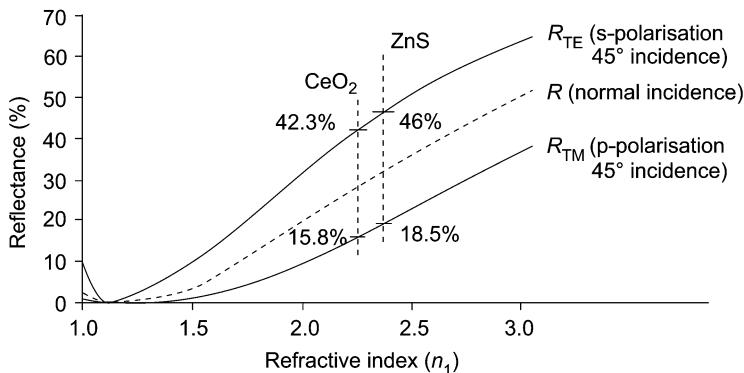


Figure 4.11. Peak reflectance in air of a quarter-wave of index n_1 on glass of index 1.52 at normal and 45° incidence.

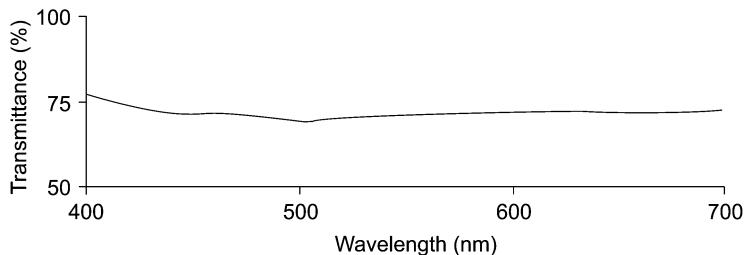


Figure 4.12. Measured transmittance curve of a dielectric 70/30 beam splitter at 45° angle of incidence. (Courtesy of Sir Howard Grubb, Parsons & Co. Ltd.)

$$(TR)_{TM} = (0.158 \times 0.842) = 0.133$$

$$(TR)_{\text{unpolarised}} = 0.189.$$

Clearly the dielectric beam splitter, even if it does tend to have characteristics which more nearly correspond to 70/30 rather than 50/50, has a considerably better performance than the metallic beam splitter. The reflectance curve of a typical 70/30 beam splitter in figure 4.12 shows how the reflectance varies on either side of the peak.

Beam splitters with 55/45 characteristics can be made by evaporating pure titanium in a good vacuum and subsequently oxidising it to TiO_2 by heating at 420°C in air at atmospheric pressure. The titanium oxide thus formed has rutile structure and a refractive index of 2.8. Titanium films produced in a poor vacuum oxidise subsequently to the anatase form, having rather lower refractive index. The production of very large beam splitters, of this type, 17 × 13 inches, is described in a paper by Holland *et al* [17].

The single-layer beam splitter suffers from a fall in reflectance on either side

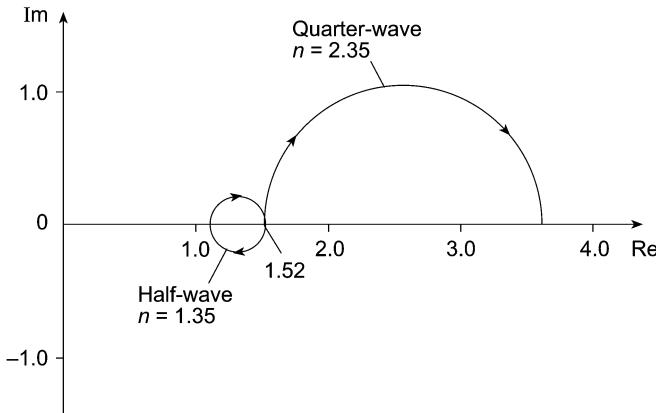


Figure 4.13. Admittance diagram at λ_0 of a two-layer beam splitter. The high-index quarter-wave layer gives the required high reflectance. The low-index half-wave layer flattens the performance over the visible region.

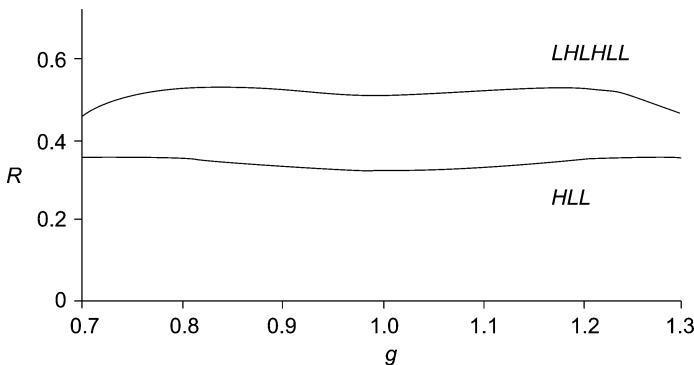


Figure 4.14. (a) The performance of the beam splitter shown in figure 4.13. Design: Air (1.00)|*HLL*|Glass (1.52) with *L* a quarter-wave of index 1.35 and *H* of 2.35. (b) The performance of a beam splitter of design: Air (1.00)|*LHLHLL*|Glass (1.52) with indices as for (a).

of the central wavelength. In the same way that single-layer antireflection coatings can be broadened by adding a half-wave layer, so the single quarter-wave beam splitter can be broadened. The same basic pattern of admittance circles can be achieved either by a low-index half-wave layer between the high-index quarter-wave and the glass substrate or an even higher index half-wave deposited over the quarter-wave. Since no suitable materials for the latter solution exist in practice, the low-index half-wave is the only feasible approach. The admittance diagram is shown in figure 4.13 and the performance in figure 4.14.

The technique is effective also for multilayer systems to give a higher reflectance. Approximately 50% reflectance can be achieved by a four-layer coating, Air |*LHLH*| Glass, and this can be flattened by an additional low-index half-wave at the glass end of the multilayer, that is, Air |*LHLHLL*| Glass. Figure 4.14 shows the performance calculated for this design of beam splitter.

A detailed discussion of the role of half-wave layers is given by Knittl [18].

As mentioned above, beam-splitting cubes must be used in some applications where plate beam splitters are unsuitable. Unfortunately, the main problem connected with dielectric beam splitters, the low reflectance for TM waves, becomes even worse with cube beam splitters. The reason for this is simply that 45° incidence in glass is effectively a much greater angle of incidence than 45° in air. Consequently, the polarisation splitting is even greater and the TM performance becomes so poor that the beam splitter is unusable in most applications. Metal layers are, therefore, the only ones which can be used in the straightforward cube beam splitter and combiner. This disadvantage of the dielectric layer can, however, be turned to advantage in the construction of polarisers as we shall see in chapter 8.

4.3 Neutral-density filters

A filter which is intended to reduce the intensity of an incident beam of light evenly over a wide spectral region is known as a neutral-density filter.

The performance of neutral-density filters is usually defined in terms of the optical density, D :

$$D = \log_{10}(I_0/I_T)$$

where I_0 is the incident intensity and I_T is the transmitted intensity measured either at one particular wavelength or integrated over a region.

Absorption and absorptance are terms which are not correctly used of neutral-density filters because they represent the fraction of energy which is actually absorbed in the film, and in neutral-density filters a proportion of the incident energy is removed by reflection.

The advantage of using the logarithmic term is that the effect of placing two or more neutral-density filters in series is easily calculated. The overall density is simply the sum of the individual densities (provided that multiple reflections are not permitted to occur between the individual filters, which would affect the result in the way shown in chapter 2, p 69, equation (2.139)).

Thin-film neutral-density filters consist of single metallic layers with thicknesses chosen to give the correct transmission values. Rhodium, palladium, tungsten, chromium, as well as other metals, are all used to some extent, but the best performance is obtained by the evaporation of a nickel chromium alloy, approximately 80% nickel and 20% chromium. Chromel A or Nichrome are standard resistance wires which have this composition and can be readily

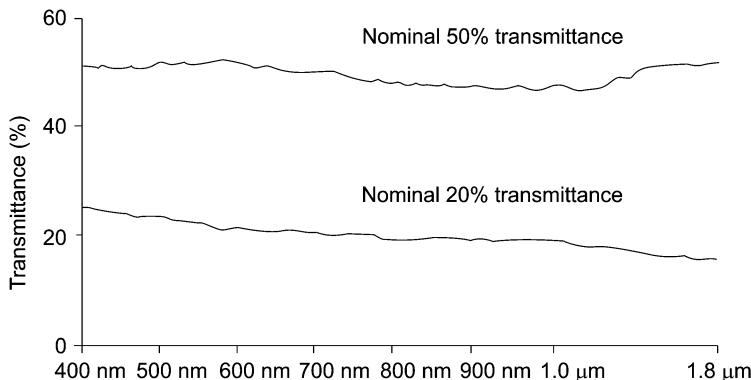


Figure 4.15. Measured transmittance curves of neutral-density filters consisting of Nichrome films on glass substrates. (Courtesy of Sir Howard Grubb, Parsons & Co. Ltd.)

obtained. The method is described by Banning [19]. The Chromel or Nichrome should be evaporated at 10^{-4} torr or better from a thick tungsten spiral. Neutral films, having densities up to around 1.5, corresponding to a transmission of 3%, can be manufactured in this way. If the films are made thicker, they are not as neutral and tend to have a higher transmission in the red, owing to excess chromium. The films are very robust and do not need any protection, especially if they are heated to around 200 °C after evaporation.

Figure 4.15 shows some response curves of neutral-density filters made from Nichrome on glass. The filters are reasonably neutral over the visible and near infrared out to 2 μm . In fact, if quartz substrates are used the filters will be good over the range 0.24–2 μm .

References

- [1] Hass G 1955 Filmed surfaces for reflecting optics *J. Opt. Soc. Am.* **45** 945–52
- [2] Park K C 1964 The extreme values of reflectivity and the conditions for zero reflection from thin dielectric films on metal *Appl. Opt.* **3** 877–81
- [3] Hass G 1972 Optical constants of metals *American Institute of Physics Handbook* ed D E Gray (New York: MacGraw-Hill) pp 6-124–56. The value used for aluminium, 0.82 – i5.99 at 546 nm, is quoted on p 6-125
- [4] Hass G, Heany J B, Herzig H, Osantowski J F and Triolo J J 1975 Reflectance and durability of Ag mirrors coated with thin layers of Al_2O_3 plus reactively deposited silicon oxide *Appl. Opt.* **14** 2639–44
- [5] Cox J T, Hass G and Hunter W R 1975 Infrared reflectance of silicon oxide and magnesium fluoride protected aluminium mirrors at various angles of incidence from 8 μm to 12 μm *Appl. Opt.* **14** 1247–50
- [6] Pellicori S F 1974 Private communication (Santa Barbara Research Center, Goleta, CA) see reference [5]

- [7] Jenkins F A 1958 Extension du domaine spectral de pouvoir réflecteur élevé des couches multiples diélectriques *J. Phys. Radium* **19** 301–6
- [8] Madden R P 1963 Preparation and measurement of reflecting coatings for the vacuum ultraviolet *Physics of Thin Films* vol 1, ed G Hass (New York: Academic) pp 123–86
- [9] Hass G and Hunter W R 1978 The use of evaporated films for space applications—extreme ultraviolet astronomy and temperature control of satellites *Physics of Thin Films* vol 10, ed G Hass and M H Francombe (New York: Academic) pp 71–166
- [10] Hass G and Tousey R 1959 Reflecting coatings for the extreme ultraviolet *J. Opt. Soc. Am.* **49** 593–602
- [11] Canfield L R, Hass G and Waylonis J E 1966 Further studies on MgF₂-overcoated aluminium mirrors with highest reflectance in the vacuum ultraviolet *Appl. Opt.* **5** 45–50
- [12] Cox J T, Hass G and Waylonis J E 1968 Further studies on LiF overcoated aluminium mirrors with highest reflectance in the vacuum ultraviolet *Appl. Opt.* **7** 1535–9
- [13] Heavens O S 1955 *Optical Properties of Thin Solid Films* (London: Butterworths) figure 6.5, p 162
- [14] Pohlack H 1953 Beitrag zur Optik dünner metallschichten *Jenaer Jahrbuch* (Jena: Zeis) pp 241–5
- [15] Shkliarevskii I N and Avdeenko A A 1959 Increasing the transparency of metallic coatings *Opt. Spectrosc.* **6** 439–43
- [16] Anders H 1965 *Dunne Schichten für die Optik* (Stuttgart: Wissenschaftliche Verlagsgesellschaft) pp 82–91
- [17] Holland L, Hacking K and Putner T 1953 The preparation of titanium dioxide beam-splitters of large surface area *Vacuum* **3** 159–61
- [18] Knittl Z 1976 *Optics of Thin Films* (London: Wiley)
- [19] Banning M 1947 Neutral density filters of Chromel A *J. Opt. Soc. Am.* **37** 686–7

Chapter 5

Multilayer high-reflectance coatings

The metal reflecting layers of the previous chapter suffer from a considerable absorption loss which, although unfortunate, still permits a high level of performance in most simple systems. There are applications where the absorption in metal layers is too high and the reflectance too low. These include multiple-beam interferometers and resonators, where the large number of successive reflections magnifies the effects of absorption, and high-power systems where the energy absorbed can be sufficient to damage the coating. One way of increasing the reflectance of an opaque metal coating, as we have seen, is to boost the reflectance by adding dielectric layers. This also reduces the absorptance, but the transmittance remains effectively zero. For high-reflecting coatings which must transmit what they do not reflect, all-dielectric multilayers are required. The description which follows is built around the most successful of the multiple-beam interferometers, the Fabry–Perot interferometer. As we shall see later, this interferometer is also of considerable importance in the development of thin-film band-pass filters, and this is a further reason for dealing with it in some detail here.

5.1 The Fabry–Perot interferometer

First described in 1899 by Fabry and Perot [1], the interferometer known by their names has profoundly influenced the development of thin-film optics. It belongs to the class of interferometers known as multiple-beam interferometers because a large number of beams is involved in the interference. The theory of each of the various types of multiple-beam interferometer is similar. They differ mainly in physical form. Their common feature is that their fringes are much sharper than those in two-beam interferometers, thus improving both measuring accuracy and resolution. Multiple-beam interferometers are described in almost all textbooks on optics, for example that by Born and Wolf [2].

A Fabry–Perot interferometer consists of two flat plates separated by a distance d_s and aligned so that they are parallel to a very high degree of accuracy.

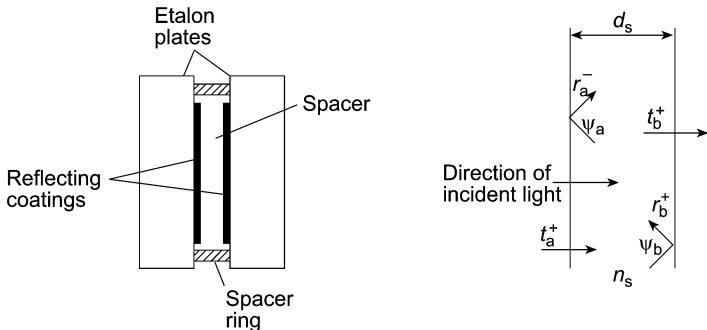


Figure 5.1. A Fabry–Perot etalon. The amplitude coefficients in the diagram are converted to the intensity coefficients of equation (5.1) as shown on p 76.

The separation is usually maintained by a spacer ring made of Invar or quartz, and the assembly of two plates and a spacer is known as an etalon. The inner surfaces of the two plates are usually coated to enhance their reflectance.

Figure 5.1 shows an etalon in diagrammatic form. The amplitude reflection and transmission coefficients are defined as shown. The basic theory has already been given in chapter 2 (p 76), where it was shown that the transmission for a plane wave is given by

$$T = \frac{T_a T_b}{[1 - R_a^- R_b^+]^{1/2}]^2} \left[1 + \frac{4(R_a^- R_b^+)^{1/2}}{[1 - (R_a^- R_b^+)^{1/2}]^2} \sin^2 \left(\frac{\phi_a + \phi_b}{2} - \delta \right) \right]^{-1} \quad (5.1)$$

where $\delta = (2\pi n_s d_s \cos \theta_s)/\lambda$, d_s and n_s being the physical thickness and refractive index of the spacer layer. This is similar to (2.150) except that δ has been modified to include oblique incidence θ_s . In order to simplify the discussion, let the reflectances and transmittances of the two surfaces be equal, let there be no phase change on reflection, i.e. let $\phi_a = \phi_b = 0$, and let n_s be unity, i.e. an air spacer. Then

$$T = \frac{T_s^2}{(1 - R_s)^2} \frac{1}{1 + [4R_s/(1 - R_s)^2] \sin^2 \delta} \quad (5.2)$$

and, writing

$$F = \frac{4R_s}{(1 - R_s)^2} \quad (5.3)$$

then

$$T = \frac{T_s^2}{(1 - R_s)^2} \frac{2}{1 + F \sin^2 \delta}. \quad (5.4)$$

If there is no absorption in the reflecting layers, then

$$1 - R_s = T_s$$

and

$$T = \frac{1}{1 + F \sin^2 \delta}. \quad (5.5)$$

The form of this function is given in figure 5.2 where T is plotted against δ . T is a maximum for $\delta = m\pi$, where $m = 0, \pm 1, \pm 2, \dots$, and a minimum halfway between these values. The successive peaks of T are known as fringes and m is known as the order of the appropriate fringe. As F increases, the widths of the fringes become very much narrower. The ratio of the separation of adjacent fringes to the halfwidth (the fringe width measured at half the peak transmission) is called the ‘finesse’ of the interferometer and is written \mathcal{F} . From equation (5.5), the value of δ corresponding to a transmission of half the peak value is given by

$$0.5 = \frac{1}{1 + F \sin^2 \delta}$$

and if δ is sufficiently small so that we can replace $\sin^2 \delta$ by δ^2 , then

$$0.5 = \frac{1}{1 + F \delta^2}$$

i.e.

$$\delta = \frac{1}{F^{1/2}}$$

which is *half* the width of the fringe. The separation between values of δ representing successive fringes is π , so that

$$\mathcal{F} = \frac{\pi F^{1/2}}{2}$$

or

$$\mathcal{F} = \frac{\pi R_s^{1/2}}{(1 - R_s)}. \quad (5.6)$$

The Fabry–Perot interferometer is used principally for the examination of the fine structure of spectral lines. The fringes are produced by passing light from the source in question through the interferometer. Measurement of the fringe pattern as a function of the physical parameters of the etalon can yield very precise values of the wavelengths of the various components of the line. The two most common arrangements are either to have the incident light highly collimated and incident normally, or at some constant angle, when the fringes can be scanned by varying the spacer thickness, or it is possible to keep the spacer thickness constant and scan the fringes by varying θ_s , the angle of incidence. Possible arrangements corresponding to these two methods are shown in figure 5.3.

Practical considerations limit the achievable finesse to a maximum normally of around 25, or perhaps 50 in exceptional cases. This is due mainly to

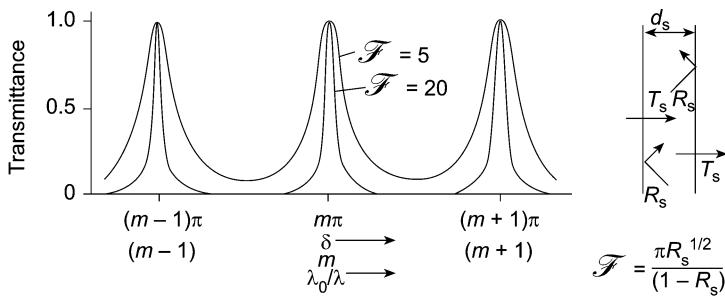


Figure 5.2. Fabry-Perot fringes.

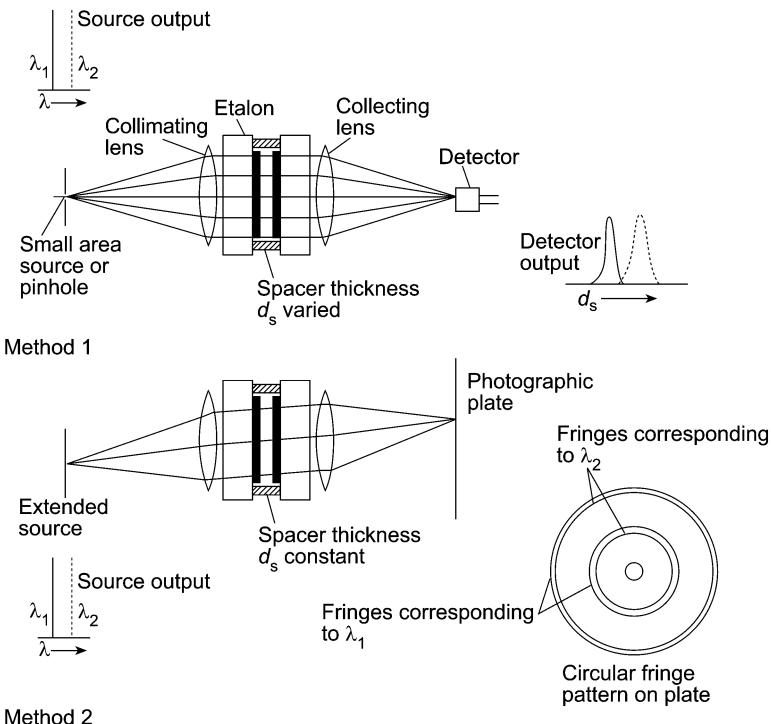


Figure 5.3. Two possible arrangements of a Fabry-Perot interferometer.

imperfections in the plates themselves. It is extremely difficult to manufacture a plate with flatness better than $\lambda/100$ at, say 546 nm. Variations in flatness of the plates give rise to local variations of d_s and hence δ , causing the fringes to shift. These variations should not be greater than the fringe width, otherwise the luminosity of the instrument will suffer. Chabbal [3] has considered this problem in great detail, but for our present purpose it is sufficient to assume that, for a

pair of $\lambda/100$ plates (i.e. having errors not greater than $\pm\lambda/200$ about the mean), the variation in thickness of the spacer layer will be of the order of $\pm\lambda/100$ about the mean. This will occur when the defects in the plates are in the form of either spherical depressions in both plates or else protrusions. This in turn means a change in δ of $\pm 2\pi/100$ corresponding to a total excursion of $2\pi/50$. Any decrease in fringe width below this will not increase the resolution of the system but merely reduce the overall luminosity, so that $2\pi/50$ represents a lower limit on the fringe width. Since the interval between fringes is π , this condition is equivalent to an upper limit on finesse of $\pi/(2\pi/50)$, i.e. 25. In more general terms, if the plates are good enough to limit the total thickness variation in the spacer to λ/p (not quite the same as saying that each plate is good to λ/p), then the finesse should be not greater than $p/2$.

The resolution of an optical instrument is normally determined by the Rayleigh criterion, which is particularly concerned with intensity distributions of the form

$$I(\delta) = \left(\frac{\sin \delta/2}{\delta/2} \right)^2 I_{\max}$$

which are of a type produced by diffraction rather than interference effects. Two wavelengths are considered just resolved by the instrument if the intensity maximum of one component falls exactly over the first intensity zero of the other component. This implies that if the two components are of equal intensity, then, in the combined fringe pattern, the minimum which will exist between the two maxima will be of intensity $8/\pi^2$ times that at either of them. In the Fabry–Perot interferometer the fringes are of rather different form, and the pattern of zeros and successively weaker maxima associated with the $[(\sin \delta/2)/(\delta/2)]^2$ function is missing. The Rayleigh criterion cannot, therefore, be applied directly. Born and Wolf [4] suggest that a suitable alternative form of the criterion, which could be applied in this case, might be that two equally intense lines are just resolved when the resultant intensity between the peaks in the combined fringe pattern is $8/\pi^2$ that at either peak. On this basis they have shown that the resolving power of the Fabry–Perot interferometer is

$$\frac{\lambda}{\Delta\lambda} = 0.97m\mathcal{F}$$

which is virtually indistinguishable from

$$\frac{\lambda}{\Delta\lambda} = m\mathcal{F}$$

and which is the ratio of the peak wavelength of the appropriate order to the halfwidth of the fringe. Thus the halfwidth of the fringe is a most useful parameter because it is directly related to the resolution of the instrument in a most simple manner. We shall make much use of the concept of halfwidth in chapter 7.

Since resolution is the product of finesse and order number, a low finesse does not necessarily mean low resolution, but it does mean that to achieve high resolution the interferometer must be used in high order. This in its turn means that the separation of neighbouring orders in terms of wavelength is small—in high order this is given approximately by λ/m . If steps are not taken to limit the range of wavelengths accepted by the interferometer then the interpretation of the fringe patterns becomes impossible. This limiting of the range can be achieved by using some sort of filter in series with the etalon. This filter could be a thin-film filter of a type discussed in chapter 7. Another method is to use, in series with the etalon, other etalons of lower order, and hence resolution, arranged so that the fringes coincide only at the wavelength of interest and at wavelengths very far removed. The wide fringe interval or, as it is also called, free spectral range, of the low-order, low-resolution instrument is thus combined with the high resolution and narrow free spectral range of the high-order instrument. A simpler and more convenient method, which is probably that most often employed, involves a spectrograph and is generally used in conjunction with the second method of scanning the interferometer: variation of θ_s keeping d_s constant. The resolution of the spectrograph need not be high and the entrance slit can be quite broad. It is usually placed where the photographic plate is in figure 5.3, so that it accepts a broad strip down the centre of the circular fringe pattern. The plate from the spectrograph then shows a low-resolution spectrum with a fringe pattern along each line corresponding to the fine-structure components within the line.

So far in our examination of the Fabry–Perot interferometer we have neglected to consider absorption in the reflecting coatings. Equation (5.4) contains the information we need.

$$T = \frac{T_s^2}{(1 - R_s)^2} \frac{1}{1 + F \sin^2 \delta}. \quad (5.4)$$

Let A_s be the absorptance of the coatings; then

$$1 = R_s + T_s + A_s$$

then equation (5.4) becomes

$$T = \frac{T_s^2}{(T_s + A_s)^2} \frac{1}{1 + F \sin^2 \delta}$$

i.e.

$$T = \frac{1}{(1 + A_s/T_s)^2} \frac{1}{1 + F \sin^2 \delta}. \quad (5.7)$$

Clearly the all-important parameter is A_s/T_s .

Curves are shown in figure 5.4 which connect the transmission of the etalon with finesse, given the absorption of the coatings. It is possible on this diagram to plot the performance of any type of coating if the way in which R_s , T_s and A_s vary

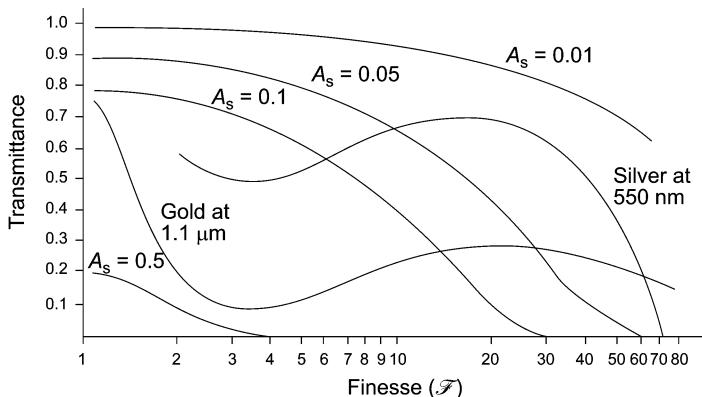


Figure 5.4. Etalon transmittance against finesse for various values of absorptance of the coatings.

is known. This has been done for silver layers at 550 nm and gold at 1.1 μm . The figures from which these curves were plotted were taken from Mayer [5]. Other sources of information, particularly on silver films, are available [6, 7] and results may differ from those plotted in some respects. However, the curves are adequate for their primary purpose, which is to show that the performance of silver, the best metal of all for the visible and near infrared, begins to fall off rapidly beyond a finesse of 20 and is inadequate for the very best interferometer plates. An enormous improvement is possible with all-dielectric multilayer coatings.

5.2 Multilayer dielectric coatings

In chapter 1 it was mentioned that a high reflectance can be obtained from a stack of quarter-wave dielectric layers of alternate high and low index. This is because the beams reflected from all the interfaces in the assembly are of equal phase when they reach the front surface, where they combine constructively. An expression is given on p 53 for the optical admittance of a series of quarter waves. If n_H and n_L are the indices of the high- and low-index layers and if the stack is arranged so that the high-index layers are outermost at both sides, then

$$Y = \left(\frac{n_H}{n_L} \right)^{2p} \frac{n_H^2}{n_s} \quad (5.8)$$

where n_s is the index of the substrate and $(2p + 1)$ the number of layers in the stack.

The reflectance in air or free space is then

$$R = \left(\frac{1 - (n_H/n_L)^{2p} (n_H^2/n_s)}{1 + (n_H/n_L)^{2p} (n_H^2/n_s)} \right)^2. \quad (5.9)$$

The greater the number of layers the greater the reflectance. Maximum reflectance for a given odd number of layers is always obtained with the high-index layers outermost.

If

$$\left(\frac{n_H}{n_L}\right)^{2p} \frac{n_H^2}{n_s} > 1$$

then

$$R \simeq 1 - 4\left(\frac{n_L}{n_H}\right)^{2p} \frac{n_s}{n_H^2}$$

and

$$T = 1 - R \simeq 4\left(\frac{n_L}{n_H}\right)^{2p} \frac{n_s}{n_H^2} \quad (5.10)$$

which shows that when reflectance is high, then the addition of two extra layers reduces the transmission by a factor of $(n_L/n_H)^2$.

Provided the materials which are used are transparent, the absorption in a multilayer stack can be made very small indeed. We shall return later to this topic, but we can note here that in the visible region of the spectrum the absorptance can be less than 0.01%.

Dielectric multilayers, however, suffer from two defects. The first, which is more of a complication than a fault, is that there is a variable change in phase associated with the reflection. The second, which is more serious, is that the high reflectance is obtained over a limited range of wavelengths.

We can see, qualitatively, how the phase shift varies, using the admittance diagram. If, as is usual, the multilayer consists of an odd number of layers with high-index layers on the outside, then at the outer surface of the final layer the admittance will be on the real axis with a high positive value. This is shown diagrammatically in figure 5.5. The quadrants are marked on the figure with reference to figure 2.9(b). Clearly the phase shift associated with the coating is π , for the reference wavelength for which all the layers are quarter-waves. For slightly longer wavelengths, the circles shrink slightly from the semicircles associated with the quarter-waves and so the terminal point of the locus moves upwards into the region associated with the third quadrant. If the wavelength decreases, the terminal point moves into the second quadrant. The phase shift, therefore, increases with wavelength. If, on the other hand, the coating ends with a quarter-wave of low-index material so that at the reference wavelength the admittance is real, but less than unity, then the phase shift on reflection will be zero, moving into the first quadrant as the wavelength increases or into the fourth as it decreases.

To investigate the effect of the phase change, and also of the dispersion of phase change, on the operation of the interferometer, we return to the original

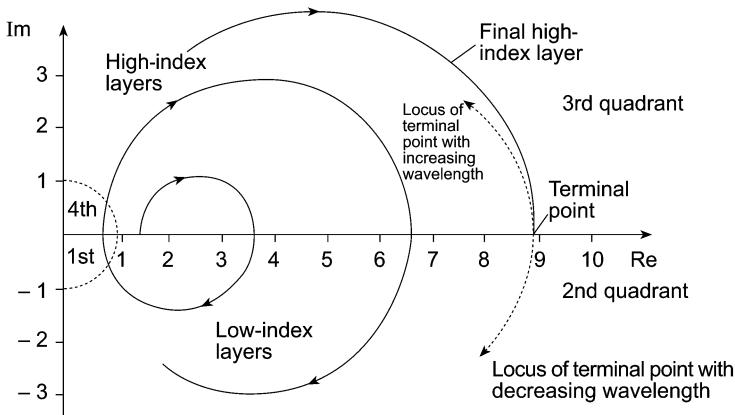


Figure 5.5. Admittance diagram for a quarter-wave stack ending with a high-index layer. The quadrants for the phase shift on reflection ϕ are marked on the diagram and correspond to those in figure 2.9(b). For decreasing wavelength the terminal point moves into the region associated with values of ϕ in the second quadrant while for increasing wavelength ϕ moves into the third quadrant.

formula, equation (5.1). In our analysis we made the assumption that the phase change on reflection was zero and concluded that transmission peaks would be obtained at wavelengths given by

$$\delta = m\pi$$

where $m = 0, \pm 1, \pm 2, \dots$. If we now permit ϕ_a and ϕ_b to be nonzero, then the positions of the transmission peaks will be given by

$$\frac{\phi_a + \phi_b - 2\delta}{2} = q\pi$$

where $q = 0, \pm 1, \pm 2, \dots$. The effect of the phase changes ϕ_a and ϕ_b is simply to shift the positions of the peak wavelengths. If the order is fairly high (and as we have seen most interferometers are used in high order), the shift is quite small. The effect of the phase change, and of any phase dispersion, can be completely eliminated from the determination of wavelength with the interferometer, by a method described by Stanley and Andrew [8] which involves the use of two spacers of different thickness.

The behaviour of a typical quarter-wave stack is shown in figure 5.6. The high-reflection zone can be seen to be limited in extent. On either side of a plateau, the reflectance falls abruptly to a low, oscillatory value. The addition of extra layers does not affect the width of the zone of high reflectance, but increases the reflectance within it and the number of oscillations outside.

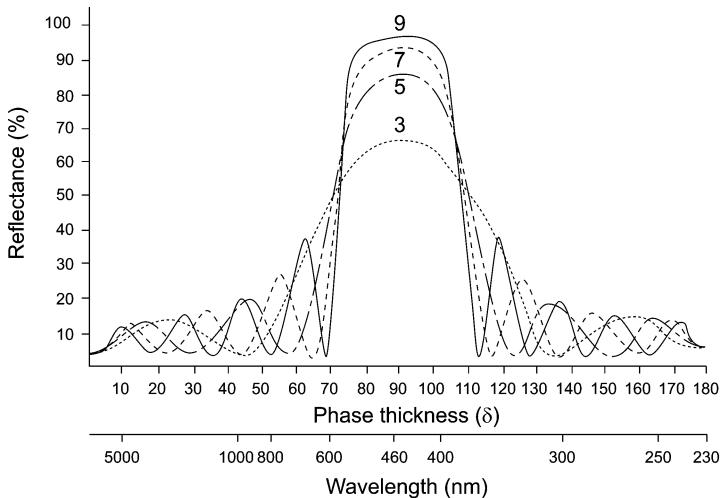


Figure 5.6. Reflectance R for normal incidence of alternating $\lambda_0/4$ layers of high- $(n_H = 2.3)$ and low-index ($n_L = 1.38$) dielectric materials on a transparent substrate ($n_s = 1.52$) as a function of the phase thickness $\delta = 2\pi nd/\lambda$ (upper scale) or the wavelength λ for $\lambda_0 = 460$ nm (lower scale). The number of layers is shown as a parameter on the curves. (After Penselin and Steudel [14].)

The width of the high-reflectance zone can be computed using the following method. If a multilayer consists of n repetitions of a fundamental period consisting of two, three or indeed any number of layers, then the characteristic matrix of the multilayer is given by

$$[\mathcal{M}] = [M]^n$$

where $[M]$ is the matrix of the fundamental period. Let $[M]$ be written

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}.$$

Then it can be shown that for wavelengths which satisfy

$$\left| \frac{M_{11} + M_{22}}{2} \right| \geq 1 \quad (5.11)$$

the reflectance increases steadily with increasing number of periods. This is therefore the condition that a high-reflectance zone should exist and the boundaries are given by

$$\left| \frac{M_{11} + M_{22}}{2} \right| = 1. \quad (5.12)$$

A rigorous proof of this result is somewhat involved. One version is given by Born and Wolf [9] and another by Welford [10]. A justification of the result, rather than a proof, was given by Epstein [11] and it is his method which is followed here.

If the characteristic matrix of a thin-film assembly on a substrate of admittance η_{n+1} is given by

$$\begin{bmatrix} B \\ C \end{bmatrix}$$

then if η_{n+1} is real, equation (2.67) shows that

$$T = \frac{4\eta_0\eta_{n+1}}{(\eta_0B + C)(\eta_0B + C)^*} = \frac{4\eta_0\eta_{n+1}}{|\eta_0B + C|^2}$$

where η_0 is the admittance of the incident medium. Let the characteristic matrix of the assembly of thin films be, as above,

$$[\mathcal{M}] = \begin{bmatrix} \mathcal{M}_{11} & \mathcal{M}_{12} \\ \mathcal{M}_{21} & \mathcal{M}_{22} \end{bmatrix}.$$

Then

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{11} & \mathcal{M}_{12} \\ \mathcal{M}_{21} & \mathcal{M}_{22} \end{bmatrix} \begin{bmatrix} 1 \\ \eta_{n+1} \end{bmatrix} = \begin{bmatrix} \mathcal{M}_{11} + \eta_{n+1}\mathcal{M}_{12} \\ \eta_{n+1}\mathcal{M}_{22} + \mathcal{M}_{21} \end{bmatrix}$$

where $[\mathcal{M}] = [M]^n$ as before and we have

$$T = \frac{4\eta_0\eta_{n+1}}{|\eta_0(\mathcal{M}_{11} + \eta_{n+1}\mathcal{M}_{12}) + \eta_{n+1}\mathcal{M}_{22} + \mathcal{M}_{21}|^2}.$$

If there is no absorption, \mathcal{M}_{11} and \mathcal{M}_{22} are real, and \mathcal{M}_{12} and \mathcal{M}_{21} are imaginary. Then

$$T = \frac{4\eta_0\eta_{n+1}}{|\eta_0\mathcal{M}_{11} + \eta_{n+1}\mathcal{M}_{22}|^2 + |\eta_0\eta_{n+1}\mathcal{M}_{12} + \mathcal{M}_{21}|^2}. \quad (5.13)$$

In the absence of the multilayer, the transmission of the substrate will be

$$T_{\text{sub}} = \frac{4\eta_0\eta_{n+1}}{(\eta_0 + \eta_{n+1})^2}. \quad (5.14)$$

To simplify the discussion, let $\eta_0 = \eta_{n+1}$. Then, from equations (5.13) and (5.14), T will be less than T_{sub} if

$$\frac{|\mathcal{M}_{11} + \mathcal{M}_{22}|}{2} \geq 1$$

regardless of the values of \mathcal{M}_{12} and \mathcal{M}_{21} . Now, if

$$\frac{|\mathcal{M}_{11} + \mathcal{M}_{22}|}{2} > 1$$

where $[M]$ is the matrix of the fundamental period in the multilayer, then, generally, as the number of periods increases, that is, as n tends to infinity,

$$\frac{|\mathcal{M}_{11} + \mathcal{M}_{22}|}{2} \rightarrow \infty.$$

That this is plausible may be seen by first of all squaring $[M]$, whence, writing M'_{pq} for the terms in $[M]^2$,

$$M'_{11} + M'_{22} = (M_{11})^2 + 2M_{12}M_{21} + (M_{22})^2.$$

Since $\det[M] = 1$,

$$2M_{12}M_{21} = 2M_{11}M_{22} - 2$$

so that

$$M'_{11} = M'_{22} = (M_{11} + M_{22})^2 - 2.$$

If

$$\frac{|\mathcal{M}_{11} + \mathcal{M}_{22}|}{2} = 1 + \delta$$

when δ is positive, then

$$M'_{11} + M'_{22} = (2 + 2\delta)^2 - 2 = 2 + 8\delta + 4\delta^2$$

so that by squaring $[M']$ and resquaring the result and so on, it can be seen that

$$\frac{|\mathcal{M}_{11} + \mathcal{M}_{22}|}{2} \rightarrow \infty \quad \text{as} \quad n \rightarrow \infty.$$

The quarter-wave stack, which we have so far been considering, consists of a number of two-layer periods, together with one extra high-index layer. Each period has a characteristic matrix:

$$[M] = \begin{bmatrix} \cos \delta & (i \sin \delta)/n_L \\ in_L \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} \cos \delta & (i \sin \delta)/n_H \\ in_H \sin \delta & \cos \delta \end{bmatrix}.$$

Since the two layers are of equal optical thickness, δ without any suffix has been used for phase thickness.

$$\frac{M_{11} + M_{22}}{2} = \cos^2 \delta - \frac{1}{2} \left(\frac{n_H}{n_L} + \frac{n_L}{n_H} \right) \sin^2 \delta.$$

The right-hand side of this expression cannot be greater than +1, and so to find the boundaries of the high-reflectance zone we must set

$$-1 = \cos^2 \delta_e - \frac{1}{2} \left(\frac{n_H}{n_L} + \frac{n_L}{n_H} \right) \sin^2 \delta_e$$

which, with some rearrangement, gives

$$\left(\frac{n_H - n_L}{n_H + n_L} \right)^2 = \cos^2 \delta_e.$$

Now,

$$\delta = \frac{\pi}{2} \frac{\lambda_0}{\lambda}$$

where λ_0 is, as usual, the wavelength for which the layers have quarter-wave optical thickness. We can also write this as

$$\delta = \frac{\pi}{2} g$$

where

$$g = \frac{\lambda_0}{\lambda}.$$

Let the edges of the high-reflectance zone be given by

$$\delta_e = \frac{\pi}{2} g_e = \frac{\pi}{2} (1 \pm \Delta g)$$

so that

$$\cos^2 \delta_e = \sin^2 \left(\pm \frac{\pi \Delta g}{2} \right)$$

and the width of the zone is $2\Delta g$. Then

$$\Delta g = \frac{2}{\pi} \sin^{-1} \left(\frac{n_H - n_L}{n_H + n_L} \right). \quad (5.15)$$

This shows that the width of the zone is a function only of the indices of the two materials used in the construction of the multilayer. The higher the ratio, the greater the width of the zone. Figure 5.7 shows Δg plotted against the ratio of refractive indices.

So far we have considered only the fundamental reflectance zone for which all the layers are one-quarter of a wavelength thick. It is obvious that high-reflectance zones will exist at all wavelengths for which the layers are an odd number of quarter wavelengths thick. That is, if the centre wavelength of the fundamental zone is λ_0 , then there will also be high-reflectance zones with centre wavelengths $\lambda_0/3, \lambda_0/5, \lambda_0/7, \lambda_0/9$, and so on.

At wavelengths where the layers have optical thickness equivalent to an even number of quarter-waves, which is the same as an integral number of half-waves, the layers will all be absentee layers and the reflectance will be that of the bare substrate.

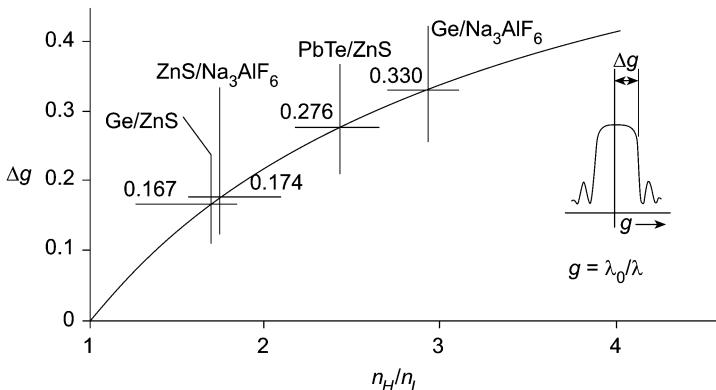


Figure 5.7. The width of the high-reflectance zone of a quarter-wave stack plotted against the ratio of the refractive indices, n_H/n_L .

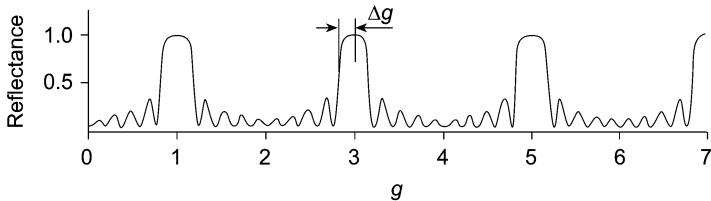


Figure 5.8. Reflectance of a nine-layer stack of zinc sulphide ($n_H = 2.35$) and cryolite ($n_L = 1.35$) on glass ($n = 1.52$) showing the high-reflectance bands.

The analysis determining Δg for the fundamental zone is valid also for all higher-order zones so that the boundaries are given by

$$g_0 \pm \Delta g, \quad 3g_0 \pm \Delta g, \quad 5g_0 \pm \Delta g$$

and so on. Higher-order reflectance curves are shown in figure 5.8.

For the visible region, the most common coating materials are zinc sulphide and cryolite. Absorption levels less than 0.5% can be achieved with ease, 0.1% with extra care and 0.001% with minute attention to detail. Neither material in thin-film form is particularly hard, but they are both easy to evaporate and give high optical performance even when evaporated onto a cold substrate. This means that the risk of distortion of very accurate interferometer plates through heating is eliminated. The layers are rather susceptible to attack by moisture and care should be taken to avoid any condensation, such as might happen when cold plates are exposed to a warmer atmosphere; otherwise, the coatings will be ruined. Touching by fingers is also to be avoided at all costs. The softness of the coatings can, however, be turned to advantage. Etalon plates are extremely expensive and if the coatings are easily removable, the plates can be recoated for use at other

wavelengths. Prolonged soaking in warm water is often sufficient to bring zinc sulphide and cryolite coatings off. In cases where the coatings are not completely removed in this way, the addition of two or three drops of hydrochloric acid to the water will quickly complete the operation. This should obviously be done with great care and the plates immediately rinsed in running water to avoid any risk of surface damage.

Where substrates are worked to somewhat lower tolerances, harder materials can be used. Oxide layers, such as titanium dioxide, zirconium dioxide or cerium dioxide, are all useful high-index materials with indices in the region of 2.2. Magnesium fluoride evaporated on to a hot substrate with an index of 1.38, or quartz, with index 1.45, or silicon oxide, with an index around 1.5, are all useful low-index layers. Such combinations will withstand handling, humidity and abrasion.

For the ultraviolet, a good combination for the 300–400 nm region is antimony trioxide with cryolite, evaporated on to a cold substrate. They should be handled as carefully as zinc sulphide and cryolite.

For the infrared, germanium for the region 1.8–2.0 μm with an index of 4.0, or lead telluride for the region 3.5–4.0 μm , with an index of 5.5, are good high-index materials. Zinc sulphide, with an index of 2.35, is a useful low-index material out to 20 μm . In the near infrared, silicon monoxide, calcium fluoride, magnesium fluoride, cerium fluoride, or thorium fluoride are all good low-index materials. More details of these and of all the other materials mentioned in this chapter will be found in chapter 8.

The losses experienced in the coatings are as much a function of the technique used as of the materials themselves. Great care in preparing the plant and substrates is needed. Everything should be scrupulously clean. Two papers which will be found useful if the maximum performance is required are by Perry [12] and Heitmann [13]. Both these authors are concerned with laser mirrors, where losses must be of an even lower order than in the case of the Fabry–Perot interferometer.

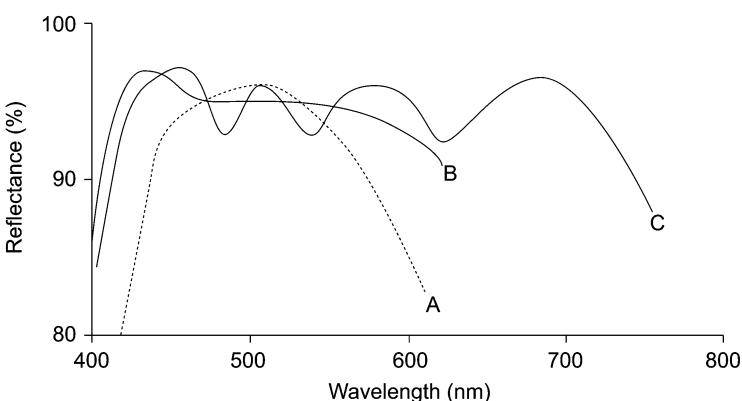
5.2.1 All-dielectric multilayers with extended high-reflectance zones

The limited range over which high reflectance can be achieved with a quarter-wave stack is a difficulty in some applications, and a number of attempts have been made to extend the range by altering the design. Most of these have involved the staggering of the thicknesses of successive layers throughout the stack to form a regular progression, the aim being to ensure that at any wavelength in a fairly wide range, enough of the layers in the stack have optical thickness sufficiently near a quarter-wave to give high reflectance.

Penselin and Steudel [14] were probably the first workers to try this method. They produced a number of multilayers where the layer thicknesses were in a harmonic progression. The best 13-layer results which they published were obtained with the scheme in table 5.1. See also figure 5.9.

Table 5.1. The performance is shown as curve B in figure 5.9.

Number of layers	Material	Index	Wavelength for which layer is a quarter-wave (nm)
	Quartz substrate	1.45	Massive
1	PbCl ₂	2.20	330
2	MgF ₂	1.38	344
3	PbCl ₂	2.20	360
4	MgF ₂	1.38	377
5	ZnS	2.35	396
6	Na ₃ AlF ₆	1.35	417
7	ZnS	2.35	440
8	Na ₃ AlF ₆	1.35	466
9	ZnS	2.35	495
10	Na ₃ AlF ₆	1.35	528
11	ZnS	2.35	566
12	Na ₃ AlF ₆	1.35	609
13	ZnS	2.35	660
	Air	1.00	Massive

**Figure 5.9.** Broadband multilayer reflectors. A, computed curve for a seven-layer quarter-wave stack. B, measured reflectance of a broadband design (Penselin and Steudel [14]). C, measured reflectance of an alternative design (Baumeister and Stone [16]).

Heavens and Liddell [15] used a similar approach. They computed a large number of reflection curves for assemblies of layers for which the thicknesses were in either arithmetic or geometric progression. With the same number of

Table 5.2.

	Number of layers	High-reflectance region (nm)	Wavelength of first-layer quarter-wave (nm)
Arithmetic filters	15	419–625	600
	25	418–725	700
	35	330–840	800
Geometric filters	15	394–625	600
	25	342–730	700
	35	300–826	800

layers the geometric progression gave very slightly broader reflection zones. In the computations the high index was assumed to be 2.36 (zinc sulphide), the low index 1.39 (magnesium fluoride) and the substrate index 1.53 (glass). Values of common difference for the arithmetic progression ranged from -0.05 to $+0.05$, and for the common ratio of the geometric progression from 0.95 to 1.05. Their results for -0.02 and 0.97 respectively are summarised in table 5.2.

The monitoring wavelengths for which each layer is a quarter-wave are given for the arithmetic filters by

$$t, t(1+k), \dots, t[1+(q-2)k], t[1+(q-1)k]$$

and for the geometric filters by

$$t, kt, \dots, k^{q-2}t, k^{q-1}t$$

where q is the number of layers, t the monitoring wavelength for the first layer, and k the common difference or common ratio respectively. A 35-layer geometric curve is shown in figure 5.10.

As in the case of antireflection coatings, computer refinement can be used to improve an initial, less satisfactory performance. Baumeister and Stone [16, 17] pioneered the use of this technique in optical thin films. By trial and error they arrived at a preliminary 15-layer design with high reflectance over an extended range but with unacceptably large dips. The aim was to produce a reflectance of around 95% using zinc sulphide ($n = 2.3$) and cryolite ($n = 1.35$) and the final result is shown as curve C of figure 5.9 with design details listed in table 5.3. Computer limitations forced the use of a very coarse net for the relaxation—only five points were involved—and in addition, arbitrary relationships between the various layers were used to reduce the number of independent variables to five. This was in 1956. Since then, advances in the technique have kept pace with the increasing power of computers. The detailed methods are outside the scope of this book. They are considered in depth by Liddell [18]. As an illustration of what is possible, figure 5.11 shows the calculated performance of a 21-layer design

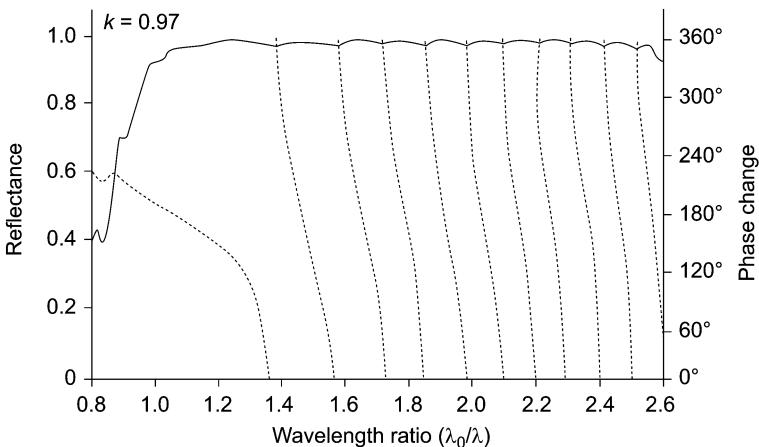


Figure 5.10. Reflectance of a 35-layer geometric stack on glass. Reflectance (full curve) and phase change on reflection (dashed curve); $n_0 = 1.00$, $n_H = 2.36$, $n_L = 1.39$, $n_s = 1.53$, common difference $k = 0.97$. (After Heavens and Liddell [15].)

giving greater than 97% reflectance over the region 400–800 nm. Dispersion of the indices of zinc sulphide and cryolite, the materials used, have been included both in the design procedure and in the performance calculation [19].

Possibly the simplest method of all is to place a quarter-wave stack for one wavelength on top of another for a different wavelength. This process has been considered in detail by Turner and Baumeister [20]. Unfortunately, if each stack consists of an odd number of layers with outermost layers of the same index, then a peak of transmission is found in the centre of the high-reflectance zone. This peak arises because the two stacks act in much the same way as Fabry–Perot reflectors. In a Fabry–Perot interferometer, as we have seen, provided the reflectances and transmittances of the structures on either side of the spacer layer are equal in magnitude, then the transmittance of the assembly will be unity for

$$\frac{\phi_a + \phi_b - 2\delta}{2} = q\pi$$

where $q = 0, \pm 1, \pm 2, \dots$

The situation is sketched in figure 5.12. The assembly of the two stacks is divided at the boundary between them and spaced apart leaving a layer of free space forming a spacer layer. The phase angle ϕ associated with each reflection coefficient is also shown. At one wavelength, given by the mean of the centre wavelengths of the stacks, it can be seen that

$$\phi_a + \phi_b = 2\pi.$$

Also by symmetry, at this wavelength the reflectances of both stacks are equal and, therefore, the condition for unity transmittance will be completely

Table 5.3.

Number of layers	Substance	Index	Wavelength for which layer is a quarter-wave (nm)
Glass substrate			
1	ZnS	2.30	690.8
2	Na ₃ AlF ₆	1.35	690.8
3	ZnS	2.30	690.8
4	Na ₃ AlF ₆	1.35	666.7
5	ZnS	2.30	575.7
6	Na ₃ AlF ₆	1.35	701.3
7	ZnS	2.30	626.2
8	Na ₃ AlF ₆	1.35	517
9	ZnS	2.30	520.5
10	Na ₃ AlF ₆	1.35	463.7
11	ZnS	2.30	463.7
12	Na ₃ AlF ₆	1.35	434.8
13	ZnS	2.30	414
14	Na ₃ AlF ₆	1.35	414
15	ZnS	2.30	414
Air			

satisfied if $2\delta = 0$, that is if the spacer layer of free space is allowed to shrink until it vanishes completely. A peak of transmission will always exist, therefore, if two stacks are deposited so that they are overlapping at the mean of the two monitoring wavelengths. This is shown in figure 5.13, which is reproduced from Turner and Baumeister [20]. Curves A and B are measured reflectance of two high-reflectance quarter-wave stacks, each with the same odd number of layers, starting and finishing with a high-index layer. Curve C shows the measured reflectance of a coating made by combining the two stacks. The peak of transmission can be clearly seen as a dip in the reflectance curve. Experimental errors, either in monitoring or measurement, prevent its reaching the theoretical minimum.

The dip can be removed by destroying the relationship

$$\frac{\phi_a + \phi_b - 2\delta}{2} = q\pi$$

in the region where both stacks have high reflectance. Turner and Baumeister achieved the result quite simply by adding a low-index layer, one quarter-wave thick at the mean wavelength, in between the stacks. This gave value for δ of $\pi/2$ and for $(\phi_a + \phi_b - 2\delta)/2$ of $\pi/2$, which corresponds to minimum possible transmission and maximum reflectance. This is illustrated by curve D. The dip has disappeared completely, leaving a broad flat-topped reflectance curve.

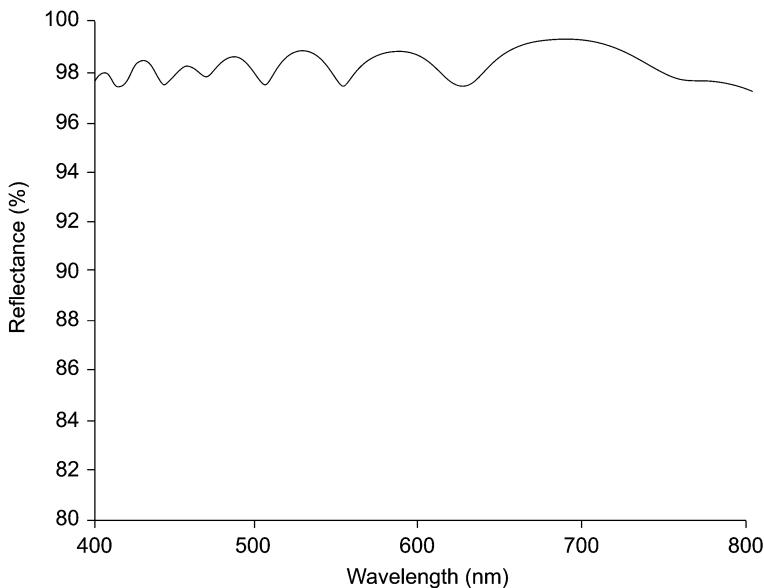


Figure 5.11.

Layer no	Material	Geometrical thickness (nm)	Layer no	Material	Geometrical thickness (nm)
0	Air	Medium	12	Na_3AlF_6	120.4
1	ZnS	41.6	13	ZnS	77.6
2	Na_3AlF_6	76.8	14	Na_3AlF_6	129.9
3	ZnS	51.4	15	ZnS	69.1
4	Na_3AlF_6	94.3	16	Na_3AlF_6	153.0
5	ZnS	49.0	17	ZnS	65.4
6	Na_3AlF_6	94.0	18	Na_3AlF_6	155.7
7	ZnS	47.9	19	ZnS	69.6
8	Na_3AlF_6	95.2	20	Na_3AlF_6	179.1
9	ZnS	58.6	21	ZnS	105.3
10	Na_3AlF_6	147.3	22	SiO_2	Substrate
11	ZnS	62.2			

The calculated performance and the design of a 21-layer high-reflectance coating for the visible and near infrared. Dispersion of the indices of the materials has been taken into account in both design by refinement and in performance calculation. (After Pelletier *et al* [19].)

Turner and Baumeister have also considered the design of broadband reflectors from a slightly different point of view and achieved similar results to the above, although the reasoning is completely different. If a stack is made up of

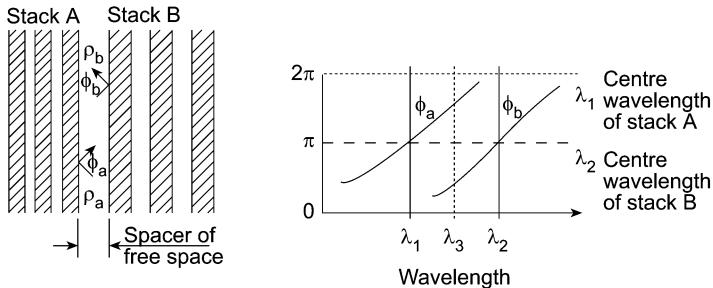


Figure 5.12. At λ_3 , $(\phi_a + \phi_b)/2 = \pi$. Also, by symmetry, at λ_3 , $(\lambda_2/\lambda_3) - 1 = 1 - (\lambda_1/\lambda_3)$, i.e. $\lambda_3 = \frac{1}{2}(\lambda_1 + \lambda_2)$.

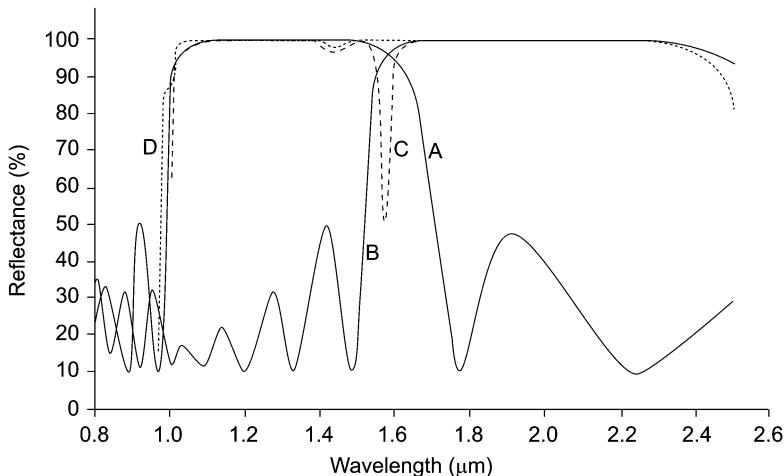


Figure 5.13. Measured reflectances of two quarter-wave stacks with slightly overlapping high-reflectance bands. Individual stacks, full curves: Curve A: A $|0.8(HLHLHLHLH)|G$. Curve B: A $|1.2(HLHLHLHLH)|G$. When these are combined in a single coating, there is a minimum in the overlap region resulting from the condition in figure 5.12: Curve C (dashed): A $|0.8(HLHLHLHLH)1.2(HLHLHLHLH)|G$. An inserted L layer eliminates the minimum by destroying the π phase shift. Curve D (dotted): A $|0.8(HLHLHLHLH)L1.2(HLHLHLHLH)|G$. G denotes the glass substrate ($n = 1.52$), A the air incident medium ($n = 1.00$), H the stibnite high-index films and L the chiolite low-index films. H and L are quarter-wave thicknesses at the reference wavelength, λ_0 , of $1.6 \mu\text{m}$. (After Turner and Baumeister [20].)

a number of symmetrical periods such as

$$\frac{H}{2}L\frac{H}{2} \quad \text{or} \quad \frac{L}{2}H\frac{L}{2}$$

it can be represented mathematically by a single layer of thickness similar to the actual thickness of the multilayer and with a real optical admittance. This relationship holds good for all regions except the zones of high reflectance where the thickness and optical admittance are both imaginary. This result has already been referred to on p 75 and will be examined in much greater detail in the following two chapters. For our present purpose it is sufficient to note that the relationship does exist. If a single layer of real refractive index is deposited on top of a 100% reflector, no interference maxima and minima can possibly exist. For reflectors falling short of the 100% condition, maxima and minima can exist, but are very weak. Thus, in the region where the overlapping stack has a real refractive index, the high reflectance of the lower stack remains virtually unchanged, provided enough layers are used. The high-reflectance zones can either just touch without overlapping, in which case no reflectance minima will exist, or overlap, in which case the minima will be suppressed because the central layer, composed of an eighth-wave from each stack, is a quarter wavelength thick at the mean of the two monitoring wavelengths, and, as has been shown above, this effectively removes any reflectance minima. Figure 5.14(a) shows the measured reflectance of two stacks,

$$\left(\frac{L}{2}H\frac{L}{2}\right)^4$$

on a barium fluoride substrate together with the measured reflectance of two similar stacks superimposed on the same substrate in such a way that the high-reflectance zones just touch.

5.2.2 Coating uniformity requirements

One feature of the broadband reflectors which we have been considering is that the change in phase on reflection varies very rapidly with wavelength, much more rapidly than in the case of the simple quarter-wave stack. The difficulty which this could cause if such coatings were used in the determination of wavelength in a Fabry-Perot interferometer has frequently been mentioned. Actually, the method proposed by Stanley and Andrew [8], which uses two spacers, completely eliminates the effect of even the most rapid phase change with wavelength, but there is another effect which is the subject of a dramatic report by Ramsay and Ciddor [21]. They used a 13-layer coating of a design similar to that of Baumeister and Stone. Their scheme is given in table 5.4.

The coating was deposited with layer uniformity in the region of 1–2 nm from centre to edge of the 75 mm diameter plates. When tested, however, after coating, the plates appeared to be $\lambda/60$ concave at 546 nm, very uniform at 588 nm and $\lambda/10$ convex at 644 nm. This curvature is, of course, only apparent. Tests on the plates using silver layers showed that they were probably $\lambda/60$ concave. The apparent curvature results from changes both in the thickness of the coatings and in the phase change on reflection.

Table 5.4.

Number of layers	Material	Wavelength for which layer is a quarter-wave (nm)
	Fused silica substrate	
1	ZnS	589
2	Na ₃ AlF ₆	671
3	ZnS	720
4	Na ₃ AlF ₆	594
5	ZnS	562
6	Na ₃ AlF ₆	573
7	ZnS	539
8	Na ₃ AlF ₆	535
9	ZnS	571
10	Na ₃ AlF ₆	392
11	ZnS	385
12	Na ₃ AlF ₆	355
13	ZnS	454

In fact, a theory sufficient to explain the effect was published, together with some estimates of required uniformity, by Giacomo [22] in 1958. He obtained the result that the apparent variation of spacer thickness (measured in units of phase) was equal to the error in uniformity of the coating (measured as the variation in physical thickness) times a factor

$$\left(\frac{\nu}{e} \frac{\partial\phi}{\partial\nu} + 4\pi\nu \right)$$

where e is the total thickness of the coating (physical thickness), $\nu = 1/\lambda$ is the wavenumber and ϕ is the phase change on reflection at the surface of the coating. Another way of stating the result is to take $\Delta\rho_m$ as the maximum allowable error in spacer thickness (measured in units of phase) due to this cause, and then the uniformity in coating must be better than

$$\frac{\Delta e}{e} = \frac{\Delta\rho_m}{[(\partial\phi/\partial\nu) + 4\pi e]\nu}.$$

Giacomo showed that the two terms in the expression, $\partial\phi/\partial\nu$ (which is generally negative) and $4\pi e$, could cancel, or partially cancel, so that some designs of coating would be more sensitive to uniformity errors than others. Ramsay and Ciddor carried this further by pointing out that the two terms in the expression vary in magnitude throughout the high-reflectance zone of the coating, and, although the cancellation or partial cancellation does occur, in addition,

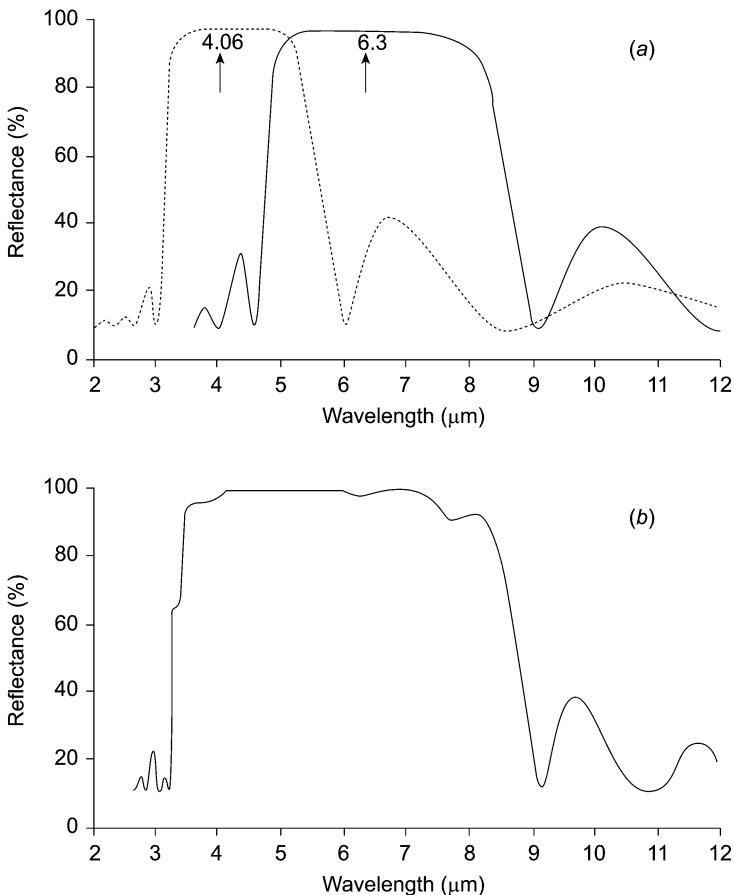


Figure 5.14. (a) Measured reflectances of two stacks $A |(0.5L H 0.5L)^4|G$ on BaF_2 substrates. G denotes the BaF_2 and A air; H and L are films of stibnite and chiolite a quarter-wave thick at reference wavelengths $\lambda_0 = 4.06 \mu\text{m}$ (dashed curve) or $6.3 \mu\text{m}$ (solid curve). (After Turner and Baumeister [20].) (b) Measured reflectance of the two stacks of (a) superimposed in a single coating for an extended high-reflectance region. (After Turner and Baumeister [20].)

the varying magnitudes mean that it is possible in some cases for the apparent curvature due to uniformity errors to vary from concave to convex or vice versa throughout the range. This is so for the particular coating they considered, and it is this change in apparent curvature which is particularly awkward, implying that the interferometer must be tested for flatness over the entire working range, not, as is normal, at one convenient wavelength.

For the conventional quarter-wave coating, the magnitude of $\partial\phi/\partial\nu$ falls

Table 5.5.

Number of layers	Index	Wavelength for which the layer is one quarter-wave thick (nm)
0	1.00	Massive—incident medium
1	1.35	309
2	2.30	866
3	1.35	969
4	2.30	436
5	1.35	521
6	2.30	369
7	1.35	484
8	2.30	441
9	1.35	795
10	2.30	768
11	1.46	Massive—substrate

far short of $4\pi e$; for example, in the case of a seven-layer coating of zinc sulphide and cryolite, for the visible region $\partial\phi/\partial\nu$ is only $-1.5 \mu\text{m}$ compared with $4\pi e$ of around $+21.5 \mu\text{m}$, and the uniformity which is required can readily be calculated from the finesse requirement and the physical thickness of the coating, neglecting the effect of the variations in phase angle altogether. In the case of the broadband multilayer however, the magnitude of $\partial\phi/\partial\nu$ is very much greater, and at some wavelengths will exceed the value of $4\pi e$. For example, Giacomo quotes a case where $\partial\phi/\partial\nu$ reached $-125 \mu\text{m}$, completely swamping the thickness effect, $4\pi e$. Heavens and Liddell, in their paper, quote values of $\partial\phi/\partial\nu$ varying from 10 to $26 \mu\text{m}$ for the staggered multilayers. The change in apparent curvature can therefore occur with these staggered systems, and it is dangerous to attempt to calculate the required uniformity simply from the coating thickness and the finesse requirement. An analysis which is very similar in certain respects, especially in the end result, has been carried out for random errors in the layers of certain types of band-pass filters, and is considered in chapter 7. One point which does arise is the possibility of designing a coating where the two terms cancel almost completely throughout the entire working range. This is mentioned by Ramsay and Ciddor. Since then, Ciddor [23] has carried this a stage further and has now produced several possible designs. Particularly successful is a design for a reflector to give approximately 75% reflectance over the major part of the visible, which is approximately three times less sensitive to thickness variations than would be the case with a reflector exhibiting no phase change at all with change in thickness. The design is intended for film indices of 2.30 and 1.35 on a substrate of index 1.46, corresponding to zinc sulphide and cryolite on fused silica. The thicknesses are given in table 5.5. The reflectance is constant within

perhaps $\pm 2\%$ over the region 650 nm to 400 nm and an interferometer plate with such a coating would behave as if it were much flatter than the purely geometrical lack of uniformity of the coating would suggest.

5.3 Losses

If lossless materials are used, then the reflectance which can be attained by a quarter-wave stack depends solely on the number of layers. If the reflectance is high then the addition of a further pair of layers reduces the transmittance by a factor $(n_L/n_H)^2$. In practice, the reflectance which can be ultimately achieved is limited by losses in the layers. These losses can be scattering or absorption.

Scattering losses are principally due to defects such as dust in the layers or to surface roughness, and techniques for reducing them are considered in chapter 10. Absorption losses are a property of the material, which may be intrinsic or due to impurities or to composition or to structure. Absorption losses are related to the extinction coefficient of the material, and it is useful to consider the absorption losses of a quarter-wave stack composed of weakly absorbing layers having small but nonzero extinction coefficients. Expressions for this have been derived by several workers. The technique we use here is adapted from an approach devised by Hemingway and Lissberger [24].

We use the concept of potential transmittance introduced in chapter 2. We split the multilayer into subassemblies of single layers each with its own value of potential transmittance. The potential transmittance of the assembly is then the product of the individual transmittances.

For the entire multilayer we can write

$$\psi = \frac{T}{1 - R}.$$

Then, if A is the absorptance,

$$1 - \psi = \frac{1 - R - T}{(1 - R)} = \frac{A}{(1 - R)}$$

and

$$A = (1 - R)(1 - \psi).$$

Now $0 \leq \psi \leq 1$ and so we can introduce a quantity \mathcal{A}_f , and write

$$\psi_f = 1 - \mathcal{A}_f$$

for each individual layer, and since we are considering only weak absorption, the potential transmittance will be very near unity and so \mathcal{A}_f will be very small. Then

the potential transmittance of the entire assembly will be given by:

$$\begin{aligned}\psi &= \prod_{f=1}^p \psi_f = \prod_{f=1}^p (1 - \mathcal{A}_f) \\ &= 1 - \sum_{f=1}^p \mathcal{A}_f + \dots\end{aligned}$$

so that, neglecting higher powers of \mathcal{A}_f ,

$$A = (1 - R)(1 - \psi) = (1 - R) \sum_{f=1}^p \mathcal{A}_f.$$

Now let us consider one single layer. The relevant parameters are contained in

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_f & i(\sin \delta_f)/y_f \\ iy_f \sin \delta_f & \cos \delta_f \end{bmatrix} \begin{bmatrix} 1 \\ y_e \end{bmatrix} \quad (5.16)$$

and

$$\psi_f = \frac{\text{Re}(y_e)}{\text{Re}(BC^*)}$$

from equation (2.110). Also

$$\begin{aligned}y_f &= n_f - ik_f \text{ (in free space units)} \\ \delta_f &= 2\pi(n_f - ik_f)d_f/\lambda \\ &= 2\pi n_f d_f/\lambda - i2\pi k_f d_f/\lambda \\ &= \alpha - i\beta\end{aligned}$$

where k_f , and hence β , is small.

If we consider layers which are approximately quarter waves, we can set

$$\alpha = [(\pi/2) + \varepsilon]$$

where ε is small. Then

$$\begin{aligned}\cos \delta_f &\approx (-\varepsilon + i\beta) \\ \sin \delta_f &= 1\end{aligned}$$

and the matrix expression becomes

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} (-\varepsilon + i\beta) & i(n - ik) \\ i(n - ik) & (-\varepsilon + i\beta) \end{bmatrix} \begin{bmatrix} 1 \\ y_e \end{bmatrix}$$

whence

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} (-\varepsilon + i\beta) + iy_e/(n - ik) \\ i(n - ik) + y_e(-\varepsilon + i\beta) \end{bmatrix}$$

so that

$$BC^* = [(-\varepsilon + i\beta) + iy_e/(n - ik)] \cdot [i(n - ik) + y_e(-\varepsilon + i\beta)]^*$$

and, assuming that y_e is real, since we are dealing with a quarter-wave stack, and neglecting terms of second order and above in k , β and ε

$$\operatorname{Re}(BC^*) = (\beta n + y_e + y_e^2 \beta/n)$$

and

$$\psi_f = \frac{y_e}{(\beta n + y_e + y_e^2 \beta/n)} = \frac{1}{1 + \beta[(n/y_e) + (y_e/n)]}.$$

Then, since β is small,

$$\psi_f = 1 - \beta[(n/y_e) + (y_e/n)]$$

and

$$A_f = 1 - \psi_f = \beta[(n/y_e) + (y_e/n)].$$

Next we must find

$$(1 - R) \sum \mathcal{A}_f.$$

For this we need the value of y_e at each interface. Let the stack of quarter-wave layers end with a high-index layer. Then the admittance of the whole assembly will be Y , where Y is large. If we denote the admittance of the incident medium by y_0 ($= n_0$ in free space units) then

$$R = \left[\frac{y_0 - Y}{y_0 + Y} \right]^2$$

where y_0 and Y are real.

If Y is sufficiently large,

$$R = 1 - 4y_0/Y$$

or

$$(1 - R) = 4y_0/Y.$$

Further, since Y is the terminating admittance and the layers are all quarter-waves, the admittances at each of the interfaces follow the pattern:

$$\begin{array}{ccccccccc} Y & \frac{y_H^2}{Y} & \frac{y_L^2 Y}{y_H^2} & \frac{y_H^4}{y_L^2 Y} & \frac{y_L^4 Y}{y_H^4} & \frac{y_H^6}{y_L^4 Y} & \frac{y_L^6 Y}{y_H^6} & & \dots \\ y_0 & | & n_H & | & n_L & | & n_H & | & n_L & | & n_H & | & n_L & | & \dots \end{array}$$

Then

$$\begin{aligned} A &= (1 - R) \sum_{f=1}^p \mathcal{A}_f \\ &= \frac{4y_0}{Y} \left[\left(\frac{y_H}{y_H^2/Y} + \frac{y_H^2/Y}{y_H} \right) \beta_H + \left(\frac{y_L}{y_L^2 Y/y_H^2} + \frac{y_L^2 Y/y_H^2}{y_L} \right) \beta_L \right. \\ &\quad \left. + \left(\frac{y_H}{y_H^4/y_L^2 Y} + \frac{y_H^4/y_L^2 Y}{y_H} \right) \beta_H + \dots \right] \end{aligned}$$

i.e.

$$A = 4y_0 \left[\left(\frac{1}{y_H} + \frac{y_H}{Y^2} \right) \beta_H + \left(\frac{y_L}{y_H^2} + \frac{y_H^2}{y_L Y^2} \right) \beta_L + \left(\frac{y_L^2}{y_H^3} + \frac{y_H^3}{y_L^2 Y^2} \right) \beta_H + \dots \right].$$

Since β_H and β_L are small and Y is large, we can neglect terms in β/Y^2 and the absorptance is then given by

$$A = 4y_0 \left[\left(\frac{1}{y_H} + \frac{y_L^2}{y_H^3} + \frac{y_L^4}{y_H^5} + \dots \right) \beta_H + \left(\frac{y_L}{y_H^2} + \frac{y_L^3}{y_H^4} + \frac{y_L^5}{y_H^6} + \dots \right) \beta_L \right].$$

$(y_L/y_H)^2$ is less than unity and, although the series are not infinite, we can assume that they have a sufficiently large number of terms so that any error involved in assuming that they are in fact infinite is very small.

Thus

$$A = 4y_0 \left(\frac{\beta_H/y_H}{1 - (y_L/y_H)^2} + \frac{y_L\beta_L/y_H^2}{1 - (y_L/y_H)^2} \right) = \frac{4y_0(y_H\beta_H + y_L\beta_L)}{(y_H^2 - y_L^2)}.$$

Now

$$y\beta = y \left(\frac{2\pi kd}{\lambda} \right) = \left(\frac{2\pi nd}{\lambda} \right) k$$

where, since we are working in free space units, we are replacing y by n . Since the layers are quarter-waves,

$$\frac{2\pi nd}{\lambda} = \frac{\pi}{2}$$

so that

$$A = \frac{2\pi n_0(k_H + k_L)}{(n_H^2 - n_L^2)} \text{ (final layer of high index).}$$

The case of a multilayer terminating with a low-index layer can be dealt with in the same way. The final low-index layer acts to reduce the reflectance and so increase the absorption, which is given by

$$A = \frac{2\pi}{n_0} \left[\frac{(n_H^2 k_L + n_L^2 k_H)}{(n_H^2 - n_L^2)} \right] \text{ (final layer of low index).}$$

As an example, we can consider a multilayer with $k_H = k_L = 0.0001$, $n_H = 2.35$ and $n_L = 1.35$, in air, i.e. $n_0 = 1.00$.

$$A = 0.03\% \text{ (high-index layer outermost)}$$

$$A = 0.12\% \text{ (low-index layer outermost).}$$

In fact, the red part of the spectrum, the losses in a zinc sulphide and cryolite stack can be less than 0.001%, indicating that the value of k must be less than 6×10^{-6} assuming that the loss is entirely in one material. For tantalum pentoxide and silicon dioxide multilayer quarter-wave stacks, losses as low as 1 ppm, i.e. 0.0001%, have been reported. This is consistent with values of k an order of magnitude lower. At this level, small amounts of contamination on the reflector surfaces become important additional sources of loss.

In absolute terms, the absorption loss affects the reflectance more than the transmittance in any given quarter-wave stack. Giacomo [25, 26] has shown that $\Delta T/T$ and $\Delta R/R$ are of the same order, and therefore, since $R \gg T$ then $\Delta R \gg T$. We will return to this question of loss later.

References

- [1] Fabry C and Perot A 1899 Théorie et applications d'une nouvelle méthode de spectroscopie interférentielle *Ann. Chim Phys. Paris* **16** 115–44
- [2] Born M and Wolf E 1975 *Principles of Optics* 5th edn (London: Pergamon)
- [3] Chabbel R 1953 Recherche des meilleures conditions d'utilisation d'un spectromètre photoélectrique *Fabry–Perot J. Rech. CNRS* **24** 138–85
- [4] Born M and Wolf E 1965 *Principles of Optics* 3rd edn (London: Pergamon) pp 333–5
- [5] Mayer H 1950 *Physik dünner Schichten* (Stuttgart: Wissenschaftliche Verlagsgesellschaft)
- [6] Kuhn H and Wilson B A 1950 Reflectivity of thin silver films and their use in interferometry *Proc. Phys. Soc. B* **63** 745–55
- [7] Oppenheim U 1956 Semi-reflecting silver films for infrared interferometry *J. Opt. Soc. Am.* **46** 628–33
- [8] Stanley R W and Andrew K L 1964 Use of dielectric coatings in absolute wavelength measurements with a Fabry–Perot interferometer *J. Opt. Soc. Am.* **54** 625–7

- [9] Born M and Wolf E 1965 *Principles of Optics* 3rd edn (London: Pergamon) pp 66–9
- [10] Welford W (writing as W Weinstein) 1954 Computations in thin film optics *Vacuum* **4** 3–19 (The proof is on page 10)
- [11] Epstein L I 1955 Improvements in heat-reflecting filters *J. Opt. Soc. Am.* **45** 360–2
- [12] Perry D L 1965 Low loss multilayer dielectric mirrors *Appl. Opt.* **4** 987–91
- [13] Heitmann W 1966 Extrem hochreflektierende dielektrische spiegelschichten mit zinccelenid *Z. Angew. Phys.* **21** 503–8
- [14] Penselin S and Steudel A 1955 Fabry–Perot interferometersverspiegelungen aus dielektrischen vielfachschichten *Z. Phys.* **142** 21–41
- [15] Heavens O S and Liddell H M 1966 Staggered broad-band reflecting multilayers *Appl. Opt.* **5** 373–6
- [16] Baumeister P W and Stone J M 1956 Broad-band multilayer film for Fabry–Perot interferometers *J. Opt. Soc. Am.* **46** 228–9 (More information about this design technique is given in [17])
- [17] Baumeister P W 1958 Design of multilayer filters by successive approximations *J. Opt. Soc. Am.* **48** 955–8
- [18] Liddell H M 1981 *Computer-Aided Techniques for the Design of Multilayer Filters* (Bristol: Adam Hilger)
- [19] Pelletier E, Klapisch M and Giacomo P 1971 Synthèse d'empilements de couches minces *Nouv. Rev. Opt. Appl.* **2** 247–54
- [20] Turner A F and Baumeister P W 1966 Multilayer mirrors with high reflectance over an extended spectral region *Appl. Opt.* **5** 69–76
- [21] Ramsay J V and Ciddor P E 1967 Apparent shape of broad-band, multilayer reflecting surfaces *Appl. Opt.* **6** 2003–4
- [22] Giacomo P 1958 Propriétés chromatiques des couches réfléchissantes multidiélectriques *J. Phys. Rad.* **19** 307–11
- [23] Ciddor P E 1968 Minimization of the apparent curvature of multilayer reflecting surfaces *Appl. Opt.* **7** 2328–9
- [24] Hemingway D J and Lissberger P H 1973 Properties of weakly absorbing multilayer systems in terms of the concept of potential transmittance *Opt. Acta* **20** 85–96
- [25] Giacomo P 1956 Les couches réfléchissantes multidiélectriques appliquées à l'interféromètre de Fabry–Perot. Etude théorique et expérimentale des couches réelles *Rev. Opt.* **35** 317–54
- [26] Giacomo P 1956 Les couches réfléchissantes multidiélectriques appliquées à l'interféromètre de Fabry–Perot. Etude théorique et expérimentale des couches réelles. II *Rev. Opt.* **35** 442–67

Chapter 6

Edge filters

Filters in which the primary characteristic is an abrupt change between a region of rejection and a region of transmission are known as edge filters. Edge filters are divided into two main groups, longwave-pass and shortwave pass. The operation may depend on many different mechanisms and the construction may take a number of different forms. The following account is limited to thin-film edge filters. These rely for their operation on absorption or interference or both.

6.1 Thin-film absorption filters

A thin-film absorption filter consists of a thin film of material which has an absorption edge at the required wavelength and is usually longwave-pass in character. Semiconductors which exhibit a very rapid transition from opacity to transparency at the intrinsic edge are particularly useful in this respect, making excellent longwave-pass filters. The only complication which usually exists is a reflection loss in the pass region due to the high refractive index of the film. Germanium, for example, with an edge at $1.65 \mu\text{m}$, has an index of 4.0, and, as the thickness of germanium necessary to achieve useful rejection will be at least several quarter-waves, there will be prominent interference fringes in the pass zone showing variations from substrate level, at the half-wave positions, to a reflectance of 68% (in the case of a glass substrate) at the quarter-wave position. The problem can be readily solved by placing antireflection coatings between the substrate and the germanium layer, and between the germanium layer and the air. Single quarter-wave antireflection coatings are usually quite adequate. For optimum matching the values required for the indices of the antireflecting layers are 2.46 between glass and germanium, and 2.0 between germanium and air. The index of zinc sulphide, 2.35, is sufficiently near to both values and, with it, the reflectance near the peak of the quarter-wave coatings will oscillate between

$$\left(\frac{1 - (2.35^4)/(4^2 \times 1.52)}{1 + (2.35^4)/(4^2 \times 1.52)} \right)^2 = 1.3\%$$

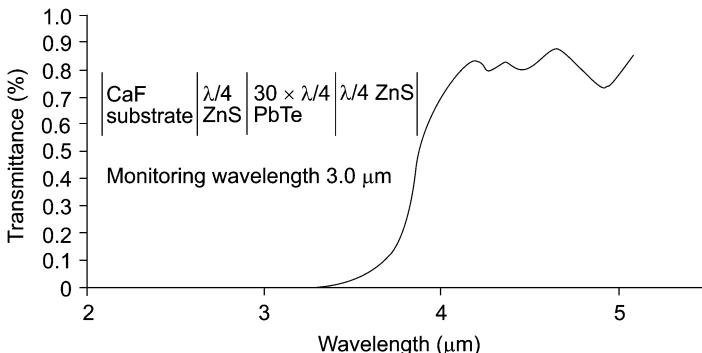


Figure 6.1. The measured characteristic of a lead telluride filter. The small dip at $4.25 \mu\text{m}$ is probably due to atmospheric CO_2 causing a slight unbalance of the measuring spectrometer. (Courtesy of Sir Howard Grubb, Parsons & Co. Ltd.)

for wavelengths where the germanium layer is equal to an integral odd number of quarter-waves, and 4%, that is the reflectance of the bare substrate, where the germanium layer is an integral number of half-waves thick (for at such a wavelength the germanium layer acts as an absentee layer and the two zinc sulphide layers combine also to form a half-wave and, therefore, an absentee layer).

Other materials used to form single-layer absorption filters in this way include cerium dioxide, giving an ultraviolet rejection-visible transmitting filter, silicon, giving a longwave-pass filter with an edge at $1 \mu\text{m}$, and lead telluride, giving a longwave-pass filter at $3.4 \mu\text{m}$.

A practical lead telluride filter characteristic is shown in figure 6.1, which also gives the design. The two zinc sulphide layers were arranged to be quarter-waves at $3.0 \mu\text{m}$. Better results would probably have been obtained if the thicknesses had been increased to quarter-waves at $4.5 \mu\text{m}$.

6.2 Interference edge filters

6.2.1 The quarter-wave stack

The basic type of interference edge filter is the quarter-wave stack of the previous chapter. As was explained there, the principal characteristic of the optical transmission curve plotted as a function of wavelength is a series of high-reflection zones, i.e. low transmission, separated by regions of high transmission. The shape of the transmission curve of a quarter-wave stack is shown in figure 6.2. The particular combination of materials shown is useful in the infrared beyond $2 \mu\text{m}$, but the curve is typical of any pair of materials having a reasonably high ratio of refractive indices.

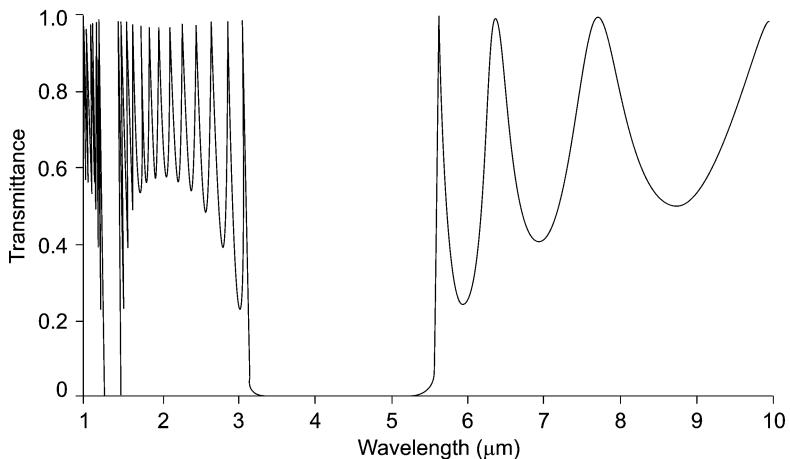


Figure 6.2. Computed characteristic of a 13-layer quarter-wave stack of germanium (index 4.0) and silicon monoxide (index 1.70) on a substrate of index 1.42. The reference wavelength, λ_0 , is 4.0 μm .

The system of figure 6.2 can be used either as a longwave-pass filter with an edge at 5.0 μm or a shortwave-pass filter with an edge at 3.3 μm . These wavelengths can be altered at will by changing the monitoring wavelength.

It sometimes happens that the width of the rejection zone is adequate for the particular application, as, for example, where light of a particularly narrow spectral region only is to be eliminated, or where the detector itself is insensitive to wavelength beyond the opposite edge of the rejection zone. In most cases, however, it is desirable to eliminate all wavelengths shorter than, or longer than, a particular value. The rejection zone, shown in figure 6.2, must somehow be extended. This is usually done by coupling the interference filter with one of the absorption type.

Absorption filters usually have very high rejection in the stop region, but, as they depend on the fundamental optical properties of the basic materials, they are inflexible in character and the edge positions are fixed. Using interference and absorption filters together combines the best properties of both, the deep rejection of the absorption filter with the flexibility of the interference filter. The interference layers can be deposited on an absorption filter, which acts as the substrate, or the interference section can sometimes be made from material which itself has an absorption edge within the interference rejection zone. Within the absorption region the filter behaves in much the same way as the single layers of the previous section.

Other methods of improving the width of the rejection zone will be dealt with shortly, but now we must turn our attention to the more difficult problem created by the magnitude of the ripple in transmission in the pass region. As the

curve of figure 6.2 shows, the ripple is severe and the performance of the filter would be very much improved if somehow the ripple could be reduced.

Before we can reduce the ripple we must first investigate the reason for its appearance, and this is not an easy task, because of the complexity of the mathematics. A paper published by Epstein [1] in 1952 is of immense importance, in that it lays the foundation of a method which gives the necessary insight into the problem to enable the performance to be not only predicted but also improved.

6.2.2 Symmetrical multilayers and the Herpin index

The paper written by Epstein [1] in 1952 dealt with the mathematical equivalent of a symmetrical combination of films and a single layer, and was the beginning of what has become the most powerful design method to date for thin-film filters.

Any thin-film combination is known as symmetrical if each half is a mirror image of the other half. The simplest example of this is a three-layer combination in which a central layer is sandwiched between two identical outer layers. If a multilayer can be split into a number of equal symmetrical periods, then it can be shown that it is equivalent in performance to a single layer having a thickness similar to that of the multilayer and an optical admittance that can be calculated. This is a most important result. Unfortunately, the accurate calculation of the equivalent optical admittance is rather involved, but the basic form of the result can be established relatively easily and used as a qualitative guide. Once the basic form of a filter has been established, computer techniques can be used to finalise the design.

Consider first a symmetrical three-layer period pqp , made up of dielectric materials free from absorption. The characteristic matrix of the combination is given by

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} \cos \delta_p & (i \sin \delta_p) / \eta_p \\ i \eta_p \sin \delta_p & \cos \delta_p \end{bmatrix} \begin{bmatrix} \cos \delta_q & (i \sin \delta_q) / \eta_q \\ i \eta_q \sin \delta_q & \cos \delta_q \end{bmatrix} \times \begin{bmatrix} \cos \delta_p & (i \sin \delta_p) / \eta_p \\ i \eta_p \sin \delta_p & \cos \delta_p \end{bmatrix} \quad (6.1)$$

(where we have used the more general optical admittance η rather than the refractive index n). By performing the multiplication we find:

$$M_{11} = \cos 2\delta_p \cos \delta_q - \frac{1}{2} \left(\frac{\eta_q}{\eta_p} + \frac{\eta_p}{\eta_q} \right) \sin 2\delta_p \sin \delta_q \quad (6.2a)$$

$$M_{12} = \frac{i}{\eta_p} \left[\sin 2\delta_p \cos \delta_q + \frac{1}{2} \left(\frac{\eta_q}{\eta_p} + \frac{\eta_p}{\eta_q} \right) \cos 2\delta_p \sin \delta_q + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \sin \delta_q \right] \quad (6.2b)$$

$$M_{21} = i\eta_p \left[\sin 2\delta_p \cos \delta_q + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \cos 2\delta_p \sin \delta_p - \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \sin \delta_q \right] \quad (6.2c)$$

and

$$M_{22} = M_{11}. \quad (6.2d)$$

It is this last relationship which permits the next step.

Now, let

$$M_{11} = \cos \gamma = M_{22} \quad (6.3)$$

and if we set

$$M_{12} = \frac{i \sin \gamma}{E} \quad (6.4)$$

then, since $M_{11}M_{22} - M_{12}M_{21} = 1$

$$M_{21} = iE \sin \gamma. \quad (6.5)$$

These quantities have exactly the same form as a single layer of phase thickness γ and admittance E . The equations can be solved for γ and E , choosing the particular value of γ which is nearest to the total phase thickness of the period. γ is then the equivalent phase thickness of the three-layer combination and E is the equivalent optical admittance, also known sometimes as the Herpin index. M_{11} does not equal M_{22} in an unsymmetrical arrangement and such a combination cannot, therefore, be replaced by a single layer.

It can easily be shown that this result can be extended to cover any symmetrical period consisting of any number of layers. First the central three layers which, by definition, will form a symmetrical assembly on their own can be replaced by a single layer. This equivalent layer can then be taken along with the next layers on either side as a second symmetrical three-layer combination, which can, in its turn, be replaced by a single layer. The process can be repeated until all the layers have been replaced and a single equivalent layer found.

The importance of this result lies both in the ease of interpretation (the properties of a single layer can be visualised much more readily than those of a multilayer) and in the ease with which the result for a single period may be extended to that for a multilayer consisting of many periods.

If a multilayer is made up of, say, S identical symmetrical periods, each of which has an equivalent phase thickness γ and equivalent admittance E , then physical considerations show that the multilayer will be equivalent to a single layer of thickness $S\gamma$ and admittance E . This result also follows because of an easily derived result:

$$\begin{bmatrix} \cos \gamma & i \sin \gamma / E \\ iE \sin \gamma & \cos \gamma \end{bmatrix}^S = \begin{bmatrix} \cos S\gamma & i \sin S\gamma / E \\ iE \sin S\gamma & \cos S\gamma \end{bmatrix}. \quad (6.6)$$

It should be noted that the equivalent single layer is not an exact replacement for the symmetrical combination in every respect physically. It is merely a mathematical expression of the product of a number of matrices. The effect of changes in angle of incidence, for instance, cannot be estimated by converting the multilayer to a single layer in this way.

In any practical case when the matrix elements are computed it will be found that there are regions where $M_{11} < -1$, i.e. $\cos \gamma < -1$. This expression cannot be solved for real γ , and in this region γ and E are both imaginary. The physical significance of this was explained in the previous chapter, where it was shown that as the number of basic periods is increased the reflectance of a multilayer tends to unity in regions where $|M_{11} + M_{22}|/2 > 1$, M_{11} and M_{22} being elements of the matrix of the basic period. In the present symmetrical case this is equivalent to

$$|M_{11}| = |M_{22}| > 1$$

which therefore denotes a region of high reflectance, i.e. a stop band. Inside the stop band, the equivalent phase thickness and the equivalent admittance are both imaginary. Outside the stop band the phase thickness and admittance are real and these regions are known as pass regions or pass bands. The edges of the pass bands and stop bands are given by $M_{11} = -1$.

6.2.2.1 Application of the Herpin index to the quarter-wave stack

Returning for the moment to our quarter-wave stack, we see that it is possible to apply the above results directly if a simple alteration to the design is made. This is simply to add a pair of eighth-wave layers to the stack, one at each end. Low-index layers are required if the basic stack begins and ends with quarter-wave high-index layers and vice versa. The two possibilities are

$$\frac{H}{2} L H L H L H \dots H L \frac{H}{2}$$

and

$$\frac{L}{2} H L H L H L \dots L H \frac{L}{2}.$$

These arrangements we can replace immediately by

$$\frac{H}{2} L \frac{H}{2} \frac{H}{2} L \frac{H}{2} \frac{H}{2} L \frac{H}{2} \frac{H}{2} L \frac{H}{2} \frac{H}{2} L \frac{H}{2} \dots \frac{H}{2} L \frac{H}{2}$$

and

$$\frac{L}{2} H \frac{L}{2} \frac{L}{2} H \frac{L}{2} \frac{L}{2} H \frac{L}{2} \frac{L}{2} H \frac{L}{2} \frac{L}{2} H \frac{L}{2} \dots \frac{L}{2} H \frac{L}{2}$$

respectively which can then be written as

$$[\frac{H}{2} L \frac{H}{2}]^S \quad \text{and} \quad [\frac{L}{2} H \frac{L}{2}]^S$$

$(H/2)L(H/2)$ and $(L/2)H(L/2)$ being the basic periods in each case. The results in equations (6.1)–(6.6) can then be used to replace both the above stack by single layers making the performance in the pass bands and also the extent of the stop bands easily calculable. We shall examine first the width of the stop bands. As mentioned above, the edges of the stop bands are given by $M_{11} = -1$. Using equation (6.2a) this is equivalent to

$$\cos^2 \delta_{qe} - \frac{1}{2} \left(\frac{\eta_q}{\eta_p} + \frac{\eta_p}{\eta_q} \right) \sin^2 \delta_{qe} = -1$$

which is exactly the same expression as was obtained in the previous chapter for the width of the unaltered quarter-wave stack. There, δ was replaced by $(\pi/2)g$, where $g = \lambda_0/\lambda$ (or v/v_0 , where v is the wavenumber), and the edges of the stop band were defined by

$$\delta_e = \frac{\pi}{2}(1 \pm \Delta g).$$

The width is therefore

$$2\Delta g = 2\Delta \left(\frac{\lambda_0}{\lambda} \right)$$

where, if $\eta_p < \eta_q$,

$$\Delta g = \frac{2}{\pi} \sin^{-1} \left(\frac{\eta_q - \eta_p}{\eta_q + \eta_p} \right) \quad (6.7)$$

or, if $\eta_q < \eta_p$,

$$\Delta g = \frac{2}{\pi} \sin^{-1} \left(\frac{\eta_p - \eta_q}{\eta_p + \eta_q} \right). \quad (6.8)$$

These expressions are plotted in figure 5.7. The width of the stop band is therefore exactly the same regardless of whether the basic period is $(H/2)L(H/2)$, or $(L/2)H(L/2)$. Of course, it is possible to have other three-layer combinations where the width of the central layer is not equal to twice the thickness of the two outer layers, and some of the other possible arrangements will be examined, both in this chapter and the next, as they have some interesting properties, but, as far as the width of the stop band is concerned, it has been shown by Vera [2] that the maximum width for a three-layer symmetrical period is obtained when the central layer is a quarter-wave and the outer layers an eighth-wave each.

Let us now turn our attention to the pass band; first the equivalent admittance and then the equivalent optical thickness. The expression for the equivalent admittance in the pass band is quite a complicated one. From equations (6.2b), (6.2c), (6.4) and (6.5)

$$\begin{aligned} E &= + \left(\frac{M_{21}}{M_{12}} \right)^{1/2} \\ &= + \left(\frac{\eta_p^2 [\sin 2\delta_p \cos \delta_q + \frac{1}{2}(\eta_p/\eta_q + \eta_q/\eta_p) \cos 2\delta_p \sin \delta_q - \frac{1}{2}(\eta_p/\eta_q - \eta_q/\eta_p) \sin \delta_q]}{\sin 2\delta_p \cos \delta_q + \frac{1}{2}(\eta_p/\eta_q + \eta_q/\eta_p) \cos 2\delta_p \sin \delta_q + \frac{1}{2}(\eta_p/\eta_q - \eta_q/\eta_p) \sin \delta_q} \right)^{1/2}. \end{aligned} \quad (6.9)$$

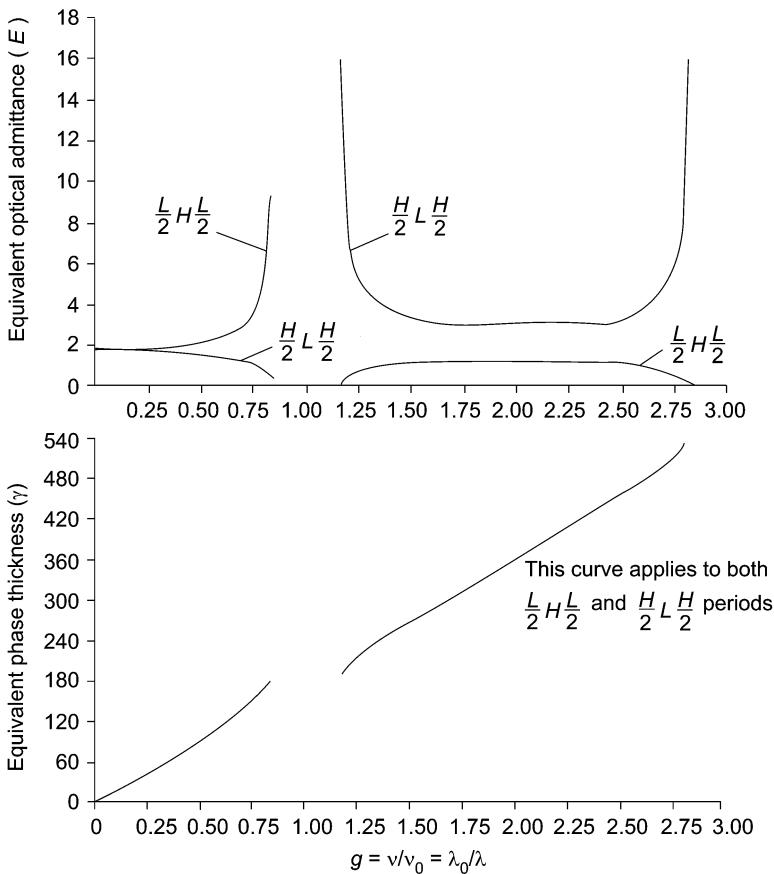


Figure 6.3. Equivalent optical admittance, E , and phase thickness, γ , of a symmetrical period of zinc sulphide ($n = 2.35$) and cryolite ($n = 1.35$) at normal incidence.

This is not a particularly easy expression to handle analytically, but evaluation is straightforward, either by computer or even a programmable calculator. Figure 6.3 shows the equivalent admittance and optical thickness of combinations of zinc sulphide and cryolite. The form of this curve is quite typical of such periods. Once the equivalent admittance and thickness have been evaluated, the calculation of the performance of the filter in the pass region, and its subsequent improvement, become much more straightforward. They are dealt with in greater detail later in this chapter. First we shall examine some of the properties of the expression for the equivalent optical admittance.

We can normalise expression (6.9) by dividing both sides by η_p . E/η_p is then solely a function of δ_p , δ_q and the ratio η_p/η_q . Next, we can make the further simplification, which we have not so far, that $2\delta_p = \delta_q$. The expression

for E/η_p then becomes

$$\frac{E}{\eta_p} = + \left(\frac{\{1 + \frac{1}{2}[\rho + (1/\rho)]\} \cos \delta_q \sin \delta_q - \frac{1}{2}[\rho - (1/\rho)] \sin \delta_q}{\{1 + \frac{1}{2}[\rho + (1/\rho)]\} \cos \delta_q \sin \delta_q + \frac{1}{2}[\rho - (1/\rho)] \sin \delta_q} \right)^{1/2} \quad (6.10)$$

where $\rho = \eta_p/\eta_q$.

It is now easy to see that the following relationships are true. We write $(E/\eta_p)(\rho, \delta_q)$ to indicate that it is a function of the variables ρ and δ_q .

$$\frac{E}{\eta_p}(\rho, \pi - \delta_q) = \frac{1}{(E/\eta_p)(\rho, \delta_q)} \quad (6.11)$$

$$\frac{E}{\eta_p}\left(\frac{1}{\rho}, \delta_q\right) = \frac{1}{(E/\eta_p)(\rho, \delta_q)}. \quad (6.12)$$

These relationships are, in fact, true for all symmetrical periods, even ones which involve inhomogeneous layers, and general statements and proofs of these and other theorems are given by Thelen [3].

Thelen has shown how these relationships may be used to reduce the labour in calculating the equivalent admittance over a wide range. Figure 6.4 shows a set of curves giving the equivalent admittance for various values of the ratio of admittances. The vertical scale has been made logarithmic which has the advantage of making the various sections of the curve repetitions of the first section. This follows directly from the relationships (6.11) and (6.12). The values of the ratios of optical admittances which have been used are all greater than unity. Values less than unity can be derived from the plotted curves using relation (6.12). Again the logarithmic scale means that it is necessary only to reorient the curve for $\eta_p/\eta_q = k$ to give that for $\eta_p/\eta_q = 1/k$. All the information necessary to plot the curves is therefore given in the enlarged version of the first section of figure 6.4 which is reproduced in figure 6.5. Figures 6.4 and 6.5 are both taken from the paper by Thelen [3].

It is also useful to note the limiting values of E :

$$\begin{aligned} E &\text{ tends to } (\eta_p \eta_q)^{1/2} & \text{as } \delta_q &\text{ tends to zero} \\ &\text{and} \\ E &\text{ tends to } \eta_p(\eta_p/\eta_q)^{1/2} & \text{as } \delta_q &\text{ tends to } \pi. \end{aligned} \quad (6.13)$$

The equivalent phase thickness of the period is given by (6.2a) and (6.3) as

$$\gamma = \cos^{-1} \left[\cos 2\delta_p \cos \delta_q - \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \sin 2\delta_p \sin \delta_q \right]. \quad (6.14)$$

This expression for γ is multivalued, and the value chosen is that nearest to $2\delta_p + \delta_q$, the actual sum of the individual phase thicknesses, which is the most easily interpreted value. It is clear from the expression for γ that it does not

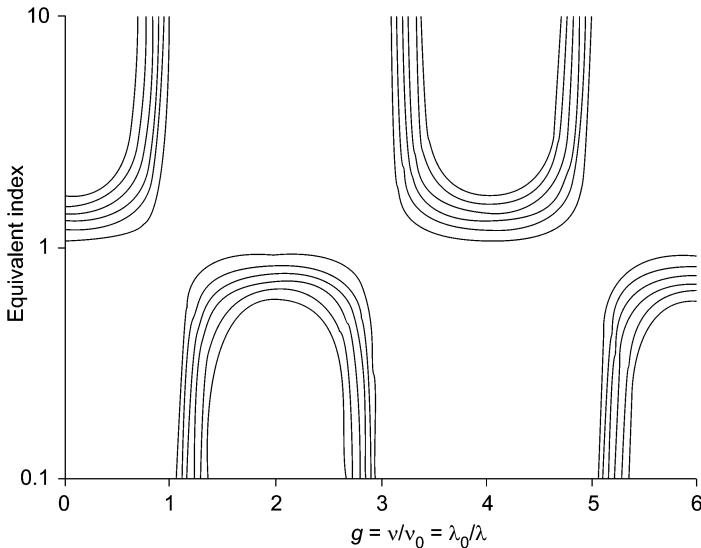


Figure 6.4. Equivalent admittance for the system $(L/2)H(L/2)$. $n_L = 1.00$ and n_H/n_L is a parameter with values 1.23, 1.50, 1.75, 2.0, 2.5, 3.0. The curves with the wider stop bands have the higher n_H/n_L values. (After Thelen [3].)

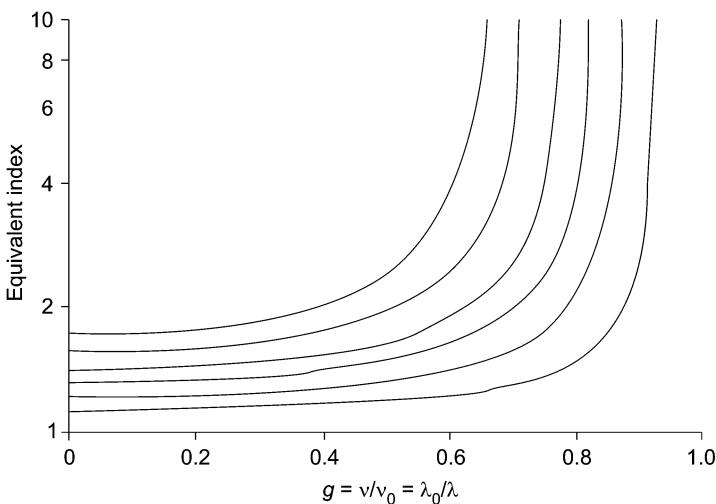


Figure 6.5. Enlarged first part of figure 6.4. (After Thelen [3].)

matter whether the ratio of the admittances is greater or less than unity. The phase thickness for ρ is the same as that for $1/\rho$. Figure 6.6, which is also taken from Thelen's paper, shows the phase thickness of the combinations in figures 6.4

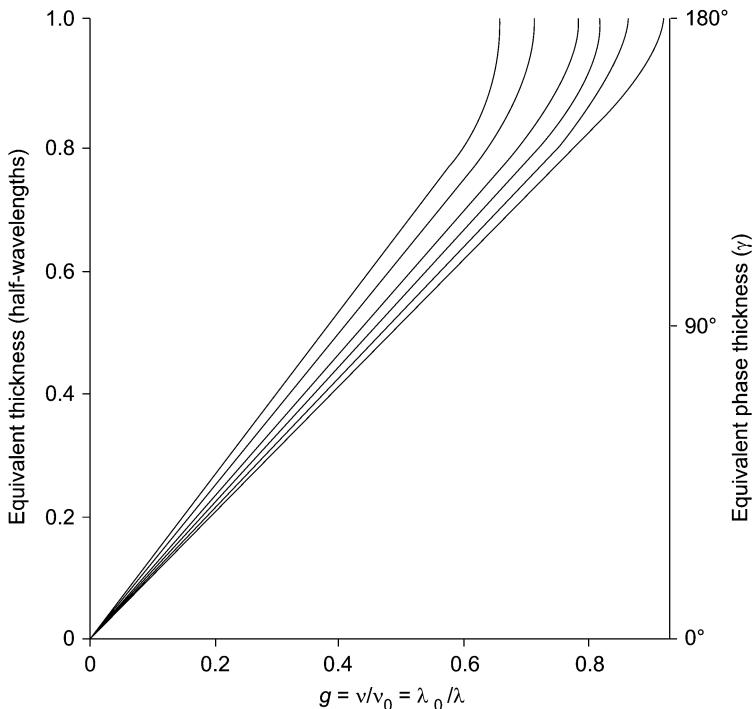


Figure 6.6. Equivalent thickness of the system described in figure 6.4. (After Thelen [3].)

and 6.5. Because of the obvious symmetries, all the information necessary for the complete curve of the equivalent phase thickness is given in this diagram. The equivalent thickness departs significantly from the true thickness only near the edge of the high-reflectance zone. At any other point in the pass bands the equivalent phase thickness is almost exactly equal to the actual phase thickness of the combination.

6.2.2.2 Application of the Herpin index to multilayers of other than quarter-waves

All the curves shown so far are for |eighth-wave| quarter-wave |eighth-wave| periods. If the relative thicknesses of the layers are varied from this arrangement then the equivalent admittance is altered. It has already been mentioned that the reflectance zones for a combination other than the above must be narrower. Some idea of the way in which the equivalent admittance alters can be obtained from

the value as $g \rightarrow 0$. Let $2\delta_p/\delta_q = \psi$. Then, from equation (6.9)

$$E = +\eta_p^2 \left[\frac{\sin 2\delta_p}{\sin 2\delta_q} \cos \delta_q + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \cos 2\delta_p - \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \right]^{1/2} \\ \times \left[\frac{\sin 2\delta_p}{\sin \delta_q} \cos \delta_q + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \cos 2\delta_p + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \right]^{-1/2}. \quad (6.15)$$

Now $\sin 2\delta_p / \sin \delta_q \rightarrow \psi$ as $g \rightarrow 0$, since $\delta_q \rightarrow 0$, $\delta_p \rightarrow 0$, i.e.

$$E \rightarrow \eta_p \left[\psi + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) - \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \right]^{1/2} \\ \times \left[\psi + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \right]^{-1/2}.$$

Rearranging this we obtain

$$\frac{E}{\eta_p} \rightarrow \left(\frac{\psi + (\eta_q/\eta_p)}{\psi + (\eta_p/\eta_q)} \right)^{1/2}. \quad (6.16)$$

This result shows that, for small g , it is possible to vary the equivalent admittance throughout the range of values between η_p and η_q but not outside that range. This result has already been referred to in the chapter on antireflection coatings, where it was shown how to use the concept of equivalent admittance to create replacements for layers having indices difficult to reproduce.

Epstein [1] has considered in more detail the variation of equivalent admittance by altering the thickness ratio and gives tables of results of zinc sulphide/cryolite multilayers.

Ufford and Baumeister [4] give sets of curves which assist in the use of equivalent admittance in a wide range of design problems.

Some results which are at first sight rather surprising are obtained when the value of the equivalent admittance around $g = 2$ is investigated. As $g \rightarrow 2$, $2\delta_p \rightarrow \pi$ and $\delta_q \rightarrow \pi$ so that, from equation (6.15)

$$\frac{E}{\eta_p} \rightarrow \left(\frac{-1 - \frac{1}{2}[(\eta_p/\eta_q) + (\eta_q/\eta_p)] - \frac{1}{2}[(\eta_p/\eta_q) - (\eta_q/\eta_p)]}{-1 - \frac{1}{2}[(\eta_p/\eta_q) + (\eta_q/\eta_p)] + \frac{1}{2}[(\eta_p/\eta_q) - (\eta_q/\eta_p)]} \right)^{1/2} = \left(\frac{\eta_p}{\eta_q} \right)^{1/2}. \quad (6.17)$$

This is quite a straightforward result. Now let $2\delta_p/\delta_q = \psi$, as in the case just considered where $g \rightarrow 0$. Let $g \rightarrow 2$ so that

$$2\delta_p + \delta_q \rightarrow 2\pi.$$

(This is really how, in this case, we define $g = \lambda_0/\lambda$ by defining λ_0 as that wavelength which makes $2\delta_p + \delta_q = \pi$.)

We have, as $g \rightarrow 2$

$$\begin{aligned}\cos 2\delta_p &\rightarrow \cos(2\pi - \delta_q) = \cos \delta_q \\ \sin 2\delta_p &\rightarrow -\sin(2\pi - \delta_q) = -\sin \delta_q\end{aligned}$$

and $\delta_q \rightarrow 2\pi/(1 + \psi)$ so that

$$\begin{aligned}\frac{E}{\eta_p} &\rightarrow \left[-\sin \delta_q \cos \delta_q + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \cos \delta_q \sin \delta_q - \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \sin \delta_q \right]^{1/2} \\ &\quad \times \left[-\sin \delta_q \cos \delta_q + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \cos \delta_q \sin \delta_q \right. \\ &\quad \left. + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \sin \delta_q \right]^{-1/2} \\ &= \left\{ -\cos \delta_q \left[1 - \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \right] - \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \right\}^{1/2} \\ &\quad \times \left\{ -\cos \delta_q \left[1 - \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \right] + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \right\}^{-1/2} \quad (6.18)\end{aligned}$$

where $\cos \delta_q = \cos[2\pi/(1 + \psi)]$.

Whatever the value of ψ , the quantities within the square root brackets have opposite signs, which means that the equivalent admittance is imaginary. Even as $\psi \rightarrow 1$, where one would expect the limit to coincide with the result in equation (6.17), the admittance is still imaginary.

The explanation of this apparent paradox is as follows. An imaginary equivalent admittance, as we have seen, indicates a zone of high reflectance. Consider first the ideal eighth-wave|quarter-wave|eighth-wave stack of equation (6.17). At the wavelength corresponding to $g = 2$, the straightforward theory predicts that the reflectance of the substrate shall not be altered by the presence of the multilayer, because each period of the multilayer is acting as a full wave of real admittance and is therefore an absentee layer. Looking more closely at the structure of the multilayer we can see that this can also be explained by the fact the all the individual layers are a half-wavelength thick. If the ratio of the thicknesses is altered, the layers are no longer a half-wavelength thick and cannot act as absentees. In fact, the theory of the above result shows that a zone of high reflectance occurs.

The transmission of a shortwave-pass filter at the wavelength corresponding to $g = 2$ is therefore very sensitive to errors in the relative thicknesses of the layers. Even a small error leads to a peak of reflection. The width of this spurious

high-reflectance zone is quite narrow if the error is small. Thus the appearance of a pronounced narrow dip in the transmission curve of a shortwave-pass filter is quite a common feature and is difficult to eliminate. The dip is referred to sometimes as a ‘half-wave hole’.

6.2.3 Performance calculations

We are now in a position to make some performance calculations.

6.2.3.1 Transmission at the edge of a stop band

The transmission in the high-reflectance region, or stop band, is an important parameter of the filter. Thelen [3] gives a useful method for calculating this at the edges of the band. His analysis is as follows.

Let the multilayer be made up of S fundamental periods so that the characteristic matrix of the multilayer is

$$[M]^S = \begin{bmatrix} \cos \gamma & (i \sin \gamma)/E \\ iE \sin \gamma & \cos \gamma \end{bmatrix}^S = \begin{bmatrix} \cos S\gamma & (i \sin S\gamma)/E \\ iE \sin S\gamma & \cos S\gamma \end{bmatrix}.$$

At the edges of the stop band we know that $\cos S\gamma \rightarrow 1$, $\sin S\gamma \rightarrow 0$, and $E \rightarrow 0$ or ∞ depending on the particular combination of layers. Now,

$$\frac{\sin S\gamma}{\sin \gamma} \rightarrow S \quad \text{as} \quad \sin \gamma \rightarrow 0$$

so that the matrix tends to

$$\begin{bmatrix} 1 & (iS \sin \gamma)/E \\ iES \sin \gamma & 1 \end{bmatrix} = \begin{bmatrix} 1 & SM_{12} \\ SM_{21} & 1 \end{bmatrix}$$

at the stop band limits. Either M_{12} or M_{21} will also tend to zero because

$$M_{11}M_{22} - M_{12}M_{21} = 1$$

and, depending on which tends to zero, we have either

$$\begin{bmatrix} 1 & SM_{12} \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 0 \\ SM_{21} & 1 \end{bmatrix}$$

for the matrix.

If η_0 is the admittance of the incident medium and η_m of the substrate, then the transmittance of the multilayer at the edge of the stop band is given by equation (2.67):

$$T = \frac{4\eta_0\eta_m}{(\eta_0B + C)(\eta_0B + C)^*}$$

where

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} 1 & SM_{12} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \eta_s \end{bmatrix} \quad \text{if } M_{21} = 0$$

or

$$\begin{bmatrix} 1 & 0 \\ SM_{21} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \eta_m \end{bmatrix} \quad \text{if } M_{12} = 0$$

i.e.

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} 1 + S\eta_m M_{12} \\ \eta_m \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 \\ \eta_m + SM_{21} \end{bmatrix}$$

so that, if there is no absorption,

$$T = \frac{4\eta_0\eta_m}{(\eta_0 + \eta_m)^2 + (S\eta_m\eta_0|M_{12}|)^2} \quad \text{when } M_{21} = 0 \quad (6.19)$$

or

$$T = \frac{4\eta_0\eta_m}{(\eta_0 + \eta_m)^2 + (S|M_{21}|)^2} \quad \text{when } M_{12} = 0 \quad (6.20)$$

(since M_{12} and M_{21} are imaginary in the absence of absorption). For M_{12} or M_{21} to be zero requires that

$$\sin 2\delta_p \cos \delta_q + \frac{1}{2} \left(\frac{\eta_p}{\eta_q} + \frac{\eta_q}{\eta_p} \right) \cos 2\delta_p \sin \delta_q = \mp \frac{1}{2} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \sin \delta_q.$$

If M_{12} is zero we can deduce that

$$|M_{21}| = \left| \eta_p \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \sin \delta_q \right| \quad (6.21)$$

or, if M_{21} is zero, that

$$|M_{12}| = \left| \frac{1}{\eta_p} \left(\frac{\eta_p}{\eta_q} - \frac{\eta_q}{\eta_p} \right) \sin \delta_q \right|. \quad (6.22)$$

At the limits of the high-reflectance zone we have already seen that

$$\cos^2 \delta = \left(\frac{\eta_q - \eta_p}{\eta_q + \eta_p} \right)^2$$

i.e.

$$\sin^2 \delta = 1 - \cos^2 \delta = \frac{4\eta_p\eta_q}{(\eta_q + \eta_p)^2}.$$

Substituting this in the expressions (6.21) and (6.22) for $|M_{21}|$ and $|M_{12}|$ we find

$$|M_{21}|^2 = \left| \frac{4\eta_p(\eta_p - \eta_q)^2}{\eta_q} \right| \quad \text{for} \quad M_{12} = 0 \quad (6.23)$$

$$|M_{12}|^2 = \left| \frac{4(\eta_p - \eta_q)^2}{\eta_p^3 \eta_q} \right| \quad \text{for} \quad M_{21} = 0. \quad (6.24)$$

To give the transmittance at the edges of the high-reflectance zone, these expressions should be used in equations (6.19) and (6.20) according to the rule:

If E , the equivalent admittance, is zero, then M_{21} is zero.

If E , the equivalent admittance, is ∞ , then M_{12} is zero.

6.2.3.2 Transmission in the centre of a stop band

For the simple quarter-wave stack an expression for transmittance at the centre of the high-reflectance zone has already been given in chapter 5. For the present multilayer, the transmittance is of a similar order of magnitude but the eighth-wave layers at the outer edges of the stack complicate matters. The stack may be represented by

$$\frac{p}{2} q \frac{p}{2} \frac{p}{2} q \frac{p}{2} \dots \frac{p}{2} q \frac{p}{2}$$

which is

$$\frac{p}{2} q p q p q p \dots q \frac{p}{2}.$$

If there are S periods, then the layer q appears S times in this expression. At the centre of the high-reflectance zone, the matrix product becomes:

$$\begin{aligned} & \left[\begin{array}{cc} 1/\sqrt{2} & i/(\eta_p \sqrt{2}) \\ i\eta_p/\sqrt{2} & 1/\sqrt{2} \end{array} \right] \left[\begin{array}{cc} 0 & i/\eta_q \\ i\eta_p & 0 \end{array} \right] \left[\begin{array}{cc} 0 & i/\eta_p \\ i\eta_q & 0 \end{array} \right] \dots \left[\begin{array}{cc} 0 & i/\eta_q \\ i\eta_q & 0 \end{array} \right] \\ & \times \left[\begin{array}{cc} 1/\sqrt{2} & i/(\eta_p \sqrt{2}) \\ i\eta_p/\sqrt{2} & 1/\sqrt{2} \end{array} \right] = \left[\begin{array}{cc} 1/\sqrt{2} & i/(\eta_p \sqrt{2}) \\ i\eta_p/\sqrt{2} & 1/\sqrt{2} \end{array} \right] \left[\begin{array}{cc} 0 & i/\eta_q \\ i\eta_q & 0 \end{array} \right] \\ & \times \left[\begin{array}{cc} -\eta_q/\eta_p & 0 \\ 0 & -\eta_p/\eta_q \end{array} \right]^{S-1} \left[\begin{array}{cc} 1/\sqrt{2} & i/(\eta_p \sqrt{2}) \\ i\eta_p/\sqrt{2} & 1/\sqrt{2} \end{array} \right] \\ & = \frac{1}{2} \left[\begin{array}{cc} (-\eta_q/\eta_p)^S + (-\eta_p/\eta_q)^S & (i/\eta_p)[(-\eta_q/\eta_p)^S - (-\eta_p/\eta_q)^S] \\ i\eta_p[(-\eta_p/\eta_q)^S - (-\eta_q/\eta_p)^S] & (-\eta_q/\eta_p)^S + (-\eta_q/\eta_p)^S \end{array} \right]. \end{aligned} \quad (6.25)$$

Let η_m be the admittance of the substrate. Then

$$\begin{bmatrix} B \\ C \end{bmatrix} = \frac{1}{2} \left[\begin{array}{c} (-\frac{\eta_p}{\eta_q})^S + (-\frac{\eta_q}{\eta_p})^S + \frac{i\eta_m}{\eta_p} [(-\frac{\eta_q}{\eta_p})^S - (-\frac{\eta_p}{\eta_q})^S] \\ \eta_m [(-\frac{\eta_p}{\eta_q})^S + (-\frac{\eta_q}{\eta_p})^S] + i\eta_p [(-\frac{\eta_p}{\eta_q})^S - (-\frac{\eta_p}{\eta_q})^S] \end{array} \right]. \quad (6.26)$$

Equation (2.67) gives

$$\begin{aligned} T &= \frac{4\eta_0\eta_m}{(\eta_0B + C)(\eta_0B + C)^*} \\ &= [16\eta_0\eta_m] \left[\{(\eta_0 + \eta_m)[(-\eta_q/\eta_p)^S + (-\eta_p/\eta_q)^S]\}^2 \right. \\ &\quad \times \left. \{[(\eta_0\eta_m/\eta_p) - \eta_p][(-\eta_q/\eta_p)^S - (-\eta_p/\eta_q)^S]\}^2 \right]^{-1}. \end{aligned} \quad (6.27)$$

If S is sufficiently large so that

$$\left(\frac{\eta_H}{\eta_L} \right)^S \gg \left(\frac{\eta_L}{\eta_H} \right)^S$$

which will usually be the case, this expression reduces to

$$T = \frac{16\eta_0\eta_m}{(\eta_H/\eta_L)^{2S} \{(\eta_0 + \eta_m)^2 + [(\eta_0\eta_m/\eta_p) - \eta_p]^2\}}. \quad (6.28)$$

6.2.3.3 Transmission in the pass band

In the pass band, the multilayer behaves as if it were a single layer of slightly variable optical thickness and admittance. Let us consider the case of $[(L/2)H(L/2)]^S$. Figure 6.7 shows part of the curve of equivalent admittance E for $[(L/2)H(L/2)]$. γ , the equivalent phase thickness, is also shown.

In the case of a real single transparent layer on a transparent substrate the reflectance oscillates between two limiting values which correspond to layer thicknesses of an integral number of quarter-waves. When the layer is equivalent to an even number of quarter-waves, that is a whole number of half-waves, it is an absentee layer and behaves as if it did not exist, so that the reflectance is that of the bare substrate. When the layer is equivalent to an odd number of quarter-waves, then, according to whether the index is higher or lower than that of the substrate, the reflectance will either be a maximum or a minimum. Thus if η_f is the admittance of the film, η_m of the substrate and η_0 of the incident medium, the reflectance will be $[(\eta_0 - \eta_m)/(\eta_0 + \eta_m)]^2$, corresponding to an even number of quarter-waves, and

$$\left(\frac{\eta_0 - (\eta_f^2/\eta_m)}{\eta_0 + (\eta_f^2/\eta_m)} \right)^2$$

corresponding to an odd number of quarter-waves. Regardless of the actual thickness of the film, we can draw two lines

$$R = \left(\frac{\eta_0 - \eta_m}{\eta_0 + \eta_m} \right)^2 \quad (6.29)$$

and

$$R = \left(\frac{\eta_0 - (\eta_f^2/\eta_m)}{\eta_0 + (\eta_f^2/\eta_m)} \right)^2 \quad (6.30)$$

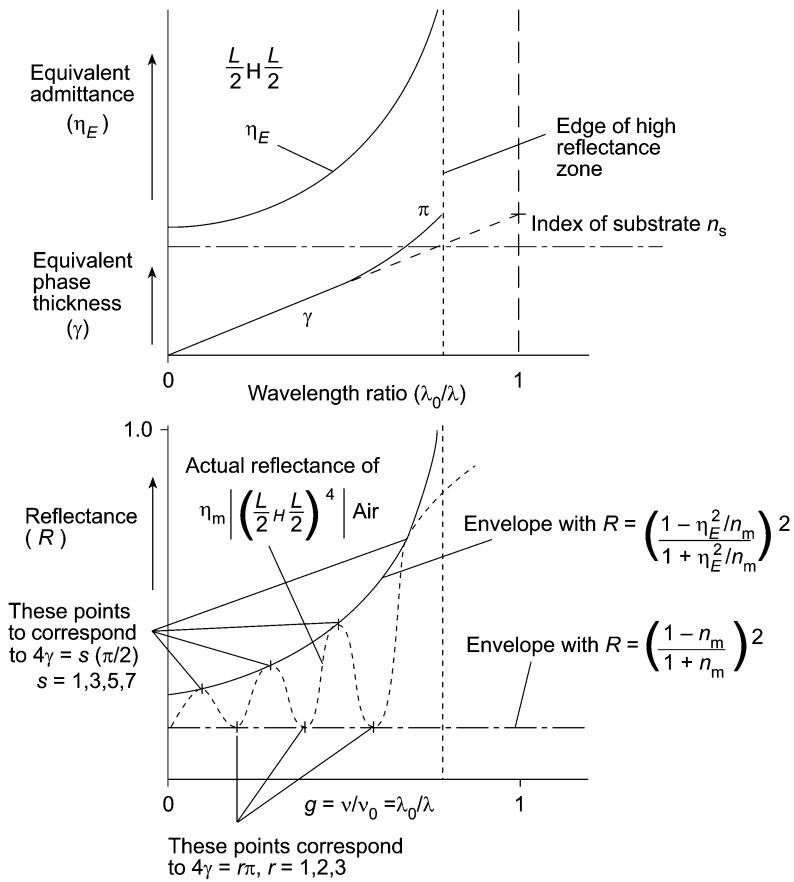


Figure 6.7. Diagram explaining the origin of the ripple in the pass band of an edge filter.

which are the loci of maximum and minimum reflectance values, that is, the envelope of the reflectance curve of the film. If the optical thickness of the film is D , then the actual positions of the turning values will be given by

$$D = 2n\lambda/4 \quad n = 0, 1, 2, 3, 4, \dots$$

for those in equation (6.29), and by

$$D = (2n + 1)\lambda/4$$

for those in equation (6.30), that is at wavelengths given by

$$\lambda = 4D/2n = 2D/n$$

and

$$\lambda = 4D/(2n + 1)$$

respectively.

We can now return to our multilayer. Since the multilayer can be replaced by a single film, the reflectance will oscillate between two values: the reflectance of the bare substrate

$$R = \left(\frac{\eta_0 - \eta_m}{\eta_0 + \eta_m} \right)^2 \quad (6.31)$$

and that given by

$$R = \frac{[\eta_0 - (E^2/\eta_m)]^2}{[\eta_0 + (E^2/\eta_m)]^2} \quad (6.32)$$

where we have replaced η_f in equation (6.29) by E , the equivalent admittance of the period. Equation (6.32) now represents a curve, since E is variable, rather than a line. To find the positions of the maxima and minima we look for values of $g = \lambda_0/\lambda$ for which the total thickness of the multilayer is a whole number of quarter-waves, which is the same as saying that the total equivalent phase thickness of the multilayer must be a whole number times $\pi/2$; an odd number corresponds to equation (6.32) and an even number to equation (6.31). If there are n periods in the multilayer, then the equivalent phase thickness will be $n\gamma$, which will be a multiple of $\pi/2$ when the equivalent phase thickness of a single period, γ , is a multiple of $\pi/2n$, i.e.

$$\gamma = s\pi/2n \quad s = 1, 3, 5, 7, \dots \quad \text{corresponding to (6.32)}$$

and

$$\gamma = r\pi/n \quad r = 1, 2, 3, 4, \dots \quad \text{corresponding to (6.31).}$$

At the very edge of the pass band, the equivalent phase thickness is π and so we might expect that the multilayer should act as an absentee layer. However, the equivalent admittance at that point is either zero or infinite and so the multilayer cannot be treated in this way, and, in fact, we apply the expressions (6.21)–(6.24), which we have already derived.

Figure 6.7 illustrates the situation where a four-period multilayer has been taken as an example. The important point, however, is that the envelopes of the reflectance curve do not vary with the number of periods.

The reason for the excessive ripple in the pass band of a filter is now clear. It is due to mismatching of the equivalent admittances of the substrate, multilayer stack, and medium. To reduce the ripple, better matching is required.

6.2.3.4 Reduction of pass-band ripple

There are a number of different approaches for reducing ripple. The simplest approach is to choose a combination which has an equivalent admittance similar to that of the substrate. Provided the reflection loss due to the bare substrate is not too great, this method should yield an adequate result. Figure 6.3 shows

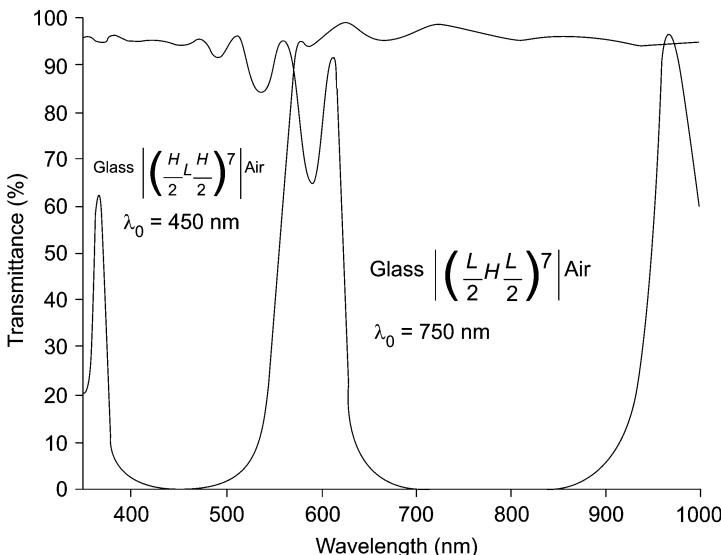


Figure 6.8. Computed transmittance of a 15-layer longwave-pass filter and a 15-layer shortwave-pass filter.

that the combination $[(H/2)L(H/2)]$ where $\eta_H = 2.35$, $\eta_L = 1.35$, should give a reasonable performance as a longwave-pass filter on glass, and this is indeed the case. The performance of such a filter is shown in figure 6.8. For a shortwave-pass filter, the combination $[(L/2)H(L/2)]$ is better and this is also shown in figure 6.8. Often, however, the materials which are available do not yield a suitable equivalent admittance and other measures to reduce ripple must be adopted.

One method which is very straightforward has been suggested by Welford [5] but does not seem to have been much used. This is simply to vary the thicknesses of the films in the basic period so that the equivalent admittance is altered to bring it nearer to the desired value. For this method to be successful, the reflectance from the bare substrate must be kept low and the substrate should have a low index. Glass in the visible region is quite satisfactory, but the method could not be used with, for example, silicon and germanium in the infrared without modification.

The more usual approach is to add matching layers at either side of the multilayer to match it to the substrate and to the medium. If a quarter-wave layer of admittance η_3 is inserted between the multilayer and substrate, and a quarter-wave layer of admittance η_1 between the multilayer and medium, then good matching will be obtained if

$$\eta_3 = (\eta_m E)^{1/2} \quad \text{and} \quad \eta_1 = (\eta_0 E)^{1/2}. \quad (6.33)$$

The layers are simply acting as antireflection layers between the multilayer and its surroundings. As a quick check that this does give the required performance we can compute the behaviour of the multilayer, considering just those wavelengths where the multilayer is equivalent either to an odd or to an even number of quarter-waves and to plot as before the envelope of the reflectance curve. At wavelengths where the multilayer acts like a quarter-wave, the equivalent admittance of the assembly is just

$$Y = \frac{\eta_1^2 \eta_3^2}{E^2 \eta_m}$$

so that the reflectance is

$$R = \left(\frac{\eta_0 - (\eta_1^2 \eta_3^2 / E^2 \eta_m)}{\eta_0 + (\eta_1^2 \eta_3^2 / E^2 \eta_m)} \right)^2 \quad (6.34)$$

which will be zero for

$$\eta_1^2 \eta_3^2 = E^2 \eta_m \eta_0. \quad (6.35)$$

When the multilayer acts like a half-wave it is an absentee, and the reflectance is

$$R = \left(\frac{\eta_0 - (\eta_1^2 \eta_m / \eta_3^2)}{\eta_0 + (\eta_1^2 \eta_m / \eta_3^2)} \right)^2 \quad (6.36)$$

which is zero if

$$\frac{\eta_1^2}{\eta_3^2} = \frac{\eta_0}{\eta_m}. \quad (6.37)$$

Solving equations (6.35) and (6.37) for η_1 and η_3 gives equation (6.33), as we expected.

If ideal matching layers do not exist, the suitability of any available materials can quickly be checked by substituting the appropriate values in equations (6.34) and (6.36).

Figure 6.9 shows a shortwave-pass filter before and after the matching layers have been added. The final reflectance envelopes are given by equations (6.34) and (6.36). The computed performance of the filter is shown in figure 6.10. As the value of g increases from 1.25, the ripple becomes a little greater than that predicted by the envelopes. This is because the envelopes were calculated on the basis of quarter-wave matching layers, and this is strictly true for $g = 1.25$ only.

6.2.3.5 Summary of design procedure so far

We have now established a simple design procedure for edge filters. First, two materials of different refractive index which are transparent in the region where transmission is required are chosen and used to form a multilayer of the form $[(L/2)H(L/2)]^S$ or $[(H/2)L(H/2)]^S$. Generally, it is better to choose as high a

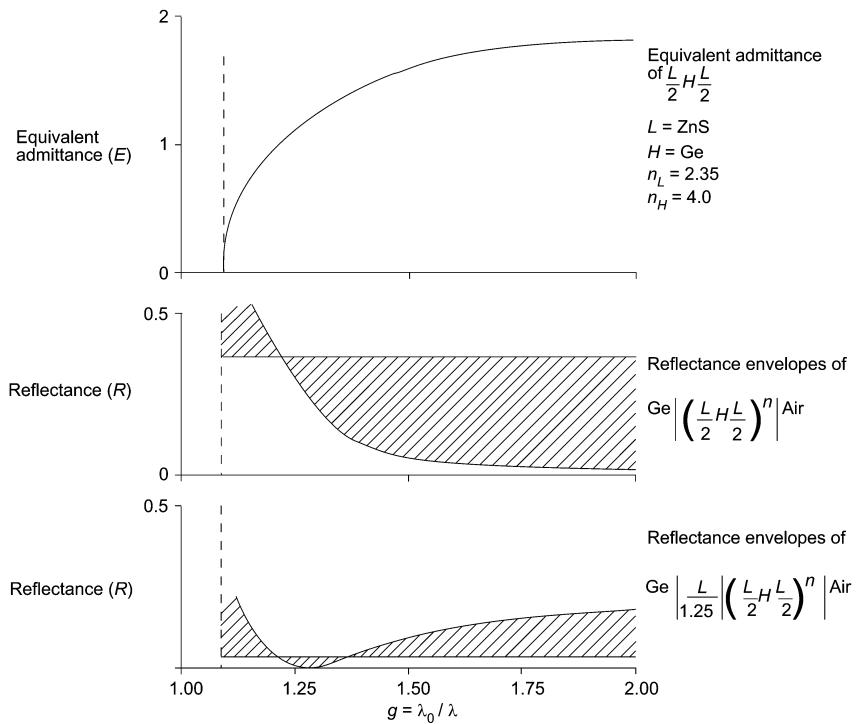


Figure 6.9. Steps in the design of a shortwave-pass filter using zinc sulphide and germanium on a germanium substrate.

ratio of refractive indices as possible to give the widest rejection zone and also the maximum rejection for a given number of periods. The width of the rejection zone is given by equation (6.7) or (6.8) and is plotted in figure 5.7. The level of rejection at the edges of the zone is given by equations (6.19), (6.20), (6.23) and (6.24) and at the centre of the zone by equation (6.28). Next, the equivalent admittance of the stack must be calculated. This can be done either by a computer or by using the design curves given in figure 6.5. The formulae given in equations (6.13) for E/η_p at $g = 2$ will be found useful as a guide to interpolating curves. The reflectance envelopes can now be drawn using the formulae (6.31) and (6.32). This will immediately give some idea of the likely ripple. The positions of the peaks and troughs of the ripple can, if necessary, be found using the curves of γ in figure 6.6 and the method given on p 237. If this ripple is adequate the next step can be omitted and the design can proceed to the final step. If the ripple is not adequate then matching layers between multilayer and substrate, and multilayer and medium should be inserted. These should be quarter wavelength films at the most important wavelength and should have admittances as nearly as possible

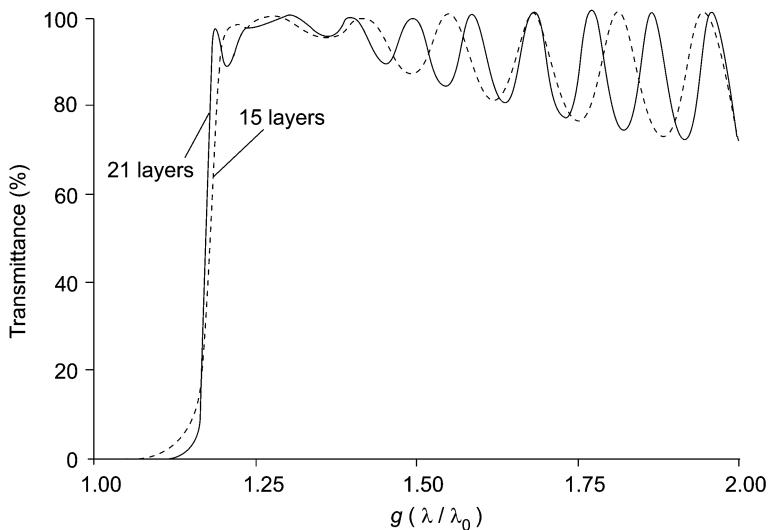


Figure 6.10. The calculated performance of filters designed according to figure 6.9 with design:

$$\text{Air} \mid (0.5L H 0.5L)^q L/1.25 \mid \text{Ge}$$

with $n_L = 2.35$, $n_H = 4.0$, $n_{\text{Ge}} = 4.0$, and $n_{\text{Air}} = 1.00$. (a) $q = 7$ (b) $q = 10$.

given by

$$\eta_1 = (\eta_0 E)^{1/2} \quad \eta_3 = (\eta_m E)^{1/2} \quad (6.33)$$

where η_1 is between the multilayer and medium and η_3 between the multilayer and substrate. Generally materials with the exact values will not be available and a compromise must be made. To test the effectiveness of the compromise the new reflectance envelope curves can be calculated using equations (6.34) and (6.36). If this is satisfactory, the next step is to calculate the actual performance on a computer. This is advisable because the quarter-wave matching layers are effective over a narrower region than assumed in equations (6.34) and (6.36). From the curve produced by the computer, the monitoring wavelength and thicknesses of the layers to position the characteristic at the correct wavelength can be calculated. The method is illustrated by the design of a shortwave-pass filter made from germanium and zinc sulphide on a germanium substrate as shown in figure 6.9 and 6.10.

A longwave-pass filter, designed by this method, with construction Air| $1.488L[(L/2)H(L/2)]^7 1.488H|\text{Ge}$ ($H = \text{PbTe}$ with $n_H = 5.3$, $L = \text{ZnS}$ with $n_L = 2.35$), is shown in figure 11.10.

6.2.3.6 More advanced procedures for eliminating ripple

At the present time, probably the most common technique for eliminating ripple, apart from that already discussed, is computer refinement. This was introduced into optical coating design by Baumeister [6] who programmed a computer to eliminate the effects of slight changes in the thicknesses of the individual layer on a merit function representing the deviation of the performance of the coating from the ideal. An initial design, not too far from ideal, was adopted and the thicknesses of the layers modified, successively, gradually to improve the performance. This is still the basis of the technique. The optimum thickness of any one layer is not independent of the thicknesses of the other layers so that the changes in thickness at each iteration cannot be large without running the risk of instability. Computer speed and capacity has increased considerably since the early work of Baumeister, but the essentials of the method are still the same. Rather than change the layers successively, it is more usual to estimate changes which should be made in all the layers. These changes are then made simultaneously and the new function of merit computed. New changes are then estimated and the process repeated. The way in which changes to be made are assessed is the principal difference between the techniques in frequent use. If the function of merit is considered as a surface in $(p + 1)$ -dimensional space with p independent variables being layer thicknesses, then a common method involves determining the direction of greatest slope of the merit surface and then altering the layer thicknesses so as to move along it, computing the new figure of merit and repeating the process. A battery of techniques for ensuring rapid convergence exists, and for further details the book by Liddell [7] should be consulted.

Less usual is complete design synthesis with no starting solution. This is still very much a research area and at the time of writing the most impressive results are those of Dobrowolski and Lowe [8].

Computer refinement is a very powerful design aid but it can only function with an initial design. It then finds a modified design with an improved performance and repeats the process until stopped or until the performance reaches a maximum. This maximum will normally be simply a local maximum rather than the best possible performance, and the most useful way of ensuring that the maximum reached will be sufficiently high is to start from an initial design which is sufficiently good. The better the performance required, the better must be the initial design. Thus the existence of efficient computer refinement techniques does not in any way imply that the analytical design methods are obsolete and can be discarded. Refinement should be looked upon as a way of making a good design better. Applied to a poor design, computer refinement techniques usually yield disappointing results. For this reason, we continue with our examination of analytical techniques. It should always be remembered, however, that the manufacture of edge filters is not altogether an easy task, and unless the design performance of the simple design is being achieved in manufacture, there is little point in attempting anything more complicated until the sources of error have

been eliminated.

The first and obvious method for improving the design is to improve the efficiency of the matching layers. In the chapter on antireflection coatings there were many multilayer coatings discussed which gave a rather better performance than the single layer. Any of these coatings can be used to eliminate the ripple. The ultimate performance is obtained with an inhomogeneous layer, but, as we have seen, the difficulty with inhomogeneous layers is that, in all practical cases, it is impossible to manufacture a layer with a graded index terminating in an index below 1.35, which means that there is always some small residual ripple. Jacobsson [9] has, however, considered briefly the matching of a multilayer longwave-pass filter $[(H/2)L(H/2)]^6$, consisting of germanium with an index of 4.0 and silicon monoxide with an index of 1.80, to a germanium substrate by means of an inhomogeneous layer. His paper shows the three curves reproduced in figure 6.11. The first curve 1 is the multilayer on a glass substrate of index 1.52. Since, in the pass band, the equivalent admittance of the multilayer falls gradually from $(1.8 \times 4.0)^{1/2} = 2.7$ to zero as the wavelength approaches the edge, it will be a value not too different from the index of the substrate in the vicinity of the edge. The transmission near the edge is, therefore, high, as we might expect. When, as in curve 2, the same multilayer is deposited on a germanium substrate of index 4.0, the severe mismatching causes a very large ripple to appear. With an inhomogeneous layer between the germanium substrate and the multilayer and with the index varying from that of germanium next to the substrate to 1.52 next to the multilayer, the performance achieved, curve 3, is almost exactly that of the original multilayer on the glass substrate.

One of the examples examined by Baumeister was a shortwave-pass filter, and the design that he eventually obtained suggested a new approach to Young and Cristal [10]. It was mentioned in chapter 3 that Young had devised a method for designing antireflection coatings based on the quarter-wave transformer used in microwave filters. The antireflection coating takes the form of a series of quarter-waves with refractive indices in steady progression from the index of one medium to the index of the other. Young has given a series of tables enabling antireflection coatings of given bandwidth and ripple to be designed.

In their paper, Young and Cristal explain that they examined Baumeister's filter, and realised that the design might be written as a series of symmetrical periods with thicknesses increasing steadily from the middle of the stack to the outside, and they were struck by the resemblance which this bore to an antireflection coating in which each layer had been replaced by a symmetrical period. They then designed a coating by microwave techniques, to match the admittance at the centre of the filter, which they arbitrarily took as 0.6, to air, with admittance 1.0, at the outside, each layer being replaced by an equivalent period. The scheme is shown as filter B in table 6.1, where the thicknesses given by Young and Cristal for one of their filters have been broken down into their symmetrical periods. The performance of the filter is shown in figure 6.12 along with one other filter of their design and Baumeister's original design. The thicknesses are

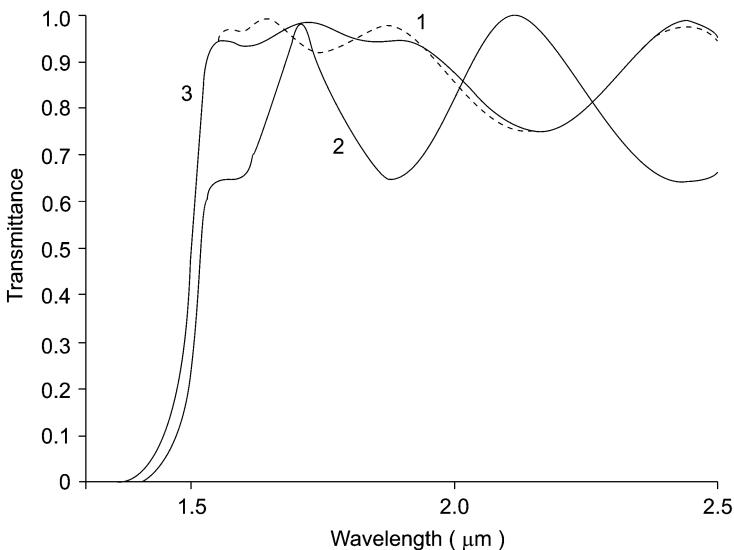


Figure 6.11. Reflectance versus wavelength of a multilayer on a substrate with index $n_{\text{sub}} = 1.52$ (curve 1), $n_{\text{sub}} = 4.00$ (curve 2) and on a substrate with $n_{\text{sub}} = 4.00$ with an inhomogeneous layer between substrate and multilayer (curve 3). (After Jacobsson [9].)

all shown in table 6.2. To simplify the discussion, Young and Cristal designed the filter to match with air on both sides of the multilayer, instead of, as is more usual, glass on one side and air on the other.

Young and Cristal do not discuss their design procedure in detail, but, from the final design of the filter, it is possible to deduce it. First, the equivalent admittance of a single period was plotted, as in figure 6.13. The wavelength corresponding to 240° was chosen for optimising. From the value of equivalent admittance at 240° the value of 0.6 was probably selected intuitively as the value to use for the centre of the stack. An antireflection coating consisting of four layers, each three-quarter wavelengths thick, was designed to match this value to air, and the admittances of the layers computed. The admittances were then matched by that of three-layer symmetrical periods by altering thicknesses of each period, following the scheme shown in figure 6.13. This meant that the admittances were ideal but the thicknesses were not. However, the antireflection coating is not very susceptible to errors in layer thickness, and as can be seen from the curve in figure 6.12, the performance achieved is excellent.

A similar approach is to use one of the multilayer antireflection coatings mentioned in chapter 3. Since the equivalent admittance of a symmetrical period varies with wavelength, any optimising at one wavelength is strictly correct over only a narrow range, and a simple approach, such as this, is probably as good as a more complicated one. Taking 240° as corresponding to the design wavelength,

Table 6.1.

Layer number	Filter B		Filter D	
	Layer thickness	Periods	Layer thickness	Periods
1 Na ₃ AlF ₆	47.50°	47.50°	48.5°	48.5°
2 ZnS	95.00°	95.00°	97.0°	97.0°
3 Na ₃ AlF ₆	93.25°	{ 47.50° 45.75° }	94.5°	{ 48.5° 46.0° }
4 ZnS	91.50°	91.50°	92.0°	92.0°
5 Na ₃ AlF ₆	90.00°	{ 45.75° 44.25° }	90.25°	{ 46.0° 44.25° }
6 ZnS	88.50°	88.50°	88.5°	88.5°
7 Na ₃ AlF ₆	87.50°	{ 44.25° 43.25° }	86.63°	{ 44.25° 42.38° }
8 ZnS	86.50°	86.50°	84.75°	84.75°
9 Na ₃ AlF ₆	86.50°	{ 43.25° 43.25° }	84.75°	{ 42.38° 42.38° }
10 ZnS	86.50°	86.50°	84.75°	84.75°
11 Na ₃ AlF ₆	87.50°	{ 43.25° 44.25° }	86.63°	{ 42.38° 44.25° }
12 ZnS	88.50°	88.50°	88.5°	88.5°
13 Na ₃ AlF ₆	90.00°	{ 44.25° 45.75° }	90.25°	{ 44.25° 46.0° }
14 ZnS	91.50°	91.50°	92.0°	92.0°
15 Na ₃ AlF ₆	93.25°	{ 45.75° 47.50° }	94.5°	{ 46.0° 48.5° }
16 ZnS	95.00°	95.00°	97.0°	97.0°
17 Na ₃ AlF ₆	47.50°	47.50°	48.5°	48.5°

The second column in each case gives the filter split into its component periods.

we find the value for equivalent admittance of the single period to be 0.8. We want the periods in the final design to be symmetrically placed around this period, so we find the starting admittance at the centre of the stack by assuming that this period should be able to act as a $3\lambda/4$ antireflection coating between the centre and the outside air. The admittance at the centre of the filter should therefore be $0.8^2 = 0.64$. Next, we design a four-layer antireflection coating to replace this basic period, using the formulae

$$\begin{aligned}\eta_1 &= \eta_0(\eta_s/\eta_0)^{1/5} & \eta_3 &= \eta_0(\eta_s/\eta_0)^{3/5} \\ \eta_2 &= \eta_0(\eta_s/\eta_0)^{2/5} & \eta_4 &= \eta_0(\eta_s/\eta_0)^{4/5}\end{aligned}$$

where η_0 is air and η_s the admittance at the centre. Taking $\eta_0 = 1.0$ and

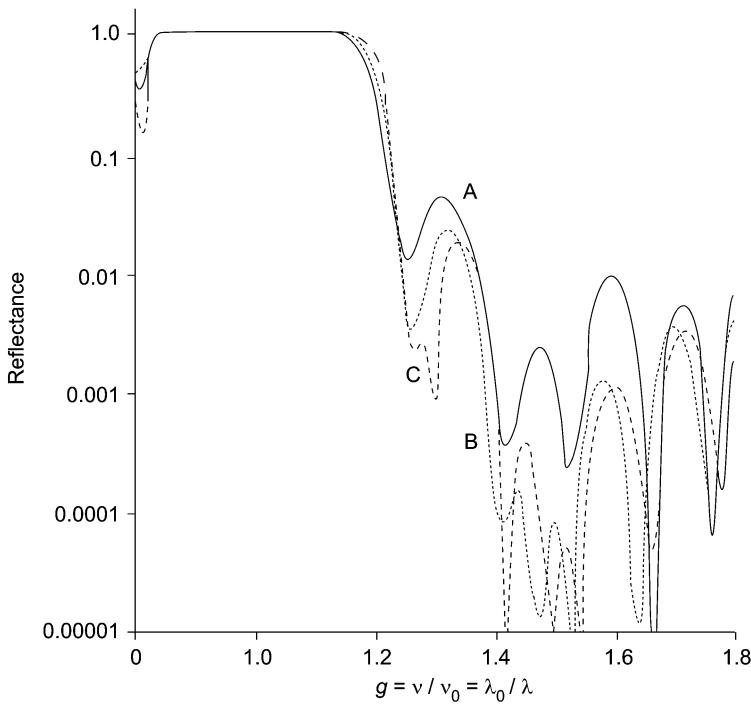


Figure 6.12. Reflectance of the three shortwave-pass filter designs, A, B and C, having unequal layer thickness. (After Young and Cristal [10].)

$\eta_s = 0.64$, these admittances are then

$$\eta_1 = 0.91 \quad \eta_2 = 0.84 \quad \eta_3 = 0.76 \quad \eta_4 = 0.70.$$

The values of total phase thickness πg at which the single period has equivalent admittance corresponding to these values are

$$\pi g_1 = 259^\circ \quad \pi g_2 = 245^\circ \quad \pi g_3 = 234^\circ \quad \pi g_4 = 226^\circ.$$

For each period to have the appropriate admittance at the design wavelength, the phase thicknesses of the layers measured at the monitoring wavelength are given by

$$\text{Period 1} \left\{ \begin{array}{l} \frac{L}{2} 45^\circ \times \frac{\pi g_1}{240^\circ} \\ H 90^\circ \times \frac{\pi g_1}{240^\circ} \\ \frac{L}{2} 45^\circ \times \frac{\pi g_1}{240^\circ} \end{array} \right. \quad \text{Period 2} \left\{ \begin{array}{l} \frac{L}{2} 45^\circ \times \frac{\pi g_2}{240^\circ} \\ H 90^\circ \times \frac{\pi g_2}{240^\circ} \\ \frac{L}{2} 45^\circ \times \frac{\pi g_2}{240^\circ} \end{array} \right.$$

Table 6.2.

Number of layers	Thickness (degrees)		
	Filter A	Filter B	Filter C
1	46.00	47.50	46.60
2	96.00	95.00	93.20
3	93.20	93.25	91.70
4	91.70	91.50	90.20
5	91.10	90.00	89.15
6	89.75	88.50	88.10
7	87.50	87.50	87.30
8	86.05	86.50	86.50
9	86.70	86.50	86.50
10	86.05	86.50	86.50
11	87.50	87.50	87.30
12	89.75	88.50	88.10
13	91.10	90.00	89.15
14	91.70	91.50	90.20
15	93.20	93.25	91.70
16	96.00	95.00	93.20
17	46.00	47.50	46.60

Filter A: The half of Baumeister's filter on the air side repeated symmetrically. (The design is referred to as design IX in Baumeister's paper.)

Filter B: New design based on a prototype transformer with a fractional bandwidth of 1.5.

Filter C: New design based on a prototype transformer with a fractional bandwidth of 1.6.

and so on. The results are shown in table 6.1, filter D. The transmission of filter D is shown in figure 6.14.

Thelen [3] has pointed out that the rapid variation of equivalent admittance near the edge of the filter is the major source of difficulty in edge filter design. It is a simple matter to match the multilayer to the substrate where the equivalent admittance curve is flat, some distance from the edge, but the variations near the edge usually give rise, with simple designs, to a pronounced dip in the transmission curve. Thelen has devised an ingenious method of dealing with this dip, involving the equivalent of a single-layer antireflection coating. Between the main or primary multilayer, which consists of a number of equal basic periods, Thelen places a secondary multilayer, similar to the first but shifted in thickness so that, in the centre of the steep portion of the admittance curve, the equivalent admittance of the secondary is made equal to the square root of the equivalent admittance of the primary times the admittance of the substrate. The number of

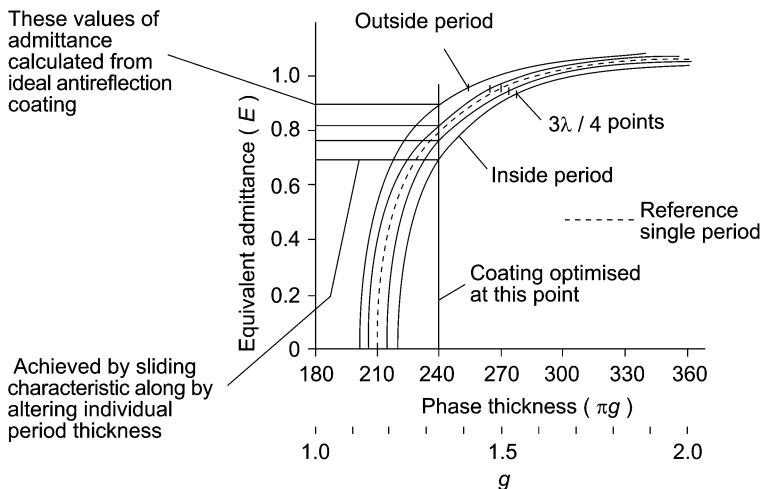


Figure 6.13. The admittance of the ideal four-layer antireflection coating to match air to the admittance of 0.6 are marked along the equivalent admittance axis. The reference single period is shown dotted and the values marked on the g axis refer to this period. By altering the total thickness of each period relative to this reference the four displaced solid line curves are obtained in such a way that the four symmetrical periods have the desired admittances at the wavelengths that correspond to a reference phase thickness of 240° .

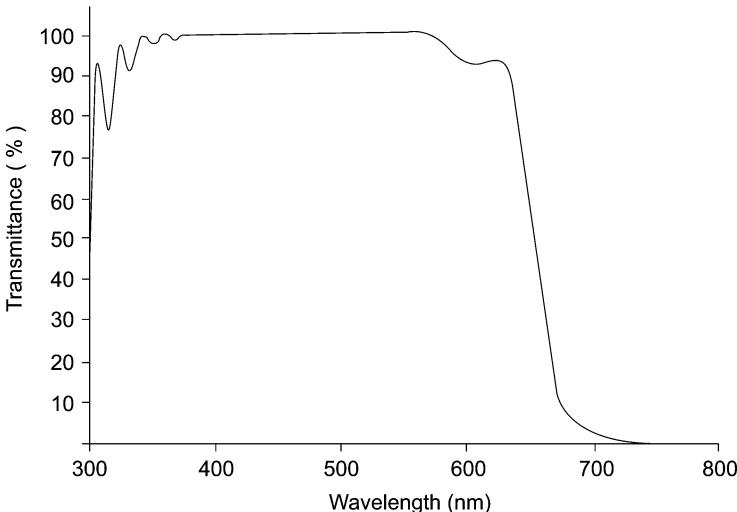


Figure 6.14. The computer transmittance of the shortwave-pass filter of design D of table 6.1. The reference wavelength, λ_0 , is 800 nm.

secondary periods is chosen to make the thickness at this point an odd number of quarter-waves and to satisfy completely the antireflection condition. Figure 6.15 shows the performance he achieved.

Seeley [11] has developed a different method of adapting results obtained in the synthesis of lumped electrical circuits for use in thin-film optical filters. One of the features of Young's method is that the refractive indices cannot be specified in advance, and as the range of available indices is limited this can lead to difficulties. In certain cases this can be avoided, as we have seen, by constructing three-layer periods with the appropriate equivalent indices, but even this has its limitations. Seeley, therefore, searched for another method which would permit the designer to specify the indices right from the start and to achieve the final performance by varying the thicknesses of the various layers. In a lumped electrical filter, consisting of inductances and capacitances, one parameter only is specified, the admittance. In the thin-film filter there are two parameters for each layer, the refractive index and the thickness. Thus it is possible for the optical designer to fix the values of the refractive indices of the multilayer filter in advance and then to compute the layer thickness by analogy with the lumped filter. As Welford [5] has pointed out, the analogy between thin-film assemblies and lumped electric filters is not exact. Thin films behave, in fact, in the same manner as lengths of waveguides. Seeley, however, devised a way of making the analogy exact, although only at one frequency. At all other frequencies, the analogy is only approximate. If the frequency chosen for exact correspondence is made the cut-off point of the filter, then the performance of the optical filter is found to be sufficiently close to that of the electrical filter over the usual working range. The techniques for optimising the performance of electrical filters are well established.

Seeley's method starts with an electrical filter of the desired type—longwave-pass, shortwave-pass or band-pass—whose performance is known to be optimum. The elements of the electrical filter are then converted by a step-by-step process into an equivalent circuit which is an exact analogue of the thin-film multilayer at one frequency. The process is shown in figure 6.16. In his design work, Seeley usually chooses electrical filters which have been designed using the Tchebyshev equal ripple polynomial. This polynomial allows the best fit to a square pass band when both edge steepness and ripple in the pass band are taken into account. From this, Seeley and Smith [12] have given simple rules for longwave-pass filters.

1. The optical admittance of the substrate n_m should lie between η_H and η_L , the admittances of the high- and low-index layers of the multilayer. If this is not satisfied, then a matching layer or combination of layers will be necessary between the substrate and the multilayer.

2. The first layer at the substrate should be high if $\eta_H/\eta_m > \eta_m/\eta_L$, and low if $\eta_m/\eta_L < \eta_H/\eta_m$.

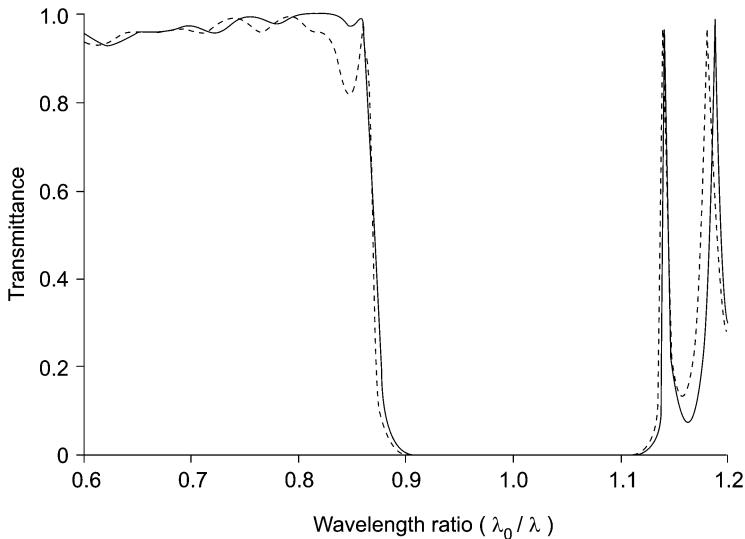


Figure 6.15. Comparison of the computed performance of the filters:

$$1.00|(0.5H\ L\ 0.5H)^{15}|1.52 \text{ (dashed line)}$$

and

$$1.00|(0.5H\ L\ 0.5H)^{12}[(1/1.05)(0.5H\ L\ 0.5)^3]|1.52 \text{ (solid line)}$$

with $n_H = 2.3$, $n_L = 1.56$. (After Thelen [3].)

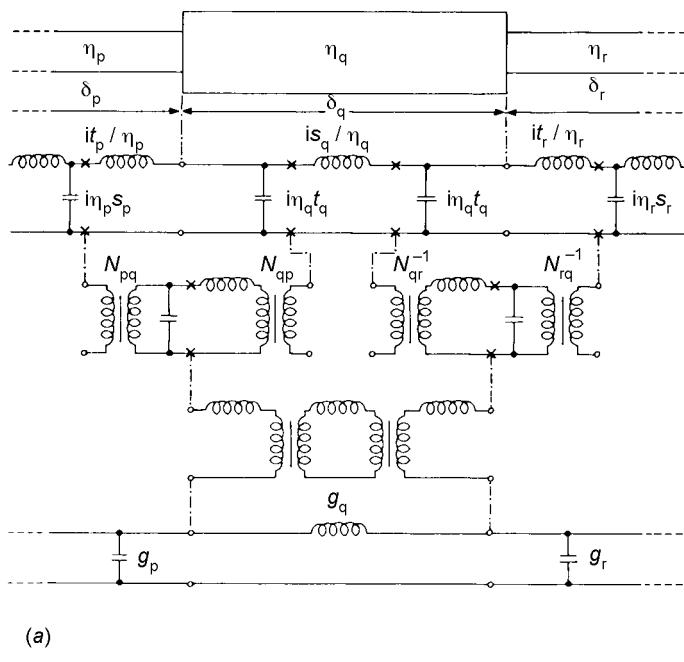
3. The fractional ripple in the pass band will be

$$\left(\frac{\eta_H}{\eta_m} - \frac{\eta_m}{\eta_L} \right)^2 \left(\frac{\eta_H}{\eta_m} + \frac{\eta_m}{\eta_L} \right)^{-2}.$$

4. For filters on germanium substrates using as layer materials lead telluride and zinc sulphide, the phase thicknesses should be in the proportions shown in table 6.3. The first layer at the substrate and all other odd layers, including the antireflection layer, are ZnS ($n = 2.2$). The remaining (even) layers are PbTe ($n = 5.1$). The substrate, germanium, has an index of 4.0.

5. Since the low-index material is usually good for matching the substrate to air, the front layer of the multilayer section of the filter should have a high index.

The computed transmittances of the designs given in table 6.3 are given in figure 6.17. The method is described in greater detail by Seeley *et al* [11].



(a)

Figure 6.16. The conversion used by Seeley in diagrammatic form. High-index layers are first replaced by a T circuit and low-index ones by a π circuit. (a) The step-by-step process by which Seeley converts a multilayer thin-film filter into a lumped electric filter in such a way that the elements of the electric filter can be identified with the optical thickness of the films, the indices of the films being specified completely independently. (Courtesy of Dr J S Seeley.)

(Opposite page) The manipulation takes place at the cut-off frequency of the lumped circuit and all variable quantities are normalised to that frequency. The scheme leads to a fairly complicated set of equations for $\dots \delta_p, \delta_q, \delta_r \dots$ in terms of $\dots g_p, g_q, g_r \dots$, which cannot be solved analytically but require iteration. Approximate solutions have been derived and are as follows:

$$\text{High-index layers: } \sin \delta_p \simeq \frac{g_p}{(\eta_H/\eta_m) + (\eta_L/\eta_m)}$$

$$\text{Low-index layers: } \sin \delta_q \simeq \frac{g_q}{(\eta_m/\eta_L) + (\eta_m/\eta_H)}$$

δ being between 0 and $\pi/2$ for longwave-pass filters and $\pi/2$ and π for short-wave-pass filters.

The admittance levels in the derivation of these two expressions have been normalised to the terminating admittance (of the substrate), so that for η_p we have written η_H/η_m and for η_q , η_L/η_m , η_H and η_L being the admittances of the high- and low-index layers respectively.

$$\begin{array}{c}
 \text{High index} \quad \text{Low index} \quad \text{High index} \\
 \left[\begin{array}{cc} \cos \delta_p & (\sin \delta_p)/\eta_p \\ i\eta_p \sin \delta_p & \cos \delta_p \end{array} \right] \times \left[\begin{array}{cc} \cos \delta_q & (\sin \delta_q)/\eta_q \\ i\eta_q \sin \delta_q & \cos \delta_q \end{array} \right] \times \left[\begin{array}{cc} \cos \delta_r & (\sin \delta_r)/\eta_r \\ i\eta_r \sin \delta_r & \cos \delta_r \end{array} \right] \\
 \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right] \left[\begin{array}{cc} 1 & i\eta_p/\eta_p \\ i\eta_p \eta_p & 1 \end{array} \right] \times \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right] \left[\begin{array}{cc} 1 & i\eta_q/\eta_q \\ i\eta_q \eta_q & 1 \end{array} \right] \times \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right] \left[\begin{array}{cc} 1 & i\eta_r/\eta_r \\ i\eta_r \eta_r & 1 \end{array} \right] \\
 \left[\begin{array}{cc} 1 - (\eta_q t_p / \eta_p) & i\eta_p / \eta_p \\ i\eta_q \eta_q & 1 \end{array} \right] \left[\begin{array}{cc} 1 & i\eta_r / \eta_r \\ i\eta_q \eta_q - (\eta_q t_r / \eta_r) & 1 \end{array} \right] \\
 \left[\begin{array}{cc} N_{pq} & 0 \\ 0 & N_{pq}^{-1} \end{array} \right] \left[\begin{array}{cc} 1 & iN_{qp}t_p/N_{pq}\eta_p \\ iN_{pq}\eta_q t_q/N_{qp} & 1 - (\eta_q t_p / \eta_p) \end{array} \right] \left[\begin{array}{cc} N_{qp} & 0 \\ 0 & N_{qp}^{-1} \end{array} \right] \left[\begin{array}{cc} N_{qr}^{-1} & 0 \\ 0 & N_{qr} \end{array} \right] \left[\begin{array}{cc} 1 - (\eta_q t_r / \eta_r) & iN_{qr}t_r/N_{qr}\eta_r \\ iN_{qr}\eta_q t_q/N_{qr} & 1 \end{array} \right] \left[\begin{array}{cc} N_{qr}^{-1} & 0 \\ 0 & N_{qr} \end{array} \right] \\
 \left[\begin{array}{cc} 1 & 0 \\ iN_{pq}\eta_q t_q/N_{qp} & 1 \end{array} \right] \times \left[\begin{array}{cc} 1 & iN_{qp}t_p/N_{pq}\eta_p \\ 0 & 1 \end{array} \right] \left[\begin{array}{cc} N_{qp} & 0 \\ 0 & N_{qp}^{-1} \end{array} \right] \left[\begin{array}{cc} 1 & i\eta_q / \eta_q \\ 0 & 1 \end{array} \right] \left[\begin{array}{cc} N_{qr}^{-1} & 0 \\ 0 & N_{qr} \end{array} \right] \left[\begin{array}{cc} 1 & iN_{qr}t_r/N_{qr}\eta_r \\ 0 & 1 \end{array} \right] \times \left[\begin{array}{cc} 1 & 0 \\ iN_{qr}\eta_q t_q/N_{qr} & 1 \end{array} \right] \\
 \left[N_{qp}/N_{qr} \right]^{-1} \left[(N_{qp}N_{qr}S_q/\eta_q) + (N_{qp}t_r/N_{qr}\eta_r) + (N_{qr}t_p/N_{qp}\eta_p) \right] \\
 \left[\begin{array}{cc} 1 & 0 \\ ig_p & 1 \end{array} \right] \times \left[\begin{array}{cc} 1 & ig_q \\ 0 & 1 \end{array} \right] \times \left[\begin{array}{cc} 1 & 0 \\ ig_r & 1 \end{array} \right]
 \end{array}$$

$$\begin{aligned}
 s &= \sin \delta \\
 t &= \tan \delta/2
 \end{aligned}$$

Figure 6.16. (b) Matrix manipulations corresponding to figure 6.16(a).

Table 6.3.

Layer number	Relative thickness	
	Longwave-pass	Shortwave-pass
1 and 14	0.55	1.25
2 and 13	0.82	1.11
3 and 12	0.92	1.05
4 and 11	0.96	1.025
5 and 10	0.98	1.015
6 and 9	0.99	1.01
7 and 8	1.0	1.0
15 (antireflection)	2.0	0.5

6.2.3.7 Practical filters

Because the stop band of the multilayer edge filter is limited in extent, it is usually necessary for practical filters to consist of a multilayer filter together with additional filters which give the broad rejection region that is almost always required. These additional filters may be multilayer and some methods of broadening the stop band in this way are mentioned in the following section. Usually they are absorption filters having wide rejection regions but inflexible characteristics. These absorption filters may be combined with multilayer filters in a number of different ways. They may simply be placed in series with the substrates carrying the multilayers, the substrates may themselves be the absorption filters or the multilayer materials may also act as thin-film absorption filters.

In the visible and near ultraviolet regions there is available a wide range of glass filters which solve most of the problems, particularly those connected with longwave-pass filters. In the infrared, the position is rather more difficult, and often the complete filter consists of several multilayers which are necessary to connect the edge of the stop band to the nearest suitable absorption filter. Figure 6.18 shows a longwave-pass filter for the infrared. Figure 6.19 gives some of the infrared absorption filters which have shortwave-pass characteristics. For longwave-pass characteristics, semiconductors such as silicon, with an edge at $1 \mu\text{m}$, and germanium, with an edge at $1.65 \mu\text{m}$, are the most suitable. Indium arsenide, with an edge at $3.4 \mu\text{m}$, and indium antimonide, with edge at $7.2 \mu\text{m}$, are also useful, but because of the rather higher absorption they can only be used in very thin slices, around 0.013 cm for indium antimonide and only a little thicker for indium arsenide. This means that they tend to be extremely fragile and can only be produced in a circular shape of rather limited diameter, not usually greater than 2.0 cm .

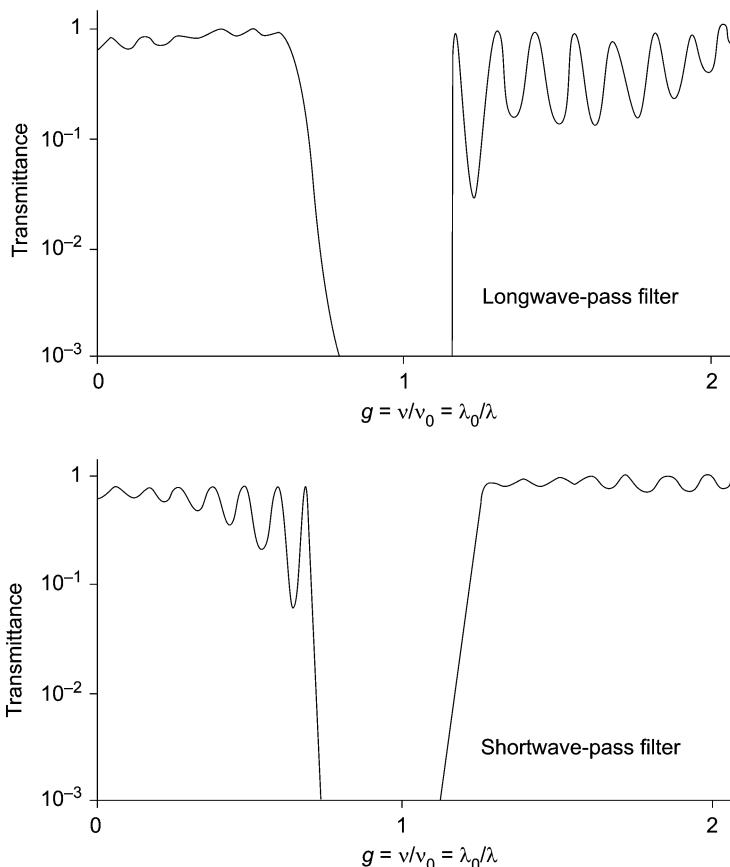


Figure 6.17. Computed transmittance of the 14-layer filters given in table 6.3. v_0 and λ_0 are the frequency and wavelength respectively at which the central layers are a quarter-wave in thickness. (After Seeley and Smith [12].)

The measured transmittance for a longwave-pass filter consisting of an edge filter together with an absorption filter is given in figure 6.20. This filter was originally designed to be used as a shortwave blocking filter with narrowband filters at $15\text{ }\mu\text{m}$. It consists of two components, a multilayer filter made from a lead telluride and zinc sulphide multilayer on a germanium substrate and placed in series with an indium antimonide filter. The very high rejection achieved can be seen from the logarithmic plot.

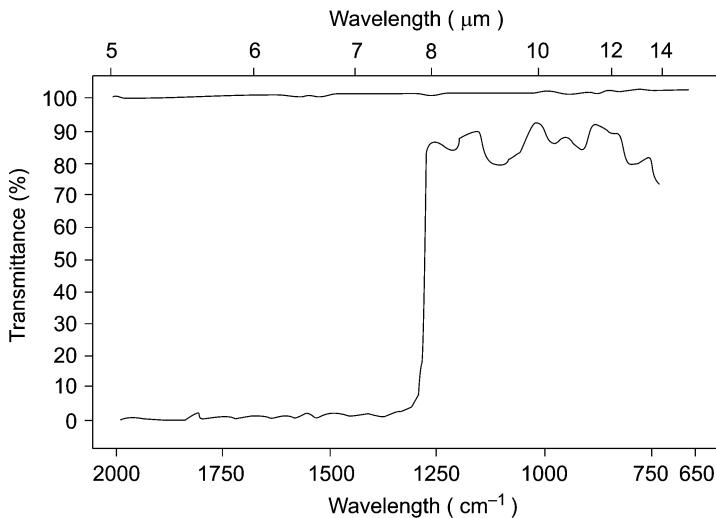


Figure 6.18. Measured transmittance of a practical longwave-pass filter with edge at 1250 cm^{-1} ($8\text{ }\mu\text{m}$). (Courtesy of OCLI Optical Coatings Ltd.)

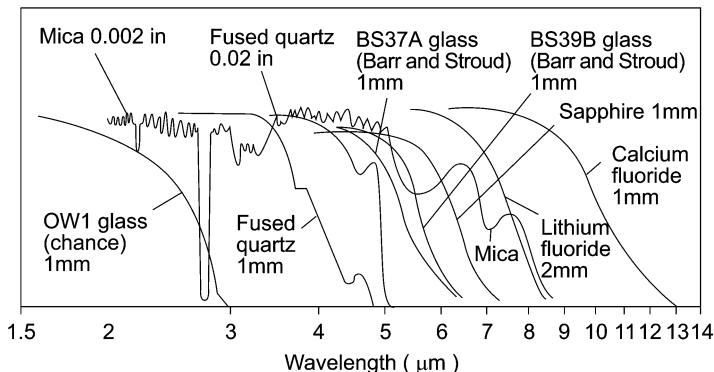


Figure 6.19. A selection of infrared materials which can be used as shortwave-pass absorption filters. (Courtesy of Sir Howard Grubb, Parsons & Co. Ltd.)

6.2.3.8 Extending the rejection zone by interference methods

The most convenient and straightforward way of extending the reflectance zone is to place a second stack in series with the first and to ensure that their rejection zones overlap. The second stack is best placed either on a second substrate or on the opposite side of the substrate from the first stack. Provided that the substrate is reasonably thick or slightly wedged, the transmission of the assembly is then

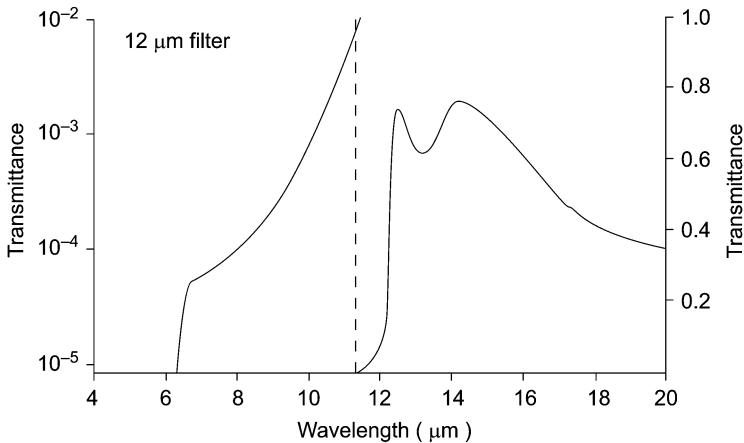


Figure 6.20. Measured transmittance of a multilayer blocking filter with edge at $12\text{ }\mu\text{m}$. A subsidiary indium antimonide filter is included to ensure good blocking at wavelengths shorter than $7\text{ }\mu\text{m}$. (After Seeley and Smith [12].)

given by equation (2.140)

$$T = \frac{1}{(1/T_a) + (1/T_b) - 1} \quad (6.38)$$

and a nomogram for calculating this is given in figure 2.15.

Occasionally it may happen that it is impossible to place the stacks on separate surfaces, and one stack must be deposited directly on top of the other. In this case it is necessary to take precautions to avoid the creation of transmission maxima. The problem has already been dealt with in chapter 5 where the extension of the high-reflectance zone of a quarter-wave stack was discussed (pp 202–9).

If we consider the assembly split into two separate multilayers, as shown in figure 5.12, then a transmission maximum will occur at any wavelength for which $(\phi_a + \phi_b)/2 = n\pi$, where $n = 0, \pm 1, \pm 2, \dots$. The height of this maximum is given by

$$T = \frac{|\tau_a^+|^2 |\tau_b^+|^2}{(1 - |\rho_a^-||\rho_b^+|)^2} = \frac{T_a T_b}{[1 - (R_a R_b)^{1/2}]^2}.$$

If there is no absorption, this expression implies that, for low transmission at the maxima, R_a and R_b should be as dissimilar as possible. This can be achieved by using many layers to keep the reflectance of one multilayer as high as possible in the pass region of the other.

In slightly more quantitative terms, from the reflectance envelope, which does not vary with the number of periods, we can find the highest reflectance in

the pass region of either multilayer making up the composite filter. If we denote this reflectance by R_p , then we can be certain that the design will be acceptable if we choose a sufficiently high number of periods to make R_s , the lowest reflectance in the stop band of the other multilayer, sufficiently high to ensure that

$$\frac{(1 - R_p)(1 - R_s)}{[1 - (R_p R_s)^{1/2}]^2} \leq T_c \quad (6.39)$$

where T_c is some acceptable level for the transmission in the rejection zone of the complete filter. This formula will give a pessimistic result; the actual transmission achieved in practice will depend on the phase change as well as the reflectance.

The only other danger area is the region where the two high-reflectance bands are overlapping. There, it must be arranged that on no account is $(\phi_a + \phi_b)/2 = n\pi$. The method for dealing with this was described in the previous chapter where a layer of intermediate thickness was placed between the two quarter-wave stacks. The result is equivalent to placing two similar multilayers, both of the form $[(L/2)H(L/2)]^n$ or $[(H/2)L(H/2)]^n$, together.

Equation (6.39) also implies that some of the sections of the composite filter should have more periods than others. In the reduction of the ripple in the pass band of the basic multilayer, the ripple on the other side of the stop band is invariably increased. Thus, in the combination of, say, two multilayers, the rejection zone of one stack will overlap a region of high ripple, while the rejection zone of the other stack will overlap a region of relatively low ripple. Since high ripple means that R_p is high, the former stack should have more periods than the latter if the same level of rejection is required throughout the combined rejection region. Figure 6.21 shows two component edge filters which are combined in a single filter in figure 6.22. The severe ripple which occurs in one of the multilayers can be seen reflected in the rejection zone of the composite filter. This ripple is limited to part of the rejection zone only, and in order to reduce the effect, more periods are necessary in the appropriate multilayer.

6.2.3.9 Extending the transmission zone

The shortwave-pass filter, as it has been described so far, possesses a limited pass band because of the higher order stop bands. These are not always particularly embarrassing, but occasionally, as for example with some types of heat reflecting filters, a much wider pass band is required. The problem was first considered by Epstein [14] and was studied more extensively by Thelen [15].

Epstein's analysis was as follows. Let the multilayer be represented by S periods each of the form

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}.$$

If a single period is considered as if it were immersed in a medium of admittance

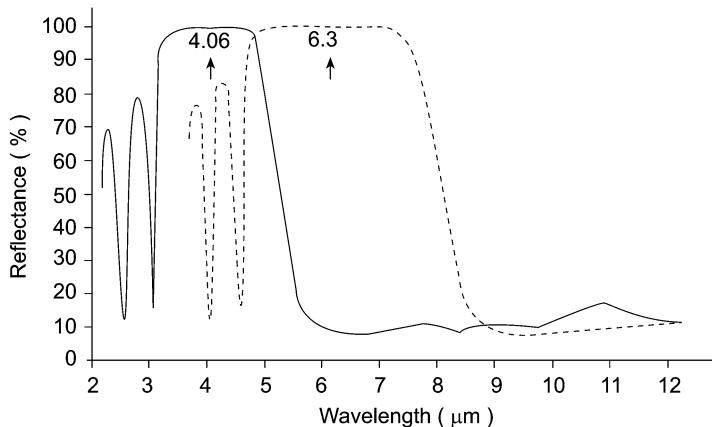


Figure 6.21. Measured reflectance of two longwave-pass stacks:

$$\text{A}|(0.5H\ L\ 0.5H)^4|\text{BaF}_2.$$

H and L are films of stibnite and chiolite a quarter-wave thick at $\lambda_0 = 4.06 \mu\text{m}$ or $6.3 \mu\text{m}$. A is air and the substrate is barium fluoride. (After Turner and Baumeister [13].)

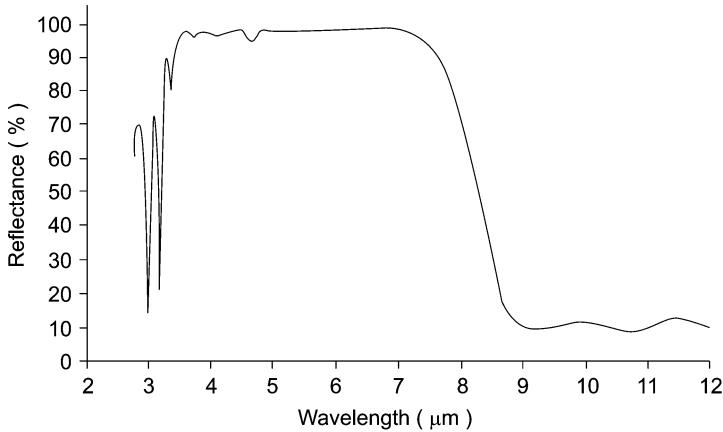


Figure 6.22. Measured reflectance of the two longwave-pass stacks of figure 6.21 superimposed in a single coating for an extended high-reflectance region. (After Turner and Baumeister [13].)

η , then the transmission coefficient of the period is given by

$$t = \frac{2\eta}{\eta\{(M_{11} + M_{22}) + [\eta M_{12} + (M_{21}/\eta)]\}}.$$

Let $t = |t|e^{i\tau}$; then

$$\frac{1}{2}\{(M_{11} + M_{22}) + [\eta M_{12} + (M_{21}/\eta)]\} = \frac{\cos \tau - i \sin \tau}{|t|}.$$

If the period is transparent, equating real parts gives

$$\frac{1}{2}(M_{11} + M_{22}) = \frac{\cos \tau}{|t|}.$$

Now, if light which has suffered two or more reflections at interfaces within the period is ignored, then

$$\tau \simeq \sum \delta$$

the total phase thickness of the period.

When $\sum \delta = n\pi$, $\cos \tau = \pm 1$, and, if $|t| < 1$, then

$$\left| \frac{1}{2}(M_{11} + M_{22}) \right| > 1$$

and a high-reflectance zone results. If, however, $|t| = 1$, then

$$\left| \frac{1}{2}(M_{11} + M_{22}) \right| = 1$$

and the high-reflectance zone is suppressed. In the simple form of stack,

$$\begin{aligned} [(L/2)H(L/2)]^S &\quad \text{or} \quad [(H/2)L(H/2)]^S \\ |t| = 1 &\quad \text{for} \quad \tau = 2r\pi \quad r = 1, 2, 3, 4, \dots \end{aligned}$$

and the even-order high-reflectance zones are therefore suppressed. As noted earlier, only a slight change in the relative thicknesses of the layers is enough to reduce t and turn the band into a high-reflectance zone.

Putting this result in another way, a zone of high reflectance potentially exists whenever the total optical thickness of an individual period of the multilayer is an integral number of half-waves, and the high-reflectance zone is prevented from appearing if, and only if, $|\tau| = 1$. This result has been used by Epstein in his paper to design a multilayer in which the fourth- and fifth-order reflectance bands were suppressed. Thelen has extended Epstein's analysis to deal with cases where any two and any three successive orders are suppressed and it is this method which we shall follow.

Following Epstein, Thelen [15] assumed a five-layer form, $ABCBA$, which involves three materials, for the basic period of the multilayer, and noted that if the period is thought of as immersed in a medium M , the combination AB becomes an antireflection coating for C in M at the wavelengths where suppression is required. In the construction of the final multilayer, the medium M can be considered first to exist between successive periods and then to suffer a progressive decrease in thickness until it just vanishes. The shrinking procedure

leaves unchanged the suppression of the various orders which has been arranged. M can therefore be chosen quite arbitrarily during the design procedure to be discarded later. The antireflection coating AB is of a type studied originally by Muchmore [16] and Thelen adapted his results as follows.

The various parameters of the layers are denoted by the usual symbols with the appropriate suffixes A , B , C and M .

Let layers A and B be of equal optical thickness, i.e.

$$\delta_A = \delta_B \quad (6.40)$$

and let

$$\eta_A \eta_B = \eta_C \eta_M. \quad (6.41)$$

Then the wavelengths for which unity transmittance will be achieved will be given by

$$\tan^2 \delta'_A = \frac{\eta_A \eta_B - \eta_C^2}{\eta_B^2 - (\eta_A \eta_C^2 / \eta_B)}. \quad (6.42)$$

(This result can be derived from equations (3.4) and (3.5). If we replace, in these equations, suffixes 1, 2, m and 0 by A , B , C and M respectively, then the condition for $\delta_A = \delta_B$ is, from equation (3.5): $\eta_A \eta_B = \eta_C \eta_M$ and equation (6.42) then follows immediately from equation (3.4).)

Two solutions given by equation (6.42), δ'_A and $(\pi - \delta'_A)$, are possible. We can specify that δ'_A corresponds to λ_1 and $(\pi - \delta'_A)$ to λ_2 where λ_1 and λ_2 are the two wavelengths where suppression is to be obtained. Solving these two equations for δ'_A gives

$$\delta'_A = \frac{\pi}{1 + (\lambda_1 / \lambda_2)} \quad (6.43)$$

which can be entered in equation (6.42), whence

$$\tan^2 \frac{\pi}{1 + (\lambda_1 / \lambda_2)} = \frac{\eta_A \eta_B - \eta_C^2}{\eta_B^2 - (\eta_A \eta_C^2 / \eta_B)}. \quad (6.44)$$

This determines the complete design of the coating. The optical thickness of the layer A can be found from equation (6.43) to be

$$\frac{\lambda_1 \lambda_2}{2(\lambda_1 + \lambda_2)}. \quad (6.45)$$

The only other quantity to be found is the optical thickness of layer C and we note first that the total optical thickness of the period is $\lambda_0/2$, where λ_0 is the wavelength of the first high-reflectance zone. The optical thicknesses of layers A and B have already been defined as equal, so that the optical thickness of layer C is

$$\frac{\lambda_0}{2} - \frac{2\lambda_1 \lambda_2}{2(\lambda_1 + \lambda_2)}. \quad (6.46)$$

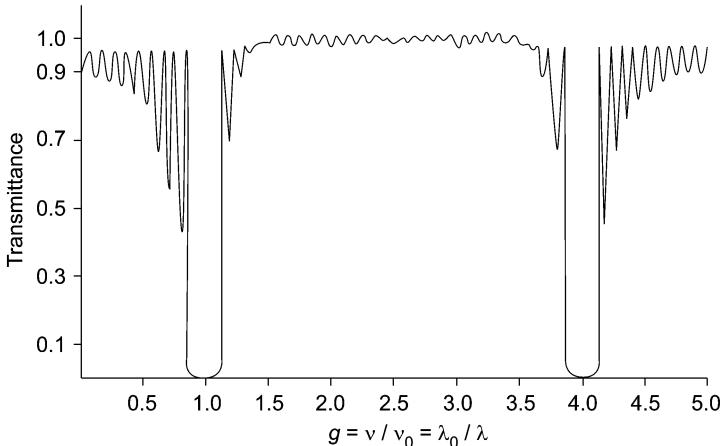


Figure 6.23. Calculated transmittance as a function of g of the design:

$$M|(ABCBA)^{10}A|S$$

with $n_S = 1.50$, $n_M = 1.00$, $n_A = 1.38$, $n_B = 1.90$ and $n_C = 2.30$. (After Thelen [15].)

This medium M which was introduced as an artificial aid to calculation, disappears and does not figure at all in the results. Any two of the optical admittances η_A , η_B and η_C can be chosen at will. The third one is then found from equation (6.44).

Thelen in his paper, gives a large number of examples of multilayers with various zones suppressed. Particularly useful is a multilayer with the second- and third-order zones suppressed. For this,

$$\lambda_1 = \lambda_0/2 \quad \lambda_2 = \lambda_0/3$$

and all the layers are found to be of equal optical thickness $\lambda_0/10$. Two of the refractive indices of the layers are then chosen and equation (6.44) solved for the remaining one. For rapid calculation Thelen gives a nomogram connecting the three quantities. The transmittance of a multilayer with the second and third orders suppressed is given in figure 6.23.

Thelen also considered a multilayer in which the second, third and fourth orders were all suppressed and found the conditions to be as follows.

Layer thicknesses:

$$A : \lambda_0/12$$

$$B : \lambda_0/12$$

$$C : \lambda_0/6.$$

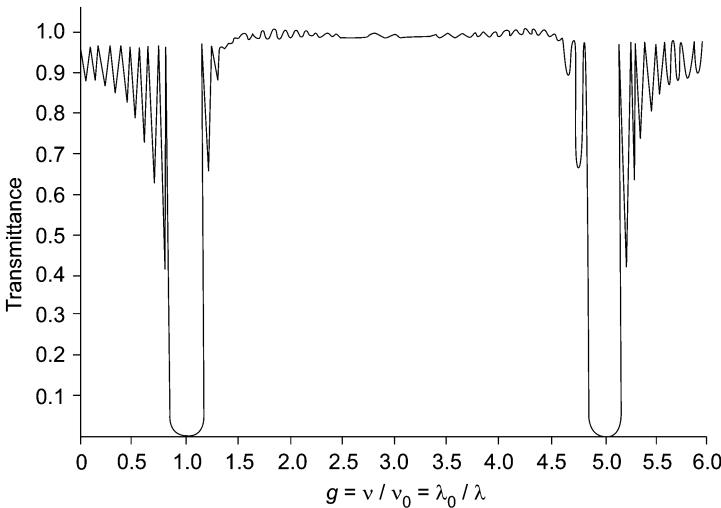


Figure 6.24. Calculated transmittance as a function of g of the design:

$$M|(AB2CBA)^{10}A|S$$

with $n_S = 1.50$, $n_M = 1.00$, $n_A = 1.38$, $n_B = 1.781$ and $n_C = 2.30$. (After Thelen [15].)

The indices are given by

$$\eta_B = (\eta_A \eta_C)^{1/2}.$$

Figure 6.24 shows the transmittance of a multilayer where the second, third and fourth orders have been suppressed in this way.

A heat-reflecting filter using a combination of stacks in which the second and third, and second, third and fourth orders have been suppressed, together with the normal quarter-wave stacks, has been designed. The calculated transmittance spectrum is shown in figure 6.25. The production of such a coating would indeed be a formidable task.

6.2.3.10 Reducing the transmission zone

The simple quarter-wave multilayer has the even-order high-reflectance bands missing. Sometimes it is useful to have these high-reflectance bands present. The method of the previous section can also be applied to this problem and the enhancement of the reflectance at the even orders is a relatively simple business.

Because it makes the analysis simpler, we assume that the basic period is of the form AB rather than $(A/2)B(A/2)$. Once the basic result is established, it

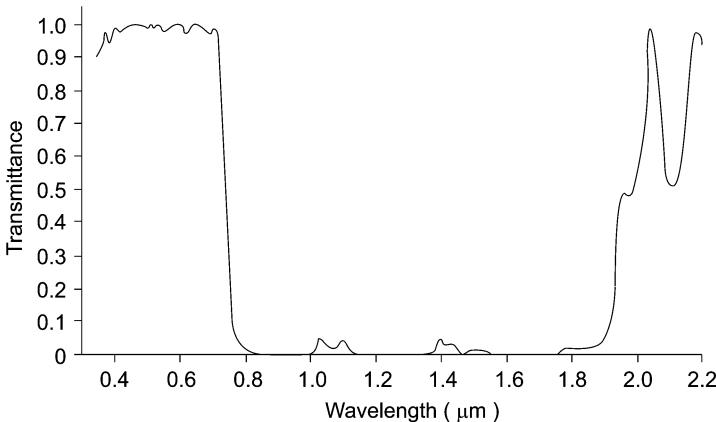


Figure 6.25. Calculated transmittance of a triple-stack heat reflector. Design:

$$M \left| [1.1(\frac{1}{2}AC\frac{1}{2}A)](\frac{1}{2}AC\frac{1}{2}A)^5[1.25(\frac{1}{2}AC\frac{1}{2}A)] - [0.57(ADCDA)]^8[0.642(AB2CBA)]^8 \frac{1}{2}A \right| S$$

with $\lambda_0 = 860$ nm, $n_S = 1.50$, $n_M = 1.00$, $n_A = 1.38$, $n_B = 1.781$, $n_C = 2.30$ and $n_D = 1.90$. (After Thelen [15].)

can easily be converted to the form $(A/2)B(A/2)$ if required. The reason that the even-order peaks are suppressed in the ordinary quarter-wave stack is that each of the layers is an integral number of half-waves thick and so $|t| = 1$ for the basic period. All that is required for a reflectance peak to appear is the destruction of this condition. To achieve this, the thickness of one of the layers must be increased and the other decreased, keeping the overall optical thickness constant. The greater the departure from the half-wave condition, the more pronounced the reflectance peak.

Consider the case where reflectance bands are required at λ_0 , $\lambda_0/2$, and $\lambda_0/3$, but not necessarily at $\lambda_0/4$. This will be satisfied by making $n_A d_A = n_B d_B / 3$ and $n_A d_A = \lambda_0/8$ so that the basic stack becomes either

$$\frac{H}{2} \frac{3L}{2} \frac{H}{2} \frac{3L}{2} \cdots \frac{3L}{2}$$

or

$$\frac{L}{2} \frac{3H}{2} \frac{L}{2} \frac{3H}{2} \cdots \frac{3H}{2}.$$

The reflectance peak at $\lambda_0/4$ will be suppressed because the layers at that wavelength have integral half-wave thicknesses.

The method can be used to produce any number of high-reflectance zones. However, it should be noted that the further the thicknesses depart from ideal quarter-waves at λ_0 , the narrower will be the first-order reflectance band.

6.2.3.11 Edge steepness

In long- and shortwave-pass filters, the steepness of edge is not usually a parameter of critical importance. The number of layers necessary to produce the required rejection in the stop band of the filter will generally produce an edge steepness which is quite acceptable.

If, however, an exceptional degree of edge steepness is required, then the easiest way of improving it is to use still more layers. Increasing the number of layers will cause an apparent increase in the ripple in the pass band, because the first minimum in the pass band will be brought nearer to the edge, and usually will be on a part of the reflectance envelope which is increasing in width towards the edge. If the increase in number of layers is considerable, then it will probably be advisable to use one of the more advanced techniques for reducing ripple.

An alternative method for increasing the steepness of edge without major alterations to the basic design concept is the use of higher-order stacks. The steepness of edge for a given number of layers will increase in proportion with the order. There are two snags here. The first is that the rejection zone width varies inversely with the order number. This can be dealt with by adding a further first-order stack to extend the rejection zone. The second snag is more serious. The permissible errors in layer thickness are also reduced in inverse proportion with the order number. This is because the performance does not depend directly on the phase thickness of the layers but rather on the sines and cosines of the layer thicknesses, and in the case of the fifth order, for example, these are layer thicknesses greater than 2π . Thus, while for a first-order edge filter, as we shall see in chapter 9, the random errors in layer thickness which can be tolerated are of the order of 5% or even 10%, those which are tolerable in the fifth order are of the order of 1% or possibly 2%. A possible further practical difficulty with higher-order filters is that considerably more material is required for each layer.

References

- [1] Epstein L I 1952 The design of optical filters *J. Opt. Soc. Am.* **42** 806–10
- [2] Vera J J 1964 Some properties of multilayer films with periodic structure *Opt. Acta* **11** 315–31
- [3] Thelen A 1966 Equivalent layers in multilayer filters *J. Opt. Soc. Am.* **56** 1533–8
- [4] Ufford C and Baumeister P W 1974 Graphical aids in the use of equivalent index in multilayer-filter design *J. Opt. Soc. Am.* **64** 329–34
- [5] Welford W T (writing as W Weinstein) 1954 Computations in thin film optics *Vacuum* **4** 3–19
- [6] Baumeister P W 1958 Design of multilayer filters by successive approximations *J. Opt. Soc. Am.* **48** 955–8
- [7] Liddell H M 1981 *Computer-Aided Techniques for the Design of Multilayer Filters* (Bristol: Adam Hilger)
- [8] Dobrowolski J A and Lowe D 1978 Optical thin film synthesis program based on the use of Fourier transforms *Appl. Opt.* **17** 3039–50

- [9] Jacobsson R 1964 Matching a multilayer stack to a high-refraction-index substrate by means of an inhomogeneous layer *J. Opt. Soc. Am.* **54** 422–3
- [10] Young L and Cristal E G 1966 On a dielectric fiber by Baumeister *Appl. Opt.* **5** 77–80
- [11] Seeley J S, Liddell H M and Chen T C 1973 Extraction of Tschebysheff design data for the lowpass dielectric multilayer *Opt. Acta.* **20** 641–61
- [12] Seeley J S and Smith S D 1966 High performance blocking filters for the region 1 to 20 microns *Appl. Opt.* **5** 81–5
- [13] Turner A F and Baumeister P W 1966 Multilayer mirrors with high reflectance over an extended spectral region *Appl. Opt.* **5** 69–76
- [14] Epstein L I 1955 Improvements in heat reflecting filters *J. Opt. Soc. Am.* **45** 1360–2
- [15] Thelen A 1963 Multilayer filters with wide transmittance bands *J. Opt. Soc. Am.* **53** 1266–70
- [16] Muchmore R B 1948 Optimum band width for two layer and anti-reflection films *J. Opt. Soc. Am.* **38** 20–6

Chapter 7

Band-pass filters

A filter which possesses a region of transmission bounded on either side by regions of rejection is known as a band-pass filter. For the broadest band-pass filters, the most suitable construction is a combination of longwave-pass and shortwave-pass filters, which we discussed in chapter 6. For narrower filters, however, this method is not very successful because of difficulties associated with obtaining both the required precision in positioning and the steepness of edges. Other methods are therefore used, involving a single assembly of thin films to produce simultaneously the pass and rejection bands. The simplest of these is the thin-film Fabry–Perot filter, a development of the interferometer already described in chapter 5. The thin-film Fabry–Perot filter has a pass band shape which is triangular and it has been found possible to improve this by coupling simple filters in series in much the same way as tuned circuits. These coupled arrangements are known as multiple cavity filters or multiple half-wave filters. If two simple Fabry–Perot filters are combined, the resultant becomes a double cavity or double half-wave filter, abbreviated to DHW filter, while, if three Fabry–Perot filters are involved, we have a triple cavity filter, abbreviated normally to THW for triple half-wave. In the earlier part of this chapter, we consider single cavity filters. First of all, we examine combinations of edge filters.

7.1 Broadband-pass filters

Band-pass filters can be very roughly divided into broadband-pass filters and narrowband-pass filters. There is no definite boundary between the two types and the description of one particular filter usually depends on the application and the filters with which it is being compared. For the purpose of the present work, by broadband filters we mean filters with bandwidths of perhaps 20% or more which are made by combining longwave-pass and shortwave-pass filters. The best arrangement is probably to deposit the two components on opposite sides of a single substrate. To give maximum possible transmission, each edge filter should

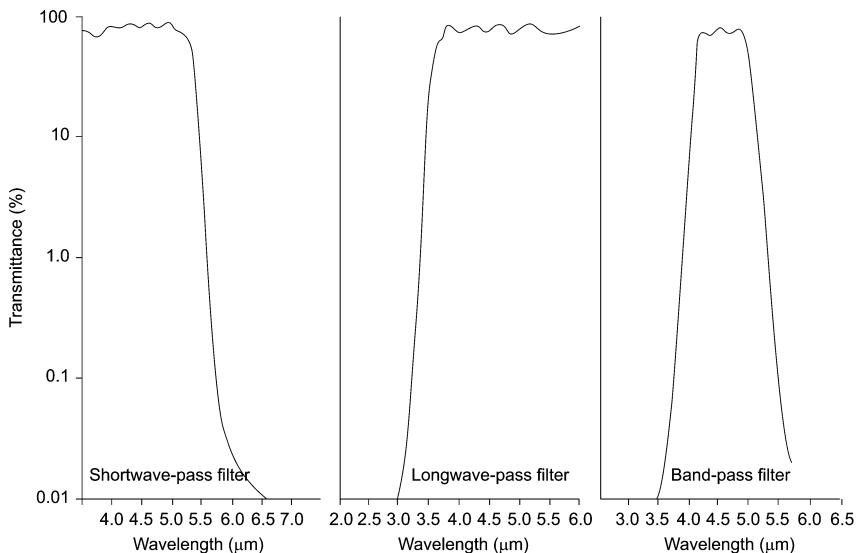


Figure 7.1. The construction of a band-pass filter by placing two separate edge filters in series. (Courtesy of Standard Telephones and Cables Ltd.)

be designed to match the substrate into the surrounding medium, a procedure already examined in chapter 6. Such a filter is shown in figure 7.1.

It is also possible, however, to deposit both components on the same side of the substrate. This was a problem which Epstein [1] examined in his early paper on symmetrical periods. The main difficulty is the combining of the two stacks so that the transmission in the pass band is a maximum and also so that one stack does not produce transmission peaks in the rejection zone of the other. The transmission in the pass band will depend on the matching of the first stack to the substrate, the matching of the second stack to the first, and the matching of the second stack to the surrounding medium. Depending on the equivalent admittances of the various stacks it may be necessary to insert quarter-wave matching layers or to adopt any of the more involved matching techniques.

In the visible region, with materials such as zinc sulphide and cryolite, the combination $[(H/2)L(H/2)]^S$ acts as a good longwave-pass filter with an equivalent admittance at normal incidence and at wavelengths in the pass region not too far removed from the edge of near unity. This can therefore be used next to the air without mismatch. The combination $[(L/2)H(L/2)]^S$ acts as a shortwave-pass filter, with equivalent admittance only a little lower than the first section, and can be placed next to it, between it and the substrate, without any matching layers. The mismatch between this second section and the substrate, which in the visible region will be glass of index 1.52, is sufficiently large to require a matching layer. Happily, the $[(H/2)L(H/2)]$ combination with a total phase

Table 7.1.[†]

Layer	Phase thickness of each layer measured at 546 nm (degrees)	Index	Phase thickness of each layer measured at) 546 nm (degrees)
1.52	Massive	1.38	55.4
1.38	67.3	2.30	33.9
2.30	134.5	1.38	67.9
1.38	122.7	2.30	67.9
2.30	110.8	1.38	67.9
1.38	110.8	2.30	67.9
2.30	110.8	1.38	67.9
1.38	110.8	2.30	67.9
2.30	110.8	1.38	67.9
1.38	110.8	2.30	33.9
2.30	110.8	1.00	Massive

[†] From Epstein [1].

thickness of 270°, i.e. effectively three quarter-waves, has an admittance exactly correct for this. The transmission of the final design is shown in figure 7.2(b) with the appropriate admittances of the two sections in figure 7.2(a). Curve A refers to a $[(L/2)H(L/2)]^4$ shortwave-pass section and B to a $[(H/2)L(H/2)]^4$ longwave-pass. The complete design is shown in table 7.1. The edges of the two sections have been chosen quite arbitrarily and could be moved as required.

To avoid the appearance of transmission peaks in the rejection zones of either component, it is safest to deposit them so that high-reflectance zones do not overlap. The complete rejection band of the shortwave-pass section will always lie over a pass region of the longwave-pass filter, but the higher-order bands should be positioned, if at all possible, clear of the rejection zone of the longwave-pass section. The combination of edge filters of the same type has already been investigated in chapter 6 and the principles discussed there apply to this present situation. It should also be remembered that, although in the normal shortwave-pass filter the second-order reflection peak is missing, a small peak can appear if any thickness errors are present. This can, if superimposed on a rejection zone of the other section, cause the appearance of a transmission peak if the errors are sufficiently pronounced. The expression for maximum transmission is

$$T_{\max} = \frac{T_a T_b}{[1 - (R_a R_b)^{1/2}]^2}$$

but this only holds if the phase conditions are met.

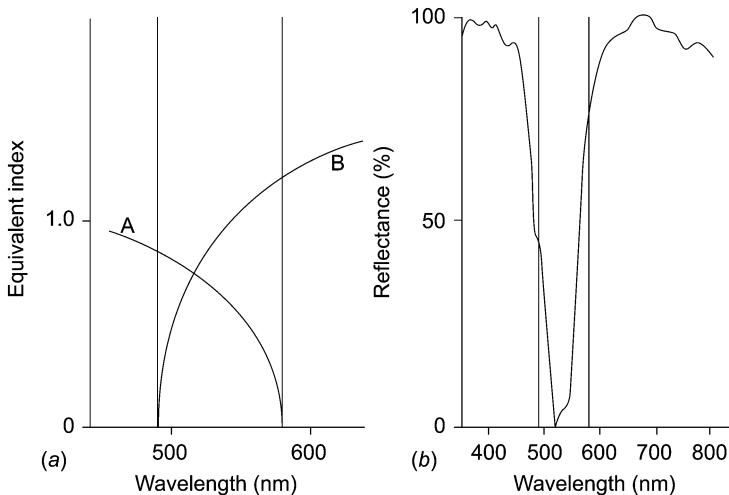


Figure 7.2. (a) Equivalent admittances of two stacks made up of symmetrical periods used to form a band-pass filter. A: $(0.5LH0.5L)$; B: $(0.5HL0.5H)$, where $n_L = 1.38$, $n_H = 2.30$. (b) Calculated reflectance curve for a band-pass filter. For the complete design of this filter, made up of two superimposed stacks, one of type A and one of type B, refer to table 7.1. (After Epstein [1].)

7.2 Narrowband filters

7.2.1 The metal–dielectric Fabry–Perot filter

The simplest type of narrowband thin-film filter is based on the Fabry–Perot interferometer discussed in chapter 5. In its original form, the Fabry–Perot interferometer consists of two identical parallel reflecting surfaces spaced apart a distance d . In collimated light, the transmission is low for all wavelengths except for a series of very narrow transmission bands spaced at intervals that are constant in terms of wavenumber. This device can be replaced by a complete thin-film assembly consisting of a dielectric layer bounded by two metallic reflecting layers (figure 7.3). The dielectric layer takes the place of the spacer and is known as the spacer layer. Except that the spacer layer now has an index greater than unity, the analysis of the performance of this thin-film filter is exactly the same as for the conventional etalon, but in other respects there are a few significant differences.

While the surfaces of the substrates should have a high degree of polish, they need not be worked to the exacting tolerances necessary for etalon plates. Provided the vapour stream in the plant is uniform, the films will follow the contours of the substrate without exhibiting thickness variations. This implies that it is possible for the thin-film Fabry–Perot filter to be used in a much lower

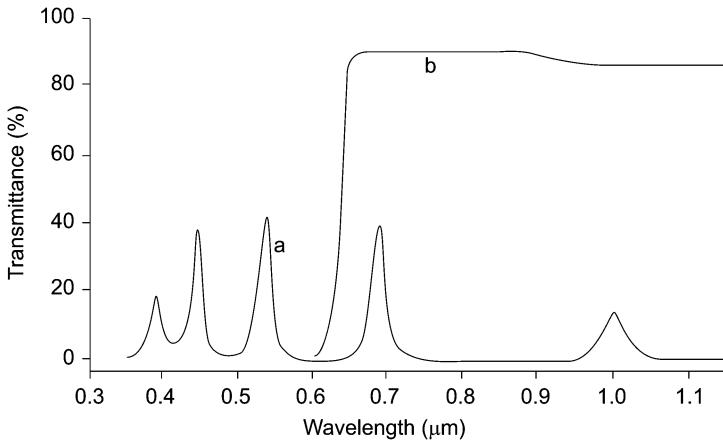


Figure 7.3. Characteristics of a metal–dielectric filter for the visible region (curve a). Curve b is the transmittance of an absorption glass filter that can be used for the suppression of the short wavelength sidebands. (Courtesy of Barr & Stroud Ltd.)

order than the conventional etalon. Indeed, it turns out in practice that lower orders must be used, because the thin-film spacer layers begin, where thicker than the fourth order or so, to exhibit roughness. This roughness broadens the pass band and reduces the peak transmittance so much that any advantage of the higher order is completely lost. This simple type of filter is known as a metal–dielectric Fabry–Perot to distinguish it from the all-dielectric one to be described later.

It is worthwhile briefly analysing the performance of the Fabry–Perot once again, this time including the effects of phase shift at the reflectors. The starting point for this analysis is equation (2.150):

$$\left. \begin{aligned} T_F &= \frac{T_a T_b}{[1 - (R_a R_b)^{1/2}]^2} \frac{1}{1 + F \sin^2[\frac{1}{2}(\phi_a + \phi_b) - \delta]} \\ F &= \frac{4(R_a R_b)^{1/2}}{[1 - (R_a R_b)^{1/2}]^2} \quad \delta = \frac{2\pi n d \cos \theta}{\lambda} \end{aligned} \right\} \quad (7.1)$$

where the notation is given in figure 2.19. We have adapted equation (2.150) slightly by removing the + and – signs on the reflectances. The analysis which follows is similar to that already performed in chapter 5 except that here we are including the effects of ϕ_a and ϕ_b . The maxima of transmission are given by

$$\frac{2\pi n d \cos \theta}{\lambda} - \frac{\phi_a + \phi_b}{2} = m\pi \quad m = 0, \pm 1, \pm 2, \pm 3, \dots \quad (7.2)$$

where we have chosen $-m$ rather than $+m$ because $(\phi_a + \phi_b)/2 < \pi$ by definition. The analysis is marginally simpler if we work in terms of wavenumber instead of

wavelength. The positions of the peaks are then given by

$$\frac{1}{\lambda} = v = \frac{m\pi + (\phi_a + \phi_b)/2}{2\pi nd \cos \theta} = \frac{1}{2nd \cos \theta} \left(m + \frac{\phi_a + \phi_b}{2\pi} \right). \quad (7.3)$$

Depending on the particular metal, the thickness, the index of the substrate and the index of the spacer, the phase shift on reflection ϕ will be either in the first or second quadrant. $(\phi_a + \phi_b)/(2\pi)$ will therefore be positive between 0 and 1 and roughly in the region of 0.5. The peak wavelength of the filter will therefore be shifted to the shortwave side of the peak which would be expected simply from the optical thickness of the spacer layer.

The resolving power of the thin-film Fabry–Perot filter may be defined in exactly the same way as for the interferometer. As we saw in chapter 5, a convenient definition is

$$\frac{\text{Peak wavelength}}{\text{Halfwidth of pass band}}$$

where the halfwidth is the width of the band measured at half the peak transmission. Now let the pass bands be sufficiently narrow, which is the same as F being sufficiently large, so that near a peak we can replace

$$\frac{\phi_a + \phi_b}{2} - \delta \quad \text{by} \quad -m\pi - \Delta\delta$$

and

$$\sin^2 \left(\frac{\phi_a + \phi_b}{2} - \delta \right) \quad \text{by} \quad (\Delta\delta)^2.$$

We are assuming here that ϕ_a and ϕ_b are constant or vary very much more slowly than δ over the pass band.

The half-peak bandwidth, or halfwidth, can be found by noting that at the half-peak transmission points

$$F \sin^2 \left(\frac{\phi_a + \phi_b}{2} - \delta \right) = 1.$$

Using the approximation given above, this becomes

$$(\Delta\delta_h)^2 = \frac{1}{F}$$

i.e. the halfwidth of the pass band

$$2\Delta\delta_h = 2/F^{1/2}.$$

The finesse is defined as the ratio of the interval between fringes to the fringe halfwidth, and is written \mathcal{F} . The change in δ in moving from one fringe to the next is just π , and the finesse, therefore, is

$$\mathcal{F} = \frac{\pi F^{1/2}}{2}. \quad (7.4)$$

Now $v_0/\Delta v_h = \delta_0/2\Delta\delta_h$ because $v \propto \delta$, where v_0 and δ_0 are respectively the values of the wavenumber and spacer layer phase thickness associated with the transmission peak, and Δv_h and $2\Delta\delta_h$ are the corresponding values of halfwidth. The ratio of the peak wavenumber to the halfwidth is then given by

$$\frac{v_0}{\Delta v_h} = \mathcal{F} \left(m + \frac{\phi_a + \phi_b}{2\pi} \right) \quad (7.5)$$

for a peak of order m , since

$$\delta_0 = m\pi + \frac{\phi_a + \phi_b}{2}.$$

The ratio of peak position to halfwidth expressed in terms of wavenumber is exactly the same in terms of wavelength,

$$\frac{v_0}{\Delta v_h} = \frac{\lambda_0}{\Delta\lambda_h} \quad (7.6)$$

where λ_0 is given by

$$\lambda_0 = \frac{2nd \cos \theta}{m + (\phi_a + \phi_b)/2\pi} \quad (7.7)$$

and this was discussed in chapter 5. The halfwidth is thus a most useful parameter with which to specify a narrowband Fabry–Perot filter since it can be converted very quickly into a measure of resolution. It has come to be used rather than resolving power for all types of narrowband filter, regardless of whether or not they are Fabry–Perot type. Usually, therefore $\Delta\lambda_h/\lambda_0$, often expressed as a percentage, is the parameter which is quoted by the manufacturers and users alike. Other measures of bandwidth sometimes quoted along with the halfwidth are the widths measured at $0.9 \times$ peak transmission, at $0.1 \times$ peak transmission, and at $0.01 \times$ peak transmission. For a Fabry–Perot filter, provided the phase shifts on reflection from the reflecting layers are effectively constant over the pass band, these widths are given respectively by one-third of the halfwidth, three times the halfwidth, and ten times the halfwidth. The other measures of bandwidth are used to give some indication of the extent to which, in any given type of filter, the sides of the pass band, compared with those of the Fabry–Perot, can be considered rectangular.

The manufacture of the metal–dielectric filter is straightforward. The main point to watch is that the metallic layers should be evaporated as quickly as

possible on to a cold substrate. In the visible and near infrared regions the best results are probably achieved with silver and cryolite, while in the ultraviolet the best combination is aluminium and either magnesium fluoride or cryolite. Wherever possible the layers should be protected by cementing a cover slip over them as soon as possible after deposition. This also serves to balance the assembly by equalising the refractive indices of the media outside the metal layers.

Turner [2] quoted some results for metal-dielectric filters constructed for the visible region which may be taken as typical of the performance to be expected. The filters were constructed from silver reflectors and magnesium fluoride spacers. For a first-order spacer a bandwidth of 13 nm with a peak transmission of 30% was obtained at a peak wavelength of 531 nm. A similar filter with a second-order spacer gave a bandwidth of 7 nm with peak transmission of 26% at 535 nm. With metal-dielectric filters the third order is usually the highest used. Because of scattering in the space layer, which becomes increasingly apparent in the fourth and higher orders, any benefit which would otherwise arise from using these orders is largely lost.

A typical curve for a metal-dielectric filter for the visible region is shown in figure 7.3. The particular peak to be used is that at $0.69 \mu\text{m}$, which is of the third order. The shortwave sidebands due to the higher-order peaks can be suppressed quite easily by the addition of an absorption glass filter, which can be cemented over the metal-dielectric element to act as a cover glass. Such a filter is also shown in the figure and is one of a wide range of absorption glasses which are available for the visible and near infrared and which have longwave-pass characteristics. There are, unfortunately, few absorption filters suitable for the suppression of the longwave sidebands. If the detector which is to be used is not sensitive to these longer wavelengths, then no problem exists and commercial metal-dielectric filters for the visible and near infrared usually possess long-wavelength sidebands beyond the limit of the photocathodes or photographic emulsions, which are the usual detectors for this region. If the longwave-sideband suppression must be included as part of the filter assembly, then there is an advantage in using metal-dielectric filters in the first order, even though the peak transmission for a given bandwidth is much lower, since they do not usually possess long-wavelength sidebands. Theoretically, there will always be a peak corresponding to the zero order at very long wavelengths, but this will not usually appear, partly because the substrate will cut off long before the zero order is reached, and also because the properties of the thin-film materials themselves will change radically. We shall discuss later a special type of metal-dielectric filter, the induced transmission filter, which can be made to have a much higher peak transmission, though with a rather broader halfwidth, without introducing long-wavelength sidebands, and which is often used as a long-wavelength suppression filter.

Silver does not have an acceptable performance for ultraviolet filters and aluminium has been found to be the most suitable metal, with magnesium fluoride as the preferred dielectric. In the ultraviolet beyond 300 nm there are few suitable cements (none at all beyond 200 nm) and it is not possible to use cover slips

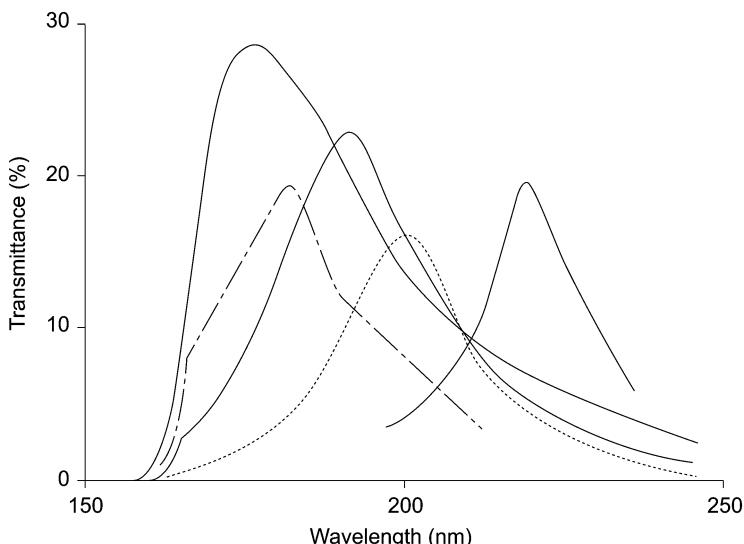


Figure 7.4. Experimental transmittance curves of first-order metal–dielectric filters for the far ultraviolet deposited on Spectrosil B substrates. (After Bates and Bradley [3].)

which are cemented over the layers in the way in which filters for the visible region are protected. The normal technique, therefore, is to attempt to protect the filter by the addition of an extra dielectric layer between the final metal layer and the atmosphere. These layers are effective in that they slow down the oxidation of the aluminium which otherwise takes place rapidly and causes a reduction in performance even at quite low pressures. This oxidation has already been referred to in chapter 4. They cannot completely stabilise the filters, however, and slight longwave drifts can occur, as reported by Bates and Bradley [3]. A second function of the final dielectric layer is to act as a reflection-reducing layer at the outermost metal surface and hence to increase the transmittance of the filter. This is not a major effect—the problem of improving metal–dielectric filter performance is dealt with later in this chapter—but any technique which helps to improve performance, even marginally, in the ultraviolet, is very welcome. Some performance curves of first-order metal–dielectric Fabry–Perot filters are shown in figure 7.4.

The formula for transmission of the Fabry–Perot filter can also be used to determine both the peak transmission in the presence of absorption in the reflectors and the tolerance which can be allowed in matching the two reflectors. First of all, let the reflectances be equal and let the absorption be denoted by A , so that

$$R + T + A = 1. \quad (7.8)$$

The peak transmission will then be given by

$$(T_F)_{\text{peak}} = \frac{T^2}{(1 - R)^2}$$

and, using equation (7.8),

$$(T_F)_{\text{peak}} = \frac{1}{(1 + A/T)^2} \quad (7.9)$$

exactly as for the Fabry–Perot interferometer, which shows that when absorption is present the value of peak transmission is determined by the ratio A/T .

To estimate the accuracy of matching which is required for the two reflectors we assume that the absorption is zero. The peak transmission is given by the expression

$$(T_F)_{\text{peak}} = \frac{T_a T_b}{[1 - (R_a R_b)^{1/2}]^2} \quad (7.10)$$

where the subscripts a and b refer to the two reflectors. Let

$$R_b = R_a - \Delta_a \quad (7.11)$$

where Δ_a is the error in matching, so that $T_b = T_a + \Delta_a$. Then we can write

$$\begin{aligned} (T_F)_{\text{peak}} &= \frac{T_a(T_a + \Delta_a)}{\{1 - [R_a(R_a - \Delta_a)]^{1/2}\}^2} \\ &= \frac{T_a(T_a + \Delta_a)}{\{1 - R_a[1 - \frac{1}{2}(\Delta_a/R_a) + \dots]\}^2}. \end{aligned} \quad (7.12)$$

Now assume that Δ_a is sufficiently small compared with R_a so that we can take only the first two terms of the expansion in equation (7.12). With some rearrangement the equation becomes

$$(T_F)_{\text{peak}} = \frac{T_a^2}{(1 - R_a)^2} \frac{1 + (\Delta_a/T_a)}{\{1 + \frac{1}{2}(\Delta_a/T_a)\}^2}. \quad (7.13)$$

The first part of the equation is the expression for peak transmission in the absence of any error in the reflectors, while the second part shows how the peak transmission is affected by errors. The second part of the expression is plotted in figure 7.5 where the abscissa is $T_b/T_a = 1 + \Delta_a/T_a$. Clearly, the Fabry–Perot filter is surprisingly insensitive to errors. Even with reflector transmittance unbalanced by a factor of 3, it is still possible to achieve 75% peak transmission.

7.2.2 The all-dielectric Fabry–Perot filter

In the same way as we found for the conventional Fabry–Perot etalon, if improved performance is to be obtained, then the metallic reflecting layers should be replaced by all-dielectric multilayers.

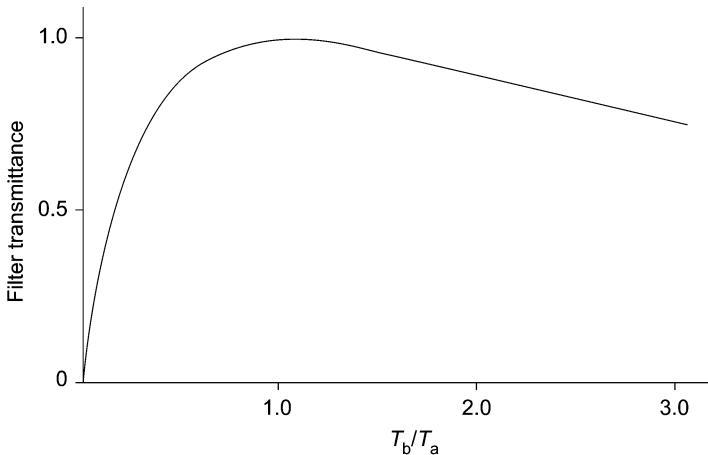


Figure 7.5. Theoretical peak transmittance of a Fabry-Perot filter with unbalanced reflectors.

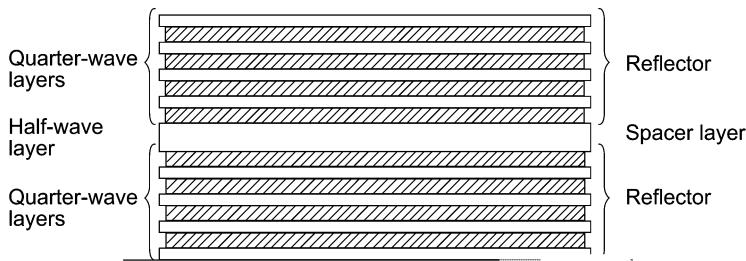


Figure 7.6. The structure of an all-dielectric Fabry-Perot filter.

An all-dielectric filter is shown in diagrammatic form in figure 7.6. Basically, this is the same as the conventional etalon with dielectric coatings and with a solid thin-film spacer, and the observations made for the metal–dielectric filter are also valid. Again, the substrate need not be worked to a high degree of flatness although the polish must be good, because, provided the plant geometry is adequate, the films will follow any contours without showing changes in thickness.

The bandwidth of the all-dielectric filter can be calculated as follows. If the reflectance of each of the multilayers is sufficiently high, then

$$F = \frac{4R}{(1-R)^2} \simeq \frac{4}{T^2}$$

and

$$\frac{\lambda_0}{\Delta\lambda_h} = m\mathcal{F} = \frac{m\pi F^{1/2}}{2} \simeq \frac{m\pi}{T}. \quad (7.14)$$

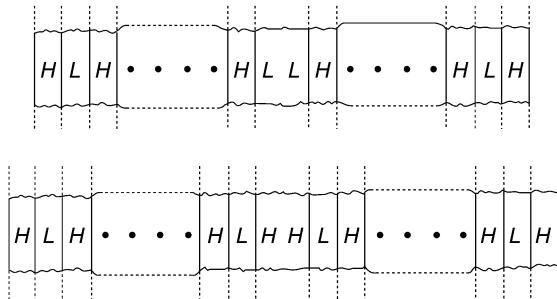


Figure 7.7. The structure of the two basic types of all-dielectric Fabry–Perot filter.

Since the maximum reflectance for a given number of layers will be obtained with a high-index layer outermost, there are really only two cases which need be considered and these are shown in figure 7.7. If x is the number of high-index layers in each stack, not counting the spacer layer, then in the case of the high-index spacer, the transmission of the stack will be given by

$$T = \frac{4n_L^{2x} \cdot n_s}{n_H^{2x+1}}$$

and in the case of the low-index spacer by

$$T = \frac{4n_L^{2x-1} n_s}{n_H^{2x}}.$$

Substituting these results into the expression for bandwidth we find, for the high-index spacer,

$$\frac{\Delta\lambda_h}{\lambda_0} = \frac{4n_L^{2x} n_s}{m\pi n_H^{2x+1}} \quad (7.15)$$

and, for the low-index spacer,

$$\frac{\Delta\lambda_h}{\lambda_0} = \frac{4n_L^{2x-1} n_s}{m\pi n_H^{2x}} \quad (7.16)$$

where we are adopting the fractional halfwidth $\Delta\lambda_h/\lambda_0$ rather than the resolving power $\lambda_0/\Delta\lambda_h$ as the important parameter. This is customary practice.

In these formulae we have completely neglected any effect due to the dispersion of phase change on reflection from a multilayer. As we have already noted in chapter 5, the phase change is not constant. The sense of the variation is such that it increases the rate of variation of $[(\phi_a + \phi_b)/2] - \delta$ with wavelength in the formula for transmission of the Fabry–Perot filter and, hence, reduces the bandwidth and increases the resolving power in equations (7.15) and (7.16).

Seeley [4] has studied the all-dielectric filter in detail and, by making some approximations in the basic expressions for the filter transmittance, has arrived at formulae for the first-order halfwidths, which, with a little adjustment, become equal to the expressions in (7.15) and (7.16) multiplied by a factor $(n_H - n_L)/n_H$. We can readily extend Seeley's analysis to all-dielectric filters of order m .

We recall that the half-peak points are given by

$$F \sin^2[(2\pi D/\lambda) - \phi] = 1 \quad (7.17)$$

where, since the filter is quite symmetrical, we have replaced $(\phi_1 + \phi_2)/2$ by ϕ . It is simpler to carry out the analysis in terms of $g = \lambda_0/\lambda = v/v_o$. At the peak of the filter we have $g = 1.0$. We can assume for small changes Δg in g that

$$2\pi D/\lambda = m\pi(1 + \Delta g)$$

and

$$\phi = \phi_0 + \frac{d\phi}{dg} \Delta g$$

so that equation (7.17) becomes

$$F \sin^2 \left(m\pi(1 + \Delta g) - \phi_0 - \frac{d\phi}{dg} \Delta g \right) = 1.$$

ϕ_0 , we know, is 0 or π , and so, using the same approximation as before,

$$F \left(m\pi \Delta g - \frac{d\phi}{dg} \Delta g \right)^2 = 1$$

or

$$\Delta g = F^{-1/2} \left(m\pi - \frac{d\phi}{dg} \right)^{-1}.$$

The halfwidth is $2\Delta g$ so that

$$\begin{aligned} 2\Delta g &= \frac{\Delta v_h}{v_0} = \frac{\Delta \lambda_h}{\lambda_0} = 2F^{-1/2} \left(m\pi - \frac{d\phi}{dg} \right)^{-1} \\ &= \frac{2}{m\pi F^{1/2}} \left(1 - \frac{1}{m\pi} \frac{d\phi}{dg} \right)^{-1}. \end{aligned} \quad (7.18)$$

We now need the quantity $d\phi/dg$. We use Seeley's technique, but, rather than follow him exactly, we choose a slightly more general approach because we shall require the results later. The matrix for a dielectric quarter-wave layer is

$$\begin{bmatrix} \cos \delta & (i \sin \delta)/n \\ i n \sin \delta & \cos \delta \end{bmatrix}$$

where, as usual, we are writing n for the optical admittance, which is in free space units. Now, for layers which are almost a quarter-wave we can write

$$\delta = \pi/2 + \varepsilon$$

where ε is small. Then

$$\cos \delta \simeq -\varepsilon \quad \sin \delta \simeq 1$$

so that the matrix can be written

$$\begin{bmatrix} -\varepsilon & i/n \\ in & -\varepsilon \end{bmatrix}.$$

We limit our analysis to quarter-wave multilayer stacks having high index next to the substrate. There are two cases, even and odd numbers of layers.

7.2.2.1 Case 1: even number ($2x$) of layers

The resultant multilayer matrix is given by

$$\begin{bmatrix} B \\ C \end{bmatrix} = [L][H][L] \dots [L][H] \begin{bmatrix} 1 \\ n_m \end{bmatrix}$$

where

$$\begin{aligned} [L] &= \begin{bmatrix} -\varepsilon_L & i/n_L \\ in_L & -\varepsilon_L \end{bmatrix} \\ [H] &= \begin{bmatrix} -\varepsilon_H & i/n_H \\ in_H & -\varepsilon_H \end{bmatrix}. \end{aligned}$$

Then

$$\begin{aligned} \begin{bmatrix} B \\ C \end{bmatrix} &= \{[L][H]\}^x \begin{bmatrix} 1 \\ n_m \end{bmatrix} \\ &= \begin{bmatrix} -(\frac{n_H}{n_L}) & -i(\frac{\varepsilon_L}{n_H} + \frac{\varepsilon_H}{n_L}) \\ -i(n_L\varepsilon_H + n_H\varepsilon_L) & -(\frac{n_L}{n_H}) \end{bmatrix}^x \begin{bmatrix} 1 \\ n_m \end{bmatrix} \\ &= \begin{bmatrix} M_{11} & iM_{12} \\ iM_{21} & M_{22} \end{bmatrix} \begin{bmatrix} 1 \\ n_m \end{bmatrix}. \end{aligned}$$

Our problem is to find expressions for M_{11} , M_{12} , M_{21} and M_{22} . In the evaluation we neglect all terms of second and higher order in ε . Terms in ε appearing in M_{11} and M_{22} are of second and higher order and therefore

$$\begin{aligned} M_{11} &= (-1)^x \left(\frac{n_H}{n_L} \right)^x \\ M_{22} &= (-1)^x \left(\frac{n_L}{n_H} \right)^x. \end{aligned}$$

M_{12} and M_{21} contain terms of first, third and higher orders in ε . The first-order terms are

$$\begin{aligned}
 M_{12} = & -\left(\frac{\varepsilon_L}{n_H} + \frac{\varepsilon_H}{n_L}\right) \left(-\frac{n_L}{n_H}\right)^{x-1} \\
 & + \left(-\frac{n_H}{n_L}\right) \left[-\left(\frac{\varepsilon_L}{n_H} + \frac{\varepsilon_H}{n_L}\right)\right] \left(-\frac{n_L}{n_H}\right)^{x-2} + \dots \\
 & + \left(-\frac{n_H}{n_L}\right)^p \left[-\left(\frac{\varepsilon_L}{n_H} + \frac{\varepsilon_H}{n_L}\right)\right] \left(-\frac{n_L}{n_H}\right)^{x-p-1} + \dots \\
 & + \left(-\frac{n_H}{n_L}\right)^{x-1} \left[-\left(\frac{\varepsilon_L}{n_H} + \frac{\varepsilon_H}{n_L}\right)\right] \\
 = & (-1)^x \left(\frac{\varepsilon_L}{n_H} + \frac{\varepsilon_H}{n_L}\right) \left[\left(\frac{n_L}{n_H}\right)^{x-1} + \left(\frac{n_L}{n_H}\right)^{x-3} + \dots + \left(\frac{n_H}{n_L}\right)^{x-1}\right] \\
 = & (-1)^x \left(\frac{\varepsilon_L}{n_H} + \frac{\varepsilon_H}{n_L}\right) \left(\frac{n_H}{n_L}\right)^{x-1} \\
 & \times \left[\left(\frac{n_L}{n_H}\right)^{2x-2} + \left(\frac{n_L}{n_H}\right)^{2x-4} + \dots + \left(\frac{n_L}{n_H}\right)^2 + 1\right] \\
 = & (-1)^x \left(\frac{\varepsilon_L}{n_H} + \frac{\varepsilon_H}{n_L}\right) \left(\frac{n_H}{n_L}\right)^{x-1} \left[1 - \left(\frac{n_L}{n_H}\right)^{2x}\right] \left[1 - \left(\frac{n_L}{n_H}\right)^2\right]^{-1}
 \end{aligned}$$

since $(n_L/n_H) < 1$.

Now, provided x is large enough and (n_L/n_H) small enough, we can neglect $(n_L/n_H)^{2x}$ in comparison with 1, and after some adjustment, the expression becomes

$$M_{12} = \frac{(-1)^x n_H n_L (n_H/n_L)^x (\varepsilon_L/n_H + \varepsilon_H/n_L)}{(n_H^2 - n_L^2)}.$$

A similar procedure yields

$$M_{21} = \frac{(-1)^x n_H n_L (n_H/n_L)^x (n_L \varepsilon_H + n_H \varepsilon_L)}{(n_H^2 - n_L^2)}.$$

7.2.2.2 Case II: odd number $(2x + 1)$ of layers

The resultant matrix is given by

$$\begin{aligned}
 \begin{bmatrix} B \\ C \end{bmatrix} &= [H][L][H]\dots[L][H] \begin{bmatrix} 1 \\ n_m \end{bmatrix} \\
 &= [H]\{[L][H]\}^x \begin{bmatrix} 1 \\ n_m \end{bmatrix}
 \end{aligned}$$

which we can denote by

$$\begin{bmatrix} N_{11} & iN_{12} \\ iN_{21} & N_{22} \end{bmatrix} \begin{bmatrix} 1 \\ n_m \end{bmatrix}$$

and which is simply the previous result multiplied by

$$\begin{bmatrix} -\varepsilon_H & i/n_H \\ in_H & -\varepsilon_H \end{bmatrix}.$$

Then

$$\begin{aligned} N_{11} &= -\varepsilon_H M_{11} - M_{21}/n_H = (-1)^{x+1} \left(\frac{n_H}{n_L} \right)^x \frac{(\varepsilon_L n_H n_L + \varepsilon_H n_H^2)}{(n_H^2 - n_L^2)} \\ N_{12} &= -\varepsilon_H M_{12} + M_{22}/n_H = (-1)^x \left(\frac{n_L}{n_H} \right)^x \frac{1}{n_H} \\ N_{21} &= n_H M_{11} - \varepsilon_H M_{21} = (-1)^x \left(\frac{n_H}{n_L} \right)^x n_H \\ N_{22} &= -\varepsilon_H M_{22} - n_H M_{12} = (-1)^{x+1} \left(\frac{n_H}{n_L} \right)^x \frac{n_H^2 n_L (\varepsilon_L/n_H + \varepsilon_H/n_L)}{(n_H^2 - n_L^2)} \end{aligned}$$

where terms in $(n_L/n_H)^x$ are neglected in comparison with $(n_H/n_L)^x$.

7.2.2.3 Phase shift: case I

We are now able to compute the phase shift on reflection. We take, initially, the index of the incident medium to be n_0 . Then

$$\begin{aligned} \begin{bmatrix} B \\ C \end{bmatrix} &= \begin{bmatrix} M_{11} & iM_{12} \\ iM_{21} & M_{22} \end{bmatrix} \begin{bmatrix} 1 \\ n_m \end{bmatrix} \\ &= \begin{bmatrix} M_{11} + in_m M_{12} \\ n_m M_{22} + iM_{21} \end{bmatrix} \\ \rho &= \frac{n_0 B - C}{n_0 B + C} = \frac{n_0(M_{11} + in_m M_{12}) - n_m M_{22} - iM_{21}}{n_0(M_{11} + in_m M_{12}) + n_m M_{22} + iM_{21}} \\ &= \frac{(n_0 M_{11} - n_m M_{22}) + i(n_0 n_m M_{12} - M_{21})}{(n_0 M_{11} + n_m M_{22}) + i(n_0 n_m M_{12} + M_{21})} \\ \tan \phi &= \frac{2n_0 n_m^2 M_{12} M_{22} - 2n_0 M_{11} M_{21}}{n_0^2 M_{11}^2 - n_m^2 M_{22}^2 + n_0^2 n_m^2 M_{12}^2 - M_{21}^2}. \end{aligned} \tag{7.19}$$

Inserting the appropriate expressions and once again neglecting terms of second and higher order in ε and terms in $(n_L/n_H)^x$, we obtain for ϕ

$$\tan \phi = \frac{-2n_H n_L (n_L \varepsilon_H + n_H \varepsilon_L)}{n_0 (n_H^2 - n_L^2)} \tag{7.20}$$

(for $LH \dots LH LH | n_m$).

7.2.2.4 Phase shift: case II

ρ is given by an expression similar to (7.19), in which M is replaced by N . Then, following the same procedure as for case I we arrive at

$$\tan \phi = \frac{-2n_0(\varepsilon_L n_L + \varepsilon_H n_H)}{(n_H^2 - n_L^2)} \quad (7.21)$$

(for $H L H \dots L H L H | n_m$).

Equations (7.20) and (7.21) are in a general form which we will make use of later. For our present purposes we can introduce some slight simplification.

$$\delta = \frac{2\pi n d}{\lambda} = 2\pi n d v = 2\pi n d v_0 (v/v_0) = (\pi/2)g$$

so that

$$\varepsilon_H = \varepsilon_L = (\pi/2)g - \pi/2 = (\pi/2)(g - 1).$$

Also, when we consider the construction of the Fabry–Perot filters we see that the incident medium in case I will be a high-index spacer layer and in case II a low-index spacer. Thus, for Fabry–Perot filters,

$$\tan \phi = \frac{-\pi n_L}{(n_H - n_L)}(g - 1)$$

for both case I and case II.

Now, ϕ is nearly π or 0. Then

$$\frac{d\phi}{dg} = \frac{-\pi n_L}{(n_H - n_L)}$$

which is the result obtained by Seeley. This can then be inserted in equation (7.18) to give

$$\frac{\Delta \nu_h}{\nu_0} = \frac{\Delta \lambda_h}{\lambda_0} = \frac{2}{m\pi F^{1/2}} \left(\frac{n_H - n_L}{n_H - n_L + n_L/m} \right).$$

Then the expressions for the halfwidth of all-dielectric Fabry–Perot filters of m th order become

High-index spacer:

$$\left(\frac{\Delta \lambda_h}{\lambda_0} \right)_H = \frac{4n_m n_L^{2x}}{m\pi n_H^{2x+1}} \frac{(n_H - n_L)}{(n_H - n_L + n_L/m)} \quad (7.22)$$

Low-index spacer:

$$\left(\frac{\Delta \lambda_h}{\lambda_0} \right)_L = \frac{4n_m n_L^{2x-1}}{m\pi n_H^{2x}} \frac{(n_H - n_L)}{(n_H - n_L + n_L/m)} \quad (7.23)$$

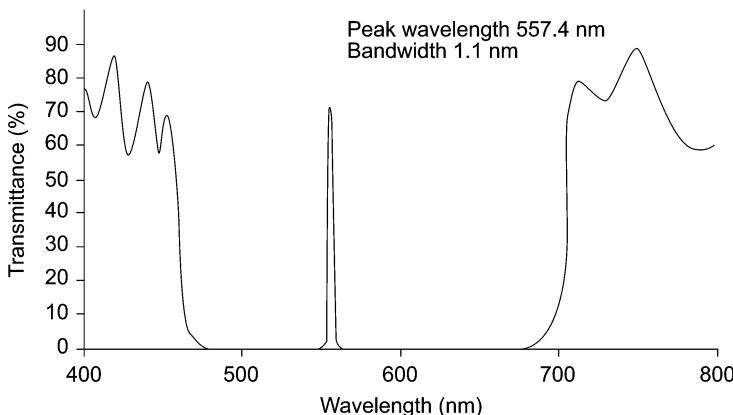


Figure 7.8. Measured transmittance of a narrowband all-dielectric filter with unsuppressed sidebands. Zinc sulphide and cryolite were the thin-film materials used. (Courtesy of Sir Howard Grubb, Parsons & Co. Ltd.)

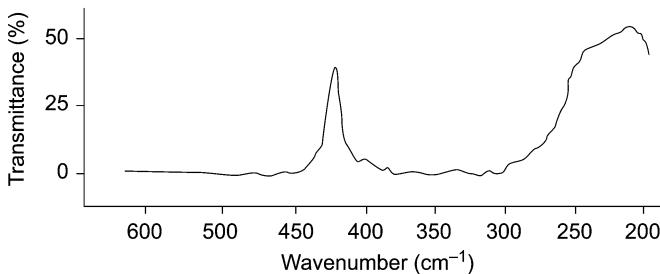


Figure 7.9. Measured transmittance of a Fabry-Perot filter for the far infrared. Design: $\text{Air}|LHLHHLH|\text{Ge}$ with H indicating a quarter-wave of germanium and L of caesium iodide. The rear surface of the substrate is unbloomed so that the effective transmission of the filter is 50%. (Courtesy of Sir Howard Grubb, Parsons & Co. Ltd.)

which are simply the earlier results multiplied by the factor $(n_H - n_L)/(n_H - n_L + n_L/m)$. It should be noted that these results are for first-order reflecting stacks and m th-order spacer. Clearly the effect of the phase is much greater the closer the two indices are in value and the lower the spacer order m . For the common visible and near infrared materials, zinc sulphide and cryolite, the factor for first-order spacers is equal to 0.43, while for infrared materials such as zinc sulphide and lead telluride it is greater, around 0.57. Figures 7.8 and 7.9 show the characteristics of typical all-dielectric narrowband Fabry-Perot filters.

Since the all-dielectric multilayer reflector is effective over a limited range only, sidebands of transmission appear on either side of the peak and in most applications must be suppressed. The shortwave sidebands can be removed very

easily by adding to the filter a longwave-pass absorption filter, readily available in the form of polished glass disks from a large number of manufacturers. Unfortunately, it is not nearly as easy to obtain shortwave-pass absorption filters and the rather shallow edges of those which are available tend considerably to reduce the peak transmission of the filter if the sidebands are effectively suppressed. The best solution to this problem is not to use an absorption type of filter at all, but to employ as a blocking filter a metal-dielectric filter of the type already discussed or of the multiple cavity type to be considered shortly. Because metal-dielectric filters used in the first order do not have longwave sidebands, they are very successful in this application. The metal-dielectric blocking filter can, in fact, be deposited over the all-dielectric filter in the same evaporation run provided that the layers are monitored using the narrowband filter itself as the test glass—this is known as direct monitoring—but more frequently a completely separate metal-dielectric filter is used. The various components which go to make up the final filter are cemented together in one assembly.

Before we leave the Fabry-Perot filters we can examine the effects of absorption losses in the layers in a manner similar to that already employed in chapter 5, where we were concerned with quarter-wave stacks. The problem has been investigated by many workers. The account which follows relies heavily on the work of Hemingway and Lissberger [5], but with slight differences.

We apply the method of chapter 5 directly. There, we recall, we showed that the loss in a weakly absorbing multilayer was given by

$$A = (1 - R) \sum \mathcal{A}$$

where, for *quarter-waves*,

$$\begin{aligned} \mathcal{A} &= \beta \left(\frac{n}{y_e} + \frac{y_e}{n} \right) \\ \beta &= \frac{2\pi k d}{\lambda} = \frac{2\pi n d}{\lambda} \frac{k}{n} = \frac{\pi k}{2 n}. \end{aligned}$$

y_e is the admittance of the structure on the emergent side of the layer, in free space units, $n - ik$ is the refractive index of the layer and d is the geometrical thickness. For quarter-waves, $nd = \lambda/4$.

The scheme is shown in table 7.2 where the admittance y_e is given at each interface and where alternative schemes for either high- or low-index spacers are included. The reflecting stacks are assumed to begin with high-index layers of which there are x per reflector, not counting the spacer.

We consider the case of low-index spacers first.

$$\begin{aligned} \sum \mathcal{A} &= \beta_H \left(\frac{n_m}{n_H} + \frac{n_H}{n_m} \right) + \beta_L \left(\frac{n_H^2}{n_L n_m} + \frac{n_L n_m}{n_H^2} \right) \\ &\quad + \beta_H \left(\frac{n_L^2 n_m}{n_H^3} + \frac{n_H^3}{n_L^2 n_m} \right) + \beta_L \left(\frac{n_H^4}{n_L^3 n_m} + \frac{n_L^3 n_m}{n_H^4} \right) + \dots \end{aligned}$$

Table 7.2.

Direction of incidence	
	n_m
n_H	n_H^2/n_m
n_L	$n_L^2 n_m / n_H^2$
n_H	$n_H^4 / (n_L^2 n_m)$
n_L	$n_L^4 n_m / n_H^4$
:	$n_H^{2x-2} / (n_L^{2x-4} n_m)$
n_L	$n_L^{2x-2} n_m / n_H^{2x-2}$
n_H	$n_H^{2x} / (n_L^{2x-2} n_m)$
$n_H^{2x} / (n_L^{2x-2} n_m)$	$n_H^{2x} / (n_L^{2x-2} n_m)$
n_L	$n_L^{2x} n_m / n_H^{2x}$
n_L	$n_L^{2x} n_m / n_H^{2x}$
Spacer	
:	
n_L	$n_L^{2x} n_m / n_H^{2x}$
n_L	$n_H^{2x} / (n_L^{2x-2} n_m)$
n_H	$n_L^{2x} n_m / n_H^{2x}$
n_H	$n_H^{2x+2} / (n_L^{2x} n_m)$
Spacer	
:	
n_H	$n_H^{2x+2} / (n_L^{2x} n_m)$
n_L	$n_L^{2x} n_m / n_H^{2x}$
n_H	$n_H^{2x} / (n_L^{2x-2} n_m)$
n_H	$n_L^{2x-2} n_m / n_H^{2x-2}$
:	
n_H	$n_H^4 / (n_L^2 n_m)$
n_H	$n_L^2 n_m / n_H^2$
n_L	n_H^2 / n_m
n_H	n_m

$$\begin{aligned}
& + \beta_L \left(\frac{n_H^{2x-2}}{n_L^{2x-3} n_m} + \frac{n_L^{2x-3} n_m}{n_H^{2x-2}} \right) + \beta_H \left(\frac{n_H^{2x-1}}{n_L^{2x-2} n_m} + \frac{n_L^{2x-2} n_m}{n_H^{2x-1}} \right) \\
& + m \left[\beta_L \left(\frac{n_H^{2x}}{n_L^{2x-1} n_m} + \frac{n_L^{2x-1} n_m}{n_H^{2x}} \right) + \beta_L \left(\frac{n_L^{2x-1} n_m}{n_H^{2x}} + \frac{n_H^{2x}}{n_L^{2x-1} n_m} \right) \right] \\
& + \beta_H \left(\frac{n_H^{2x-1}}{n_L^{2x-2} n_m} + \frac{n_L^{2x-2} n_m}{n_H^{2x-1}} \right) + \dots + \beta_H \left(\frac{n_H}{n_m} + \frac{n_m}{n_H} \right)
\end{aligned}$$

where the final set of terms is a repeat of the first and where the spacer consists of $2m$ quarter-waves. Rearranging, we find

$$\begin{aligned}
\sum \mathcal{A} = & 2\beta_H \left(\frac{n_m}{n_H} + \frac{n_L^2 n_m}{n_H^3} + \frac{n_L^4 n_m}{n_H^5} + \dots + \frac{n_L^{2x-2} n_m}{n_H^{2x-1}} \right) \\
& + 2\beta_H \left(\frac{n_H}{n_m} + \frac{n_H^3}{n_L^2 n_m} + \frac{n_H^5}{n_L^4 n_m} + \dots + \frac{n_H^{2x-1}}{n_L^{2x-2} n_m} \right) \\
& + 2\beta_L \left(\frac{n_L n_m}{n_H^2} + \frac{n_L^3 n_m}{n_H^4} + \dots + \frac{n_L^{2x-3} n_m}{n_H^{2x-2}} \right) \\
& + 2\beta_L \left(\frac{n_H^2}{n_L n_m} + \frac{n_H^4}{n_L^3 n_m} + \dots + \frac{n_H^{2x-2}}{n_L^{2x-3} n_m} \right) \\
& + 2m\beta_L \left(\frac{n_H^{2x}}{n_L^{2x-1} n_m} + \frac{n_L^{2x-1} n_m}{n_H^{2x}} \right)
\end{aligned}$$

where we have combined similar terms due to the two mirrors and where the final term is due to the spacer. The first four terms are geometric series and therefore, since $(n_L/n_H) < 1$,

$$\begin{aligned}
\sum \mathcal{A} = & 2\beta_H \frac{n_m}{n_H} \frac{[1 - (n_L/n_H)^{2x}]}{[1 - (n_L/n_H)^2]} \\
& + 2\beta_H \frac{n_H^{2x-1}}{n_L^{2x-2} n_m} \frac{[1 - (n_L/n_H)^{2x-2}]}{[1 - (n_L/n_H)^2]} \\
& + 2\beta_L \frac{n_L n_m}{n_H^2} \frac{[1 - (n_L/n_H)^{2x-2}]}{[1 - (n_L/n_H)^2]} \\
& + 2\beta_L \frac{n_H^{2x-2}}{n_L^{2x-3} n_m} \frac{[1 - (n_L/n_H)^{2x-2}]}{[1 - (n_L/n_H)^2]} \\
& + 2m\beta_L \left[\frac{n_H^{2x}}{n_L^{2x-1} n_m} + \frac{n_L^{2x-1} n_m}{n_H^{2x}} \right].
\end{aligned}$$

(n_L/n_H) will usually be rather less than unity and x will normally be large and so we can make the usual approximations and neglect terms such as $(n_L/n_H)^{2x}$

in the numerators and also those terms which have (n_m/n_H) as a factor compared with $(n_L/n_m)(n_H/n_L)^{2x-1}$ etc. Then the expression simplifies to

$$\begin{aligned}\sum \mathcal{A} &= 2\beta_H \frac{n_H^{2x-1}}{n_L^{2x-2} n_m} \frac{1}{[1 + (n_L/n_H)^2]} \\ &\quad + 2\beta_L \frac{n_H^{2x-2}}{n_H^{2x-3} n_m} \frac{1}{[1 + (n_L/n_H)^2]} \\ &\quad + 2m\beta_L \frac{n_H^{2x}}{n_L^{2x-1} n_m}.\end{aligned}$$

But

$$\begin{aligned}\beta_H &= \frac{2\pi n_H d}{\lambda} \frac{k_H}{n_H} = \frac{\pi}{2} \frac{k_H}{n_H} \\ \beta_L &= \frac{\pi}{2} \frac{k_L}{n_L}.\end{aligned}$$

Thus

$$\begin{aligned}\sum \mathcal{A} &= \frac{\pi k_H (n_H^{2x}/n_m n_L^{2x-2}) + \pi k_L (n_H^{2x}/n_m n_L^{2x-2})}{(n_H^2 - n_L^2)} + \frac{\pi m k_L n_H^{2x}}{n_L^{2x} n_m} \\ &= \frac{\pi n_H^{2x}}{n_m n_L^{2x}} \left(\frac{n_L^2 k_H + n_L^2 k_L}{(n_H^2 - n_L^2)} + m k_L \right).\end{aligned}$$

The absorption is then given by $A = (1 - R) \sum \mathcal{A}$. If the incident medium has index n_0 , then, since the terminating admittance in table 7.2 is n_m ,

$$R = \left(\frac{n_0 - n_m}{n_0 + n_m} \right)^2$$

and therefore

$$(1 - R) = \frac{4n_0 n_m}{(n_0 + n_m)^2}.$$

The above expression for $\sum \mathcal{A}$ should, therefore, be multiplied by the factor $4n_0 n_m / (n_0 + n_m)^2$ to yield the absorption. However, the filters should be designed so that they are reasonably well matched into the incident medium and therefore this factor will be unity, or sufficiently near unity. The absorption is then given by $\sum \mathcal{A}$. That is:

$$A = \frac{\pi n_H^{2x}}{n_m n_L^{2x}} \left(\frac{n_L^2 k_H + n_L^2 k_L}{(n_H^2 - n_L^2)} + m k_L \right) \quad (7.24)$$

for low-index spacers.

For high-index spacers we work through a similar scheme and, with the same approximations, we arrive at

$$A = \frac{\pi n_H^{2x}}{n_m n_L^{2x}} \left(\frac{n_L^2 k_H + n_H^2 k_L}{(n_H^2 - n_L^2)} + m k_H \right) \quad (7.25)$$

for high-index spacers.

It should be noted that, since x is the number of high-index layers, the filter represented by equation (7.25) will be narrower than that represented by equation (7.24) for equal x .

A useful set of alternative expressions can be obtained if we substitute equations (7.22) and (7.23) into equations (7.24) and (7.25) to give:

High-index spacer

$$A = 4 \frac{\lambda_0}{\Delta \lambda_h} \frac{\{k_L + k_H[m + (1-m)(n_L/n_H)^2]\}}{(n_H + n_L)[m + (1-m)(n_L/n_H)]}. \quad (7.26)$$

Low-index spacer

$$A = 4 \frac{\lambda_0}{\Delta \lambda_h} \frac{\{k_L(n_H/n_L)[m + (1-m)(n_L/n_H)^2] + (n_L/n_H)k_H\}}{(n_H + n_L)[m + (1-m)(n_L/n_H)]}. \quad (7.27)$$

Figure 7.10 shows the value of A plotted for Fabry–Perot filters with $n_H = 2.35$ and $n_L = 1.35$, typical of zinc sulphide and cryolite. $(\lambda_0/\Delta \lambda_h)$ is taken as 100 and k_H and k_L as either zero or 0.0001. The effect of other values of $(\lambda_0/\Delta \lambda_h)$ or k can be estimated by multiplying by an appropriate factor. The approximations are reasonable for $k(\lambda_0/\Delta \lambda_h)$ less than around 0.1.

It is difficult to draw any general conclusions from figure 7.10 because the results depend on the relative magnitudes of k_H and k_L . However, except in the case of very low k_L , the high-index spacer is to be preferred. There are very good reasons connected with performance when tilted, with energy grasp and with the manufacture of filters, for choosing high- rather than low-index spacers.

In the visible and near infrared regions of the spectrum, materials such as zinc sulphide and cryolite are capable of halfwidths of less than 0.1 nm with useful peak transmittance. Uniformity is, however, a major difficulty for filters of such narrow bandwidths. At the 90%-of-peak points, the Fabry–Perot filter has a width which is one-third of the halfwidth. It is a good guide that the uniformity of the filter should be such that the peak wavelength does not vary by more than one-third of the halfwidth over the entire surface of the filter. This means that the effective increase in halfwidth due to the lack of uniformity is kept within some 4.5% of the halfwidth and the reduction in peak transmittance to less than 3% (these figures can be calculated using the expressions derived later for assessing the performance of filters in uncollimated incident light). For filters of less than 0.1 nm halfwidth this rule implies a variation of not more than 0.03 nm or 0.006% in terms of layer thickness, a very severe requirement even for quite small filters.

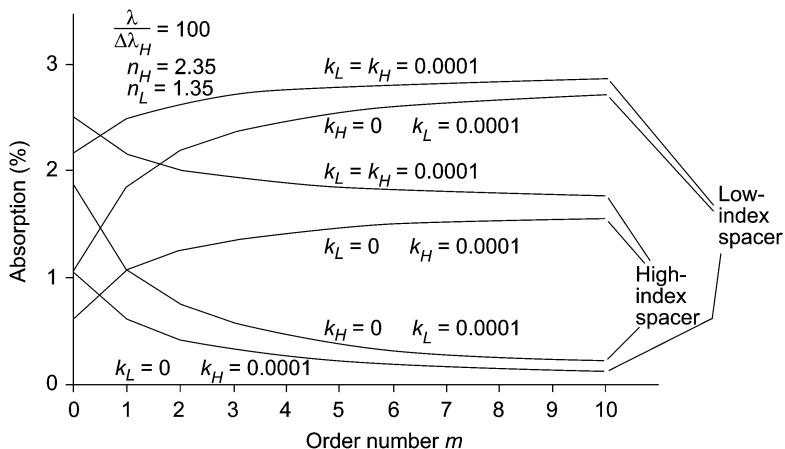


Figure 7.10. The value (expressed as a percentage) of the absorptance, as a function of the order number m , of Fabry–Perot filters with $\lambda_0/\Delta\lambda_h$ of 100 and values of extinction coefficients k_H and $k_H L$ of 0.0001 or zero. Other values can be accommodated by multiplying by an appropriate factor. n_H is taken as 2.35 and n_L as 1.35. The results are derived from equations (7.26) and (7.27).

Halfwidths of 0.3–0.5 nm are less demanding and can be produced more readily provided considerable care is taken. For narrower filters use is often made of the solid etalon filters now to be described.

7.2.3 The solid etalon filter

A solid etalon filter, or, as it is sometimes called, a solid spacer filter, is a very high-order Fabry–Perot filter in which the spacer consists of an optically worked plate or a cleaved crystal. Thin-film reflectors are deposited on either side of the spacer in the normal way, so that the spacer also acts as the substrate. The problems of uniformity which exist with all-thin-film narrowband filters are avoided and the thick spacer does not suffer from the increased scattering losses which always seem to accompany the higher-order thin-film spacers. The solid etalon filter is very much more robust and stable than the conventional air-spaced Fabry–Perot etalon, while the manufacturing difficulties are comparable. The high order of the spacer implies a small interval between orders and a conventional thin-film narrowband filter must be used in series with it to eliminate the unwanted orders.

An early account of the use of mica for the construction of filters of this type is that of Dobrowolski [6] who credits Billings with being the first to use mica in this way, achieving halfwidths of 0.3 nm. Dobrowolski obtained rather narrower pass bands and his is the first complete account of the technique. Mica

can be cleaved readily to form thin sheets with flat parallel surfaces, but there is a complication due to the natural birefringence of mica which means that the position of the pass band depends on the plane of polarisation. This splitting of the pass band can be avoided by arranging the thickness of the mica such that it is a half-wave plate, or multiple half-wave, at the required wavelength. If the two refractive indices are n_0 and n_e , this implies

$$\frac{2\pi(n_0 - n_e)d}{\lambda} = p\pi \quad p = 0, \pm 1, \pm 2, \dots$$

The order of the spacer will then be given by

$$m = \frac{n_0 p}{(n_0 - n_e)} \quad \text{or} \quad \frac{n_e p}{(n_0 - n_e)}$$

depending on the plane of polarisation. The difference between these two values is p , but, since p is small, the bandwidth will be virtually identical. The separation of orders for large m is given approximately by λ/m . Dobrowolski found that the maximum order separation, corresponding to $p = 1$, was given by 1.64 nm at 546.1 nm. With such spacers, around 60 μm thick, filters with halfwidths around 0.1 nm, the narrowest 0.085 nm, were constructed. Peak transmission ranged up to 50% for the narrower filters and up to 80% for slightly broader ones with around 0.3 nm halfwidth.

More recent work on solid etalon filters has concentrated on the use of optically worked materials as spacers. These must be ground and polished so that the faces have the necessary flatness and parallelism. The most complete account so far of the production of such filters is by Austin [7]. Fused silica spacers as thin as 50 μm have been produced with the necessary parallelism for halfwidths as narrow as 0.1 nm in the visible region, while thicker discs can give bandwidths as narrow as 0.005 nm. A 50- μm fused silica spacer gives an interval between orders of around 1.4 nm in the visible region which allows the suppression of unwanted orders to be fairly readily achieved by conventional thin-film narrowband filters.

The process of optical working tends to produce an error in parallelism over the surface of the spacer which is ultimately independent on the thickness of the spacer. Let us denote the total range of spacer thickness due to this lack of parallelism and to any deviation from flatness by Δd . This variation in spacer thickness causes the peak wavelength of the filter to vary. We can take an absolute limit for these variations as half the bandwidth of the filter. Then the resultant halfwidth will be increased by just over 10% and the peak transmittance reduced by just over 7% (once again using the expressions which we will shortly establish for filter performance in uncollimated light). We can write

$$\Delta\lambda_0/\lambda_0 = \Delta D/D = \Delta d/d \leq 0.5\Delta\lambda_h/\lambda_0$$

where D is the optical thickness nd of the spacer, $\Delta\lambda_0$ is the error in peak wavelength and $\Delta\lambda_h$ is the halfwidth. But

$$\text{Resolving power} = \lambda_0/\Delta\lambda_h = m\mathcal{F}$$

and hence, since

$$D = m\lambda_0/2$$

$$\mathcal{F} \leq \frac{0.25\lambda_0}{\Delta D}.$$

Now the attainable ΔD in the visible region is of the order of $\lambda/100$ and this means that the limiting finesse is around 25, independent of the spacer thickness. High resolving power then has to be achieved by the order number m which determines both the spacer thickness $D = m\lambda_0/2$ and the interval between orders λ_0/m . For a halfwidth of 0.01 nm at, say, 500 nm the resolving power is 50 000. The finesse of 25 implies an order number of 2000, a spacer optical thickness of 500 μm and an interval between orders of 0.25 nm. This very restricted range between orders means that it is very difficult to carry out sideband blocking by a thin-film filter directly. Instead, a broader solid etalon filter can be used with its corresponding greater interval between orders. It, in its turn, can be suppressed by a thin-film filter. For a halfwidth of 0.1 nm, a spacer optical thickness of 50 μm is required which gives an interval between orders of 2.5 nm.

The temperature coefficient of peak wavelength change of solid etalon filters with fused silica spacers is 0.005 nm $^{\circ}\text{C}^{-1}$ and the filters may be finely tuned by altering this temperature.

Candille and Saurel [8] have used Mylar foil as the spacer. Their filters were strictly of the multiple cavity type described later in this chapter. The Mylar acted as a substrate and a high-order spacer. One of the reflectors included a low-order Fabry–Perot filter which served both as blocking filter to eliminate the additional unwanted orders of the Mylar section and as an additional cavity to steepen the sides of the pass band. The position of the pass band could be altered by varying the tension in the Mylar. The filters were not as narrow as the other solid etalon filters which have been mentioned, halfwidths of 0.8–1.0 nm being obtained.

Solid etalon filters have also been constructed for the infrared. Smith and Pidgeon [9] used a polished slab of germanium some 780 μm thick working at around 700 cm^{-1} in the 400th order. Both faces were coated with a quarter-wave of zinc sulphide followed by a quarter-wave of lead telluride to give a reflectance of 62%, a fringe halfwidth of 0.1 cm^{-1} and an interval between orders of 1.6 cm^{-1} . This particular arrangement was designed so that the lines in the R-branch of the CO₂ spectrum, which are spaced at 1.6 cm^{-1} apart at around 14.5 μm , should be exactly matched by a number of adjacent orders. Order sorting was not, therefore, a problem.

Roche and Title [10] have reported a range of solid etalon filters for the infrared. These filters are some 13 mm in diameter, have resolving powers in the region of 3×10^4 and the techniques used for their construction are as reported by Austin [7]. For wavelengths equal to or shorter than 3.5 μm , fused silica spacers are quite satisfactory. For longer wavelengths Yttralox, a combination of yttrium and thorium oxides, was found most satisfactory. With this material, solid etalon filters were produced which at 3.334 μm had halfwidths as low as 0.2 nm and at

4.62 μm , 0.8 nm. At these wavelengths, the attainable finesse was 30–40 and the current limit to the halfwidth which can be achieved is the permissible interval between orders which determines the arrangement of subsidiary blocking filters.

7.2.4 The effect of varying the angle of incidence

As we have seen with other types of thin-film assembly the performance of the all-dielectric Fabry–Perot varies with angle of incidence, and this effect is particularly important when considering, for instance, the allowable focal ratio of the pencil being passed by the filter or the maximum tilt angle in any application. The variation with angle of incidence is not altogether a bad thing because the effect can be used to tune filters which would otherwise be off the desired wavelength—very important from the manufacturer's point of view because it enables him to ease a little the otherwise almost impossibly tight production tolerances.

The effect of tilting has been studied by a number of workers, particularly by Dufour and Herpin [11], Lissberger [12], Lissberger and Wilcock [13] and Pidgeon and Smith [14]. For our present purposes we follow Pidgeon and Smith since their results are in a slightly more suitable form.

7.2.4.1 Simple tilts in collimated light

The phase thickness of a thin film at oblique incidence is

$$\delta = 2\pi nd \cos \theta / \lambda$$

which can be interpreted as an apparent optical thickness of $nd \cos \theta$ which varies with angle of incidence so that layers seem thinner when tilted. Although the optical admittance changes with tilts, in narrowband filters the predominant effect is the apparent change in thickness which moves the filter pass band to shorter wavelengths.

For an ideal Fabry–Perot filter with spacer layer index n^* , where the reflectors have constant phase shift of zero or π regardless of the angle of incidence or wavelength, we can write for the position of peak wavelength in the m th order

$$2\pi n^* d \cos \theta / \lambda = m\pi$$

i.e.

$$(2\pi n^* d / \lambda_0) g \cos \theta = m\pi$$

i.e.

$$g \cos \theta = 1$$

$$\Delta g = \left(\frac{1}{\cos \theta} - 1 \right).$$

If the angle of incidence is θ_i in air then

$$\theta = \sin^{-1}(\sin \theta_i / n^*)$$

and Δg is given in terms of θ_i and n^* . The effect of tilting, then, in this ideal filter can be estimated simply from a knowledge of the index of the spacer and the angle of incidence. For small angles of incidence, the shift is given by

$$\Delta g = \Delta \nu / \nu_0 = \Delta \lambda / \lambda_0 = \theta_i^2 / 2n^{*2}. \quad (7.28)$$

The index of the spacer n^* determines its sensitivity to tilt: the higher the index, the less the filter is affected.

In the case of a real filter, the reflectors are also affected by the tilting and so the calculation of the shift in peak wavelength is more involved. It has, however, been shown by Pidgeon and Smith that the shift is similar to that which would have been obtained from an ideal filter with spacer index n^* , intermediate between the high and low indices of the layers of the filter. n^* is known as the effective index. This concept of the effective index holds good for quite high angles of incidence, up to 20° or 30° or even higher, depending on the indices of the layers making up the filter.

We can estimate the effective index for the filter by a technique similar to that already used for metal–dielectrics (equation (7.3)). We retain our assumption of small angle of incidence and small changes in g around the value which corresponds to the peak at normal incidence.

The peak position is given, as before, by

$$\sin^2[(2\pi nd \cos \theta / \lambda) - \phi] = 0 \quad (7.29)$$

with, at normal incidence

$$\sin^2[(2\pi nd / \lambda_0) - \phi_0] = 0. \quad (7.30)$$

Now ϕ_0 is 0 or π and so equation (7.30) is satisfied by

$$2\pi nd / \lambda_0 = m\pi \quad m = 0, 1, 2, \dots$$

The analysis is once again easier in terms of g ($= \lambda_0 / \lambda = \nu / \nu_0$). Equation (7.29) becomes

$$\sin^2[(2\pi nd / \lambda_0)g \cos \theta - \phi_0 - \Delta\phi] = 0. \quad (7.31)$$

We write

$$g = 1 + \Delta g \quad \text{and} \quad \cos \theta \simeq 1 - \theta^2 / 2.$$

However, we should work in terms of θ_i , the external angle of incidence, which we assume is referred to free space (if not, then we make the appropriate correction). Then

$$n \sin \theta = n_i \sin \theta_i = \sin \theta_i$$

and, using equation (7.31),

$$\sin^2[(2\pi nd/\lambda_0) - \phi_0 + m\pi\Delta g - (m\pi\theta_i^2/2n^2) - \Delta\phi] = 0$$

is the condition for the new peak position. This requires

$$m\pi\Delta g - (m\pi\theta_i^2/2n^2) - \Delta\phi = 0. \quad (7.32)$$

Now $\Delta\phi$ is a function of θ and Δg and to evaluate it we return to equations (7.20) and (7.21). The layers in the reflectors are all quarter-waves and so ε is given by

$$\pi/2 + \varepsilon = (2\pi nd/\lambda_0)g \cos\theta = (\pi/2)(1 + \Delta g)(1 - \theta^2/2)$$

but

$$\theta = \theta_i/n$$

so that

$$\varepsilon = (\pi/2)\Delta g - \pi\theta_i^2/4n^2$$

with n being either n_L or n_H for ε_L or ε_H respectively.

At this stage we are forced to consider high-index and low-index spacers separately.

7.2.4.2 Case I: high-index spacers

From equation (7.20) we have, inserting n_H for n_0 ,

$$\begin{aligned} \Delta\phi &= -\frac{2n_L^2}{(n_H^2 - n_L^2)}\varepsilon_H - \frac{2n_H n_L}{(n_H^2 - n_L^2)}\varepsilon_L \\ &= \frac{2n_L^2}{(n_H^2 - n_L^2)}\left(\frac{\pi}{2}\Delta g - \frac{\pi\theta_i^2}{4n_H^2}\right) - \frac{2n_H n_L}{(n_H^2 - n_L^2)}\left(\frac{\pi}{2}\Delta g - \frac{\pi\theta_i^2}{4n_L^2}\right) \\ &= -\frac{\pi n_L}{(n_H - n_L)}\Delta g + \frac{\pi}{2}\frac{(n_L^2 - n_L n_H + n_H^2)}{n_H^2 n_L (n_H - n_L)}\theta_i^2 \end{aligned}$$

and equation (7.32) becomes

$$m\pi\Delta g - \frac{m\pi\theta_i^2}{2n_H^2} + \frac{\pi n_L \Delta g}{(n_H - n_L)} - \frac{\pi}{2}\frac{(n_L^2 - n_L n_H + n_H^2)}{n_H^2 n_L (n_H - n_L)}\theta_i^2 = 0$$

giving, after some manipulation and simplification

$$\Delta g = \frac{1}{n_H^2} \frac{[(m-1) - (m-1)(n_L/n_H) + (n_H/n_L)]}{[m - (m-1)(n_L/n_H)]} \left(\frac{\theta_i^2}{2}\right).$$

But, comparing the expression with equation (7.28) we find

$$n^{*2} = \frac{n_H^2 [m - (m - 1)(n_L/n_H)]}{[(m - 1) - (m - 1)(n_L/n_H) + (n_H/n_L)]}$$

or

$$n^* = n_H \left(\frac{m - (m - 1)(n_L/n_H)}{(m - 1) - (m - 1)(n_L/n_H) + (n_H/n_L)} \right)^{1/2}. \quad (7.33)$$

For first-order filters

$$n^* = (n_H n_L)^{1/2} \quad (7.34)$$

which is the result obtained by Pidgeon and Smith. As $m \rightarrow \infty$ then $n^* \rightarrow n_H$, as we would expect.

7.2.4.3 Case II: low-index spacer

The analysis is exactly as for case I except that equation (7.21) is used and the n in equation (7.32) becomes n_L :

$$n^* = n_L \left(\frac{m - (m - 1)(n_L/n_H)}{m - m(n_L/n_H) + (n_L/n_H)^2} \right)^{1/2}. \quad (7.35)$$

For first-order filters

$$n^* = \frac{n_L}{[1 - (n_L/n_H) + (n_L/n_H)^2]^{1/2}} \quad (7.36)$$

which is, again, the expression given by Pidgeon and Smith and we note again that as $m \rightarrow \infty$ then $n^* \rightarrow n_L$.

Typical curves showing how the effective index n^* varies with order number for both low- and high-index spacers are given in figure 7.11.

Pidgeon and Smith made experimental measurements on narrowband filters for the infrared. The designs in question were

- (a) $L|Ge|LHLH LL H L H|Air$
- (b) $L|Ge|LHLHLL HH LHLH|Air$

where H represents a quarter-wave thickness of lead telluride and L of zinc sulphide, and where the peak wavelength was in the vicinity of 15 μm . Calculations of shift were carried out by the approximate method using n^* and by the full matrix method without approximations. The results using n^* matched the accurate calculations up to angles of incidence of 40° to an accuracy representing $\pm 2\%$ change in n^* . The experimental points showed good agreement with the theoretical estimates. Some of the results are shown in figures 7.12 and 7.13.

The angle of incidence may be in a medium other than free space, in which case equation (7.28) becomes

$$\Delta g = \Delta\lambda_0/\lambda = \Delta\nu_0/v = \frac{1}{2}(n_i\theta_i/n^*)^2 \quad (7.37)$$

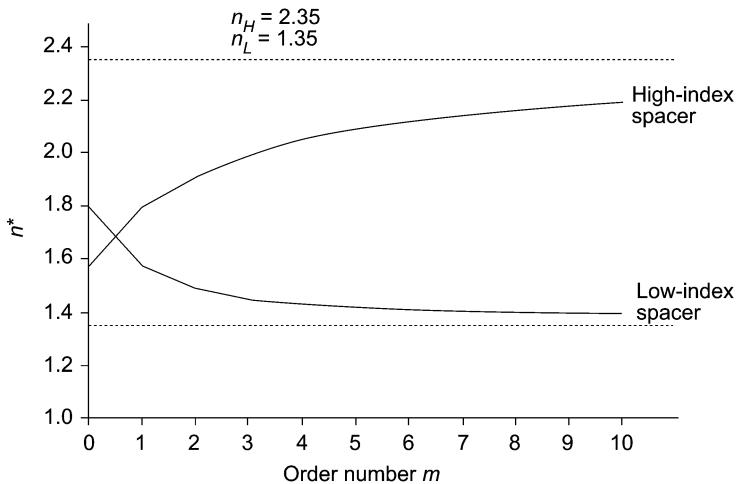


Figure 7.11. The effective index n^* plotted against order number m for Fabry–Perot filters constructed of materials such as zinc sulphide, $n = 2.35$, and cryolite, $n = 1.35$. The results were calculated from expressions (7.35) and (7.36).

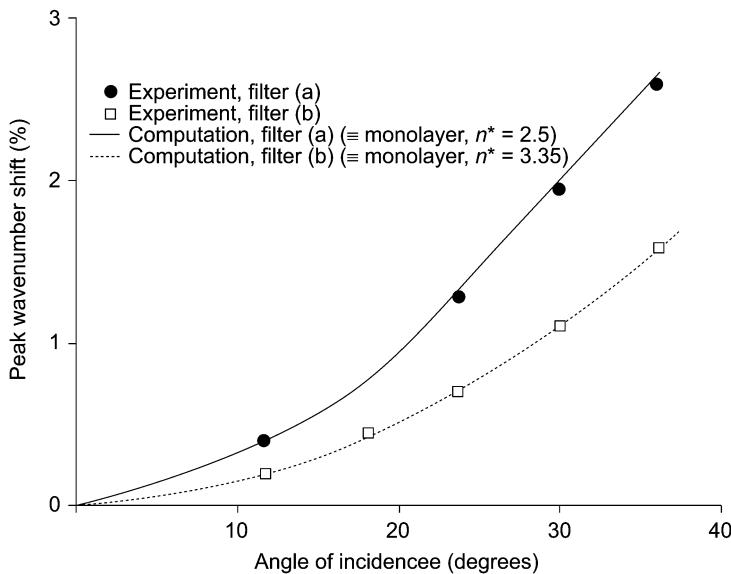


Figure 7.12. The shift of peak wavenumber with scanning angle for two Fabry–Perot filters in collimated light. In both cases the monolayer curves fit the computed curves to $\pm 2\%$ in n . (After Pidgeon and Smith [14].)

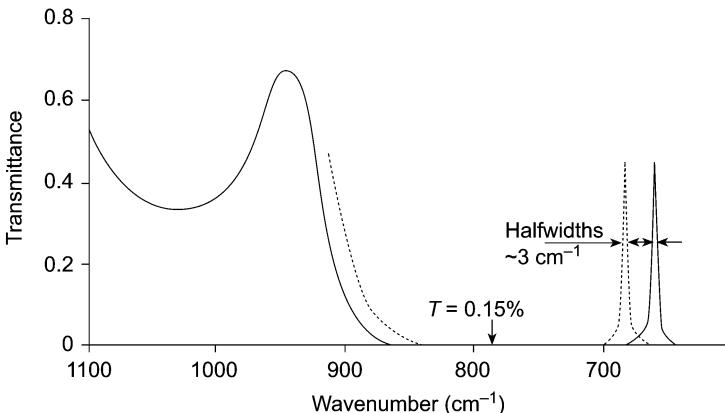


Figure 7.13. Measured transmittance of two filters of type (b). Design: Air |*HHLH LHLHLH LHLHLH*|Ge substrate |*L*|Air (*H* = PbTe, *L* = ZnS). (After Pidgeon and Smith [14].)

where θ_i is measured in radians.

If θ_i is measured in degrees, then

$$\Delta g = \Delta\lambda_0/\lambda = \Delta\nu_0/\nu_0 = 1.5 \times 10^{-4} (n_i/n^*)^2 \theta_i^2. \quad (7.38)$$

7.2.4.4 Effect of an incident cone of light

The analysis can be taken a stage further to arrive at expressions for the degradations of peak transmission and bandwidth which become apparent when the incident illumination is not perfectly collimated. Essentially the same results have been obtained by Lissberger and Wilcock [13] and by Pidgeon and Smith [14].

It is assumed first of all that, in collimated light, the sole effect of tilting a filter is a shift of the characteristic towards shorter wavelengths or greater wavenumbers, leaving the peak transmittance and bandwidth virtually unchanged. The performance in convergent or divergent light is then given by integrating the transmission curve over a range of angles of incidence. The analysis is simpler in terms of wavenumber or of g , rather than wavelength. If ν_0 is the wavenumber corresponding to the peak at normal incidence and ν_Θ to the peak at angle of incidence Θ , then it is plausible that the resultant peak, when all angles of incidence in the cone from 0 to Θ are included, should appear at a wavenumber given by the mean of the above extremes. We shall show, shortly, that this is indeed the case. The new peak is given by

$$\nu_m = \nu_0 + \frac{1}{2} \Delta\nu' \quad (7.39)$$

where

$$\Delta\nu' = \nu_\Theta - \nu_0 = \nu_0\Theta^2/2n^{*2}.$$

The effective bandwidth of the filter will, of course, appear broader and, since the process is, in effect, a convolution of a function with bandwidth W_0 , which is the width of the filter at normal incidence, and another function with bandwidth $\Delta\nu'$, the change in peak position produced by altering the angle of incidence from 0 to Θ , it seems likely that the resultant bandwidth might be given by the square root of the sum of their squares. This too is indeed the case, as we shall also show.

$$W_\Theta^2 = W_0^2 + (\Delta\nu')^2. \quad (7.40)$$

The peak transmission falls and is given by

$$\hat{T}_\Theta = \left(\frac{W_0}{\Delta\nu'} \right) \tan^{-1} \left(\frac{\Delta\nu'}{W_0} \right). \quad (7.41)$$

The analysis is as follows.

We consider incident light in the form of a cone with semiangle Θ , that is a cone of focal ratio $1/(2 \tan \Theta)$. We assume that in collimated light the effect of tilting the filter is simply to move the characteristic towards shorter wavelengths, leaving the bandwidth and peak transmittance unchanged.

For small values of θ , the flux incident on the filter is proportional to $\theta d\theta$. The resultant transmittance of the filter is then given by the total flux transmitted divided by the total flux incident.

The total flux incident is proportional to

$$\int_0^\Theta \theta d\theta = \frac{1}{2}\Theta^2.$$

The total flux transmitted is proportional to

$$\int_0^\Theta \theta T d\theta.$$

We can, for small values of θ and Δg , set

$$T = \frac{1}{1 - \{(2/\Delta g_h)[\Delta g - (\theta_i^2/2n^{*2})]\}^2}$$

where Δg_h is the halfwidth at normal incidence of the filter in units of g . This expression follows directly from the concept of n^* . The transmittance of the filter is then given by

$$T = \frac{2}{\Theta^2} \int_0^\Theta \frac{\theta d\theta}{1 + \{(2/\Delta g_h)[\Delta g - (\theta_i/2n^{*2})]\}^2}$$

$$\begin{aligned}
&= -\frac{2}{\Theta^2} \frac{n^{*2} \Delta g_h}{2} \left[\tan^{-1} \left\{ \frac{2}{\Delta g_h} \left(\Delta g - \frac{\theta_i}{2n^{*2}} \right) \right\} \right]_0^\Theta \\
&= \frac{1}{2} \frac{\Delta g_h}{(\Theta^2/2n^{*2})} \left\{ \tan^{-1} \left(2 \frac{\Delta g}{\Delta g_h} \right) - \tan^{-1} \left[2 \left(\frac{\Delta g}{\Delta g_h} - \frac{\Theta^2}{2n^{*2}} \frac{1}{\Delta g_h} \right) \right] \right\} \quad (7.42)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \frac{\Delta g_h}{(\Theta^2/2n^{*2})} \left[\tan^{-1} \left(\frac{(2/\Delta g_h)(\Theta^2/2n^{*2})}{1 + (2/\Delta g_h)^2 \{ \Delta g [\Delta g - (\Theta^2/2n^{*2})] \}} \right) \right]. \quad (7.43)
\end{aligned}$$

This is a maximum when

$$\Delta g = \frac{1}{2} \frac{\Theta^2}{2n^{*2}}.$$

But $\Theta^2/(2n^{*2})$ is the shift in the position of the peak at angle of incidence Θ . Thus in a cone of light of semiangle Θ , the peak wavelength of the filter is given by the mean of the value at normal incidence and that at the angle Θ corresponding to equation (7.39). The value of the peak transmittance is then, from equation (7.42),

$$\frac{\Delta g_h}{(\Theta^2/2n^{*2})} \tan^{-1} \left(\frac{\Theta^2/2n^{*2}}{\Delta g_h} \right)$$

which corresponds to equation (7.41).

The half-peak points are given by

$$(7.43) = \frac{1}{2}(\text{peak } T)$$

i.e.

$$\begin{aligned}
&\frac{1}{2} \cdot \frac{\Delta g_h}{(\Theta^2/2n^{*2})} \tan^{-1} \left(\frac{(2/\Delta g_h)(\Theta^2/2n^{*2})}{1 + (2/\Delta g_h)^2 \{ \Delta g [\Delta g - (\Theta^2/2n^{*2})] \}} \right) \\
&= \frac{1}{2} \frac{\Delta g_h}{(\Theta^2/2n^{*2})} \tan^{-1} \left(\frac{\Theta^2/2n^{*2}}{\Delta g_h} \right)
\end{aligned}$$

which is satisfied by

$$1 + \left(\frac{2}{\Delta g_h} \right) \left[\Delta g \left(\Delta g - \frac{\Theta^2}{2n^{*2}} \right) \right] = 2$$

i.e.

$$\Delta g \left(\Delta g - \frac{\Theta^2}{2n^{*2}} \right) - \left(\frac{\Delta g_h}{2} \right)^2 = 0.$$

We are interested in the difference between the roots of the equation which is the width of the characteristic

$$(\Delta g_1 - \Delta g_2) = \left[\left(\frac{\Theta^2}{2n^{*2}} \right)^2 + (\Delta g_h)^2 \right]^{1/2}$$

which corresponds exactly to equation (7.40).

Since

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots \quad \text{for } |x| \leq 1$$

for small values of $(\Delta\nu' / W_0)$ we can write

$$\hat{T}_\Theta = 1 - \frac{1}{3} \left(\frac{\Delta\nu'}{W_0} \right)^2. \quad (7.44)$$

If FR denotes the focal ratio of the incident light, then, for values of around 2 to infinity, it is a reasonably good approximation that

$$\Theta = 1/[2(FR)].$$

Using this, we find another expression for $\Delta\nu'$ which can be useful:

$$\Delta\nu' = \frac{\nu_0}{8n^{*2}(FR)^2}.$$

We can extend this analysis still further to the case of a cone of semiangle Θ incident at an angle other than normal, provided we make some simplifying assumptions. If the angle of incidence of the cone is χ then the range of angles of incidence will be $\chi \pm \Theta$.

If $\chi < \Theta$ then we can assume that the result is simply that for a normally incident cone of semiangle $\chi + \Theta$.

If $\chi > \Theta$ then we have three frequencies, ν_0 corresponding to normal incidence, ν_1 to angle of incidence $\chi - \Theta$, and ν_2 to angle of incidence $\chi + \Theta$. The new filter peak can be assumed to be

$$\begin{aligned} \frac{1}{2}(\nu_1 + \nu_2) &= \frac{\chi^2 + \Theta^2}{2n^{*2}} \nu_0 \quad (\chi \text{ and } \Theta \text{ in radians}) \\ &= \frac{1.52 \times 10^{-4}(\chi^2 + \Theta^2)}{n^{*2}} \nu_0 \quad (\chi \text{ and } \Theta \text{ in degrees}). \end{aligned} \quad (7.45)$$

The halfwidth is

$$[W_0^2 + (\nu_2 - \nu_1)^2]^{1/2}$$

where

$$\begin{aligned} (\nu_2 - \nu_1) &= \frac{2\chi\Theta}{n^{*2}} \nu_0 \quad (\chi \text{ and } \theta \text{ in radians}) \\ &= \frac{6.09 \times 10^{-4}\chi\Theta}{n^{*2}} \nu_0 \quad (\chi \text{ and } \Theta \text{ in degrees}) \end{aligned} \quad (7.46)$$

and the peak transmittance is

$$\frac{W_0}{(\nu_2 - \nu_1)} \tan^{-1} \left(\frac{(\nu_2 - \nu_1)}{W_0} \right) \simeq 1 - \frac{1}{3} \left(\frac{(\nu_2 - \nu_1)}{W_0} \right)^2. \quad (7.47)$$

$(\nu_2 - \nu_1)$ is proportional to $\Theta \chi$ and Hernandez [15] has found excellent agreement between measurements made on real filters and calculations from these expressions for values of $\Theta \chi$ up to 100° .

We can illustrate the use of these expressions in calculating the performance of a zinc sulphide and cryolite filter for the visible region. We assume that this is a low-index first-order filter with a bandwidth of 1%.

For this filter we calculate that $n^* = 1.55$. We take 10% reduction in peak transmittance as the limit of what is acceptable. Then, from equation (7.47)

$$(\nu_2 - \nu_1)/W_0 = 0.55$$

and the increased halfwidth which corresponds to this reduction in peak transmittance is

$$(1 + 0.55^2)^{1/2} W_0 = 1.14 W_0$$

or an increase of 14% over the basic width.

At normal incidence, the cone semiangle which can be tolerated is given by

$$1.5 \times 10^4 (\Theta^2 / n^{*2}) = \Delta \nu = 0.55 W_0 = 0.55 \times 0.01 \quad (\Theta \text{ in degrees})$$

i.e.

$$\Theta = [1.55^2 \times 0.55 \times 0.01 / (1.5 \times 10^{-4})]^{1/2} = 9.4^\circ.$$

Such a cone at normal incidence will cause a shift in the position of the peak towards shorter wavelengths or higher frequencies of

$$\frac{1}{2} (\Delta \nu' / \nu_0) = (\frac{1}{2} \times 0.55 \times 0.01) = 0.275\%.$$

Used at oblique incidence in a cone of illumination we have

$$(6.09 \times 10^{-4} \chi \Theta / n^{*2}) \nu_0 = \nu_2 - \nu_1 = 0.55 \times 0.01$$

i.e.

$$\chi \Theta = \frac{1.55^2 \times 0.55 \times 0.01}{6.09 \times 10^{-4}} = 21.7^\circ$$

which means that the filter can be used in a cone of semiangle 2° up to an angle of incidence of $21.7/2 = 10.9^\circ$ or of semiangle 3° up to an angle of incidence of 7° and so on.

One very important result is the shift in peak wavelength in a cone at normal incidence which indicates that if a filter is to be used at maximum efficiency in such an arrangement, its peak wavelength at normal incidence in collimated light should be slightly longer to compensate for this shift.

7.2.5 Sideband blocking

There is a disadvantage in the all-dielectric filter: the high-reflectance zone of the reflecting coating is limited in extent and hence the rejection zone of the filter is also limited. In the near ultraviolet, visible and near infrared regions, the transmission sidebands on the shortwave side of the peak can usually be suppressed, or blocked, by an absorption filter with a longwave-pass characteristic in the same way as for metal–dielectric filters. The longwave sidebands are more of a problem. These may be outside the range of sensitivity of the detector and therefore may not require elimination, but if they are troublesome then the usual technique for removing them is the addition of a metal–dielectric first-order filter with no longwave sidebands. It is usually very much broader than the narrowband component in order that the peak transmittance may be high. The metal–dielectric component is usually added as a separate component, but it can be deposited over the basic Fabry–Perot. Rather than a simple Fabry–Perot filter, a double cavity metal–dielectric is commonly used. Multiple cavity filters are the next topic of discussion.

7.3 Multiple cavity filters

The transmission curve of the basic all-dielectric Fabry–Perot filter is not of ideal shape. It can be shown that one half of the energy transmitted in any order lies outside the halfwidth (assuming an even distribution of energy with frequency in the incident beam). A more nearly rectangular curve would be a great improvement. Further, the maximum rejection of the Fabry–Perot is completely determined by the halfwidth and the order. The broader filters, therefore, tend to have poor rejection as well as a somewhat unsatisfactory shape.

When tuned electric circuits are coupled together, the resultant response curve is rather more rectangular and the rejection outside the pass band rather greater than a single tuned circuit, and a similar result is found for the Fabry–Perot filter. If two or more of these filters are placed in series, much the same sort of double peaked curve is obtained; it has, however, a much more promising shape than the single filter. The filters may be either metal–dielectric or all-dielectric and the basic form is

$$|\text{reflector}|\text{half-wave}|\text{reflector}|\text{half-wave}|\text{reflector}|$$

known as a double half-wave or DHW filter or as a double cavity or *two-cavity* filter. Some typical examples of all-dielectric DHW or two-cavity filters are shown in figure 7.14.

Such filters were certainly constructed by A F Turner and his co-workers at Bausch and Lomb in the early 1950s but the results were published only as quarterly reports in the Fort Belvoir Contract Series over the period 1950–68 [16]. The earliest filters were of the triple half-wave type, known at Bausch and Lomb as WADIS (wide-band all-dielectric interference filters) [17]. Double

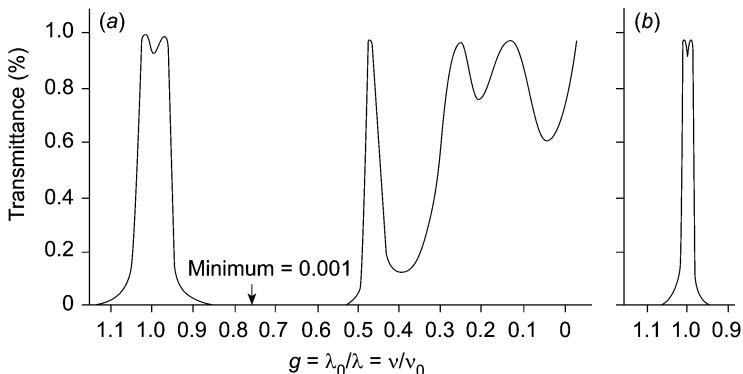


Figure 7.14. (a) Computed transmittance of *HLLHLHLLH*. (b) Computed transmittance of *HLHHLHLHHHLH*. In both cases $n_H = 4.0$ and $n_L = 1.35$. (After Smith [18].)

half-wave, or two-cavity, filters came later but were in routine use at Bausch and Lomb certainly by 1957. They were initially known as TADIs. The Fort Belvoir Contract¹ Reports make fascinating reading and show just how advanced the work at Bausch and Lomb was at that time. Use was being made of the concept of equivalent admittance for the design both of WADI filters and of the edge filters for blocking the sidebands. Multilayer antireflection coatings were also well understood.

The first complete account of a theory applicable to multiple half-wave filters was published by Smith [18] and it is his method that we follow first here.

The reflecting stacks in the classical Fabry–Perot filter have more or less constant reflectance over the pass band of the filter. A dispersion of phase change on reflection does, as we have seen, help to reduce the bandwidth, but this does so without altering the basic shapes of the pass-band shape. Smith suggested the idea of using reflectors with much more rapidly varying reflectance to achieve a better shape. The essential expression for the transmission of the complete filter has already been derived on p 75 where we have assumed $\beta = 0$, that is, no absorption in the spacer layer. From Smith's formula, equation (2.149),

$$T = \frac{|\tau_a^+|^2 |\tau_b^+|^2}{(1 - |\rho_a^-||\rho_b^+|)^2} \left[1 + \frac{4|\rho_a^-||\rho_b^+|}{(1 - |\rho_a^-||\rho_b^+|)^2} \sin^2 \frac{\phi_a + \phi_b - 2\delta}{2} \right]^{-1} \quad (7.48)$$

it can be seen that high transmission can be achieved at any wavelength if, and only if, the reflectances on either side of a chosen spacer layer are equal. Of course the phase condition must be met too, but this can be arranged by choosing

¹ These reports were obtainable from the Engineer Research and Development Laboratories, Fort Belvoir, Virginia 22060, USA, but are now out of print.

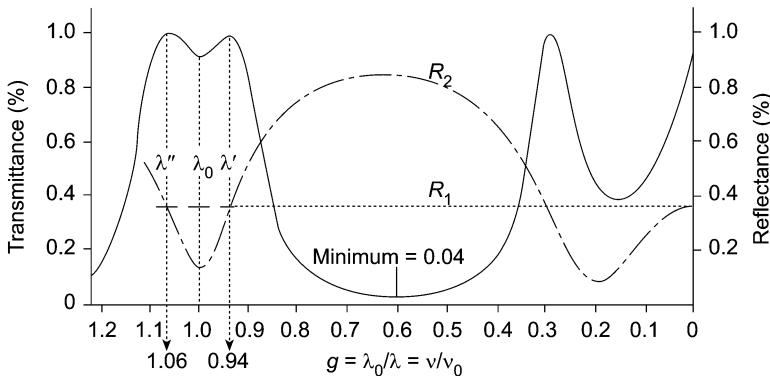


Figure 7.15. Computed transmittance of $HHLHH$ and explanatory reflectance curves R_1 and R_2 ($n_H = 4.0$, $n_L = 1.35$). (After Smith [18].)

the correct spacer thickness to make

$$\left| \frac{\phi_a + \phi_b}{2} - \delta \right| = m\pi.$$

In these expressions, the symbols have the same meanings as given in figure 2.19.

Smith pointed out the advantage of having reasonably low reflectance in the region around the peak wavelength, which means that absorption is less effective in limiting the peak transmittance. In the Fabry–Perot filter, low reflectance means wide bandwidth, but Smith limited the bandwidth by arranging for the reflectances to begin to differ appreciably at wavelengths only a little removed from the peak. This is illustrated in figure 7.15. The figure shows what is the simplest type of DHW filter, which has construction $HHLHH$. The HH layers are the two half-wave spacers and the L layer is a coupling layer. In the discussion which follows, for simplicity we shall ignore any substrate. The behaviour of the filter is described in terms of the reflectances on either side of one of the two spacers. R_1 is the reflectance of the interface between the high index and the surrounding medium, which we take as air with index unity, and is a constant. R_2 is the reflectance of the assembly on the other side of the spacer and is low at the wavelength at which the spacer is a half-wave and rises on either side. At wavelengths λ' and λ'' , the reflectances R_1 and R_2 are equal and we would expect to see high transmission if the phase condition is met, which in fact it is. The transmission of the assembly is also shown in the figure and the shape can be seen to consist of a steep-sided pass band with two peaks close together and only a slight dip in transmission between the peaks, much more like the ideal rectangle than the shape of the Fabry–Perot filter.

Smith's formula for the transmittance of a filter can be written:

$$T(\lambda) = T_0(\lambda) \frac{1}{1 + F(\lambda) \sin^2[(\phi_1 + \phi_2)/2 - \delta]} \quad (7.49)$$

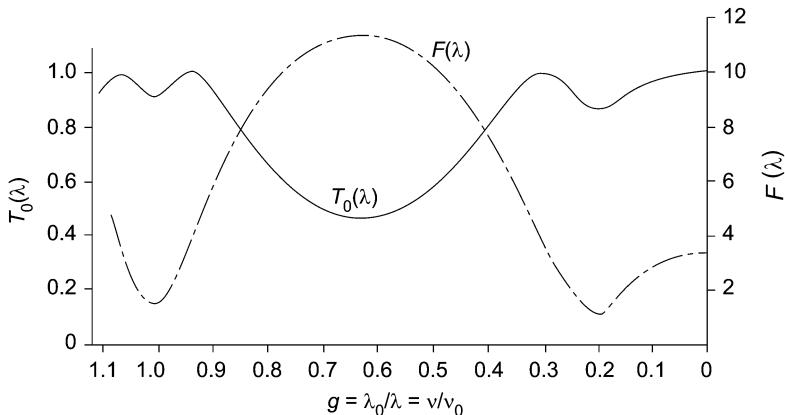


Figure 7.16. $T_0(\lambda)$ and $F(\lambda)$ for $HHLHH$. (After Smith [18].)

where

$$T_0(\lambda) = \frac{(1 - R_1)(1 - R_2)}{[1 - (R_1 R_2)^{1/2}]^2} \quad (7.50)$$

$$F(\lambda) = \frac{4(R_1 R_2)^{1/2}}{1 - (R_1 R_2)^{1/2}}. \quad (7.51)$$

Both these quantities are now variable since they involve R_2 , which is a variable. The form of the functions is also shown in figure 7.16. At wavelengths removed from the peak, $T_0(\lambda)$ is low and $F(\lambda)$ is high, the combined effect being to increase the rejection. In the region of the peak, T_0 is high, and, just as important, F is low, producing high transmittance which is not sensitive to the effects of absorption. As we have shown before, the peak transmittance is dependent on the quantity A/T , where A is the absorptance and T the transmittance of the reflecting stacks. Clearly, the greater T is, the higher A can be for the same overall filter transmittance.

The typical double-peaked shape of the double half-wave filter results from the intersection of the R_1 and R_2 curves at two separate points. Two other cases can arise. The curves can intersect at one point only, in which case the system has a single peak whose transmittance is theoretically unity, or the curves may never intersect at all, in which case the system will show a single peak of transmittance rather less than unity, the exact magnitude depending on the relative magnitudes of R_1 and R_2 at their closest approach. This third case is to be avoided in design. For the twin-peaked filter, a requirement is that the trough in the centre between the two peaks should be shallow, which means that R_1 and R_2 should not be very different at λ_0 .

Having examined the simplest type of DHW filter, we are in a position to study more complicated ones. What we have to look for is a system of two reflectors, where one of the reflectors remains reasonably constant over the range

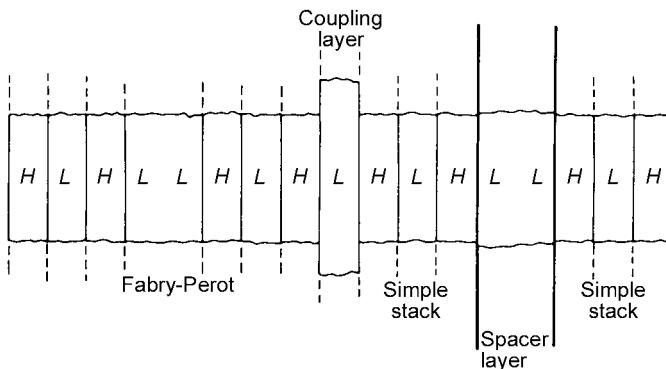


Figure 7.17. The construction of a DHW filter.

of interest and where the other should be equal, or nearly equal, to the first over the pass-band region, but should increase sharply outside the pass band. The straightforward Fabry–Perot filter has effectively zero reflectance at the peak wavelength, but the reflectance rapidly rises on either side of the peak. If, then, a simple quarter-wave stack is added to the Fabry–Perot, the resultant combination should have the desired property, that is, the reflectance equal to that of the simple stack at the centre wavelength and increasing sharply on either side. We can therefore use a simple stack as one reflector, with more or less constant reflectance, on one side of the spacer, and, on the other side, an exactly similar stack combined with a Fabry–Perot filter. This will result in a single-peaked filter since the reflectances in this way will be exactly matched at λ_0 . The double-peaked transmission curve will be obtained if the reflectance of the stack plus the Fabry–Perot filter is arranged to be just a little less than the reflectance of the stack by itself. This is the arrangement that is more often used and it involves the insertion of an extra quarter-wave layer between the stack and the Fabry–Perot. This layer appears as a sort of coupling layer in the filter. Figure 7.17 should make the situation clear.

So far we have not given any consideration to the substrate of the filter. The substrate will be on one side of the spacer and will alter the reflectance on that side. This change in reflectance can easily be calculated, particularly if the substrate is considered to be on the same side of the spacer as the simple stack. The constant reflectance R_1 of the simple stack will generally be large, and if the substrate index is given by n_s , then the transmittance of the stack on its own, $(1 - R_1)$, will become either $(1 - R_1)/n_s$ if the index of the layer next to the substrate is low, or $n_s(1 - R_1)$ if it is high.

Since this change in reflectance could be considerable, especially if n_s is large, the substrate must be taken into account in the design and this should be done right from the beginning. The substrate can be considered part of the simple

stack and R_1 can be adjusted to include it. Provided the reflectances of the two assemblies on either side of the spacer layer are arranged always to be equal at the appropriate wavelengths, the transmittance of the complete filter will be unity.

For example, let us consider the case of a filter deposited on a germanium substrate using zinc sulphide for the low-index layers and germanium for the high ones. Let the spacer be of low index and let the reflecting stack on the germanium substrate be represented by $\text{Ge}|LHLL$, where the LL layer is the spacer. The transmittance of the stack into the spacer layer will be approximately $T_1 = 4n_L^3/n_H^2 n_{\text{Ge}}$, which, since the substrate is the same material as the high-index layer, becomes $4n_L^3/n_H^3$. On the other side of the spacer layer we make a start with the combination $LLHLH|\text{air}$, representing the basic reflecting stack, where LL once again is the spacer layer. This has transmission $T_2 = 4n_L^3/n_H^4$, which is $1/n_H$ times T_1 . Clearly this is too unbalanced and an adjustment to this second stack must be made. If a low-index layer is added next to the air, then the transmission becomes $T_2 = 4n_L^5/n_H^4$. Since n_L^2 is approximately equal to the index of germanium, the transmittances T_1 and T_2 are now equal and the Fabry–Perot filter can be added to the second stack to give the desired shape to the reflectance curve. The Fabry–Perot can take any form, but it is convenient here to use a combination almost exactly the same as the combination of two stacks and a spacer layer which has already been arrived at. The complete design of the filter is then:

$$\text{Ge}|LH LL HLH L H L H LL H L H H L H|\text{air}$$

and the performance of the filter is shown in figure 7.18.

An alternative way of checking whether or not the filter is going to have high transmission uses the concept of absentee half-wave layers. The layers in DHW filters are usually either of quarter- or half-wave thickness at the centre of the pass band, as in the above filter, and we can take it as an example to illustrate the method. First we note that the two spacers are both half-wave layers and that they can be eliminated without affecting the transmission. The filter, at the centre wavelength, will have the same transmittance as

$$\text{Ge}|LH H L H L H L H H L H|\text{air}.$$

In this there are two sets of HH layers which can be eliminated in the same way, leaving two sets of LL layers which can be removed in their turn. Almost all the layers in the filter can be eliminated in this way leaving ultimately

$$\text{Ge}|L|\text{air}.$$

As we already know, a single quarter-wave of zinc sulphide is a good antireflection coating for germanium, and so the transmittance of the filter will be high in the centre of the pass band. Any type of DHW filter can be dealt with in this way.

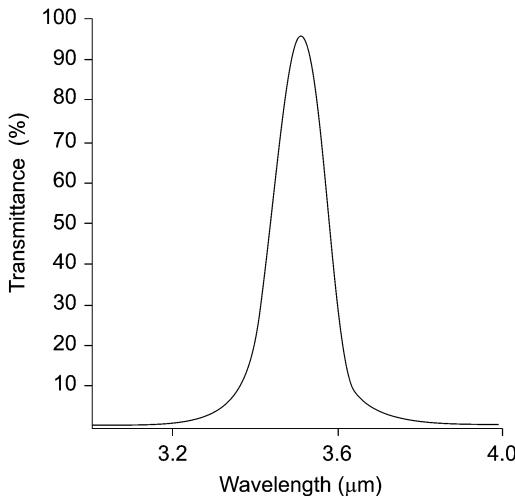


Figure 7.18. Computed transmittance of the double half-wave filter. Design: Air|LHLLHLHLLHLLH|Ge. The substrate is germanium ($n = 4.0$); H = germanium ($n = 4.0$), L = zinc sulphide ($n = 2.35$) and the incident medium is air ($n = 1.0$).

Knittl [19, 20] has used an alternative multiple beam approach to study the design of DHW filters. Basically he has applied a multiple beam summation to the first cavity, the results of which are then used in a multiple beam summation for the second cavity. This yields an expression which is not unlike Smith's, although slightly more complicated, but which has the advantage that it is only the phase which varies across the pass band. The magnitude of the reflection and transmission coefficients can be safely assumed constant and this means that the parameters which involve these quantities are also constant. The form of the expression for overall transmittance is then very much easier to manipulate so that the positions and values of maxima and minima in the pass band can be readily determined. We shall not deal further with the method here, because it is already well covered by Knittl [20].

Of course, the possible range of designs does not end with the DHW filter. Other types of filter exist involving even more half-waves. An early type of filter, which we have already mentioned, was the WADI which was devised by Turner and which consisted of a straightforward Fabry-Perot filter, to either side of which was added a half-wave layer together with several quarter-wave layers. The function of these extra layers was to alter the phase characteristics of the reflectors on either side of the primary spacer layer, so that the pass band was broadened and at the same time the sides became steeper. Similarly, it is possible to repeat the basic Fabry-Perot element used in the DHW filter once more to give a triple half-wave or THW filter, which has a similar bandwidth but steeper sides. WADI and

THW filters are much the same thing, although the original design philosophy was a little different, and usually the term THW is taken as referring to all types having three half-wave spacers. Even more spacer layers may be used giving multiple half-wave filters. The method which we have been using for the analysis of the filters becomes rather cumbersome when many half-waves are involved—even the simple method for checking that the transmittance is high in the pass band breaks down, for reasons which will be made clear in the next section, where we shall consider a very powerful design method which has been devised by Thelen.

7.3.1 Thelen's method of analysis

We have not yet arrived at any ready way of calculating the bandwidth of DHW and THW filters. The design method has merely ensured that the transmittance of the filter is high in the pass band and that the shape of the transmission curve is steep-sided. The bandwidth can be calculated, but to arrive at a prescribed bandwidth in the design has to be achieved by trial and error. It can indeed be calculated using the formula for transmittance

$$T = T_0 \frac{1}{1 + F_0 \sin^2 \delta}$$

but this can be very laborious as the phases of the reflectances have to be included in δ . This expression has been very useful in achieving an insight into the basic properties of the multiple half-wave filter, but, for systematic design, a method based on the concept of equivalent admittance will be found much more useful.

As was shown in chapter 6, any symmetrical assembly of thin films can be replaced by a single layer of equivalent admittance and optical thickness which both vary with wavelength, but which can be calculated. This concept has been used by Thelen [21] in the development of a very powerful systematic design method which predicts all the performance features of the filters including the bandwidth. The basis of the method is the splitting of the multiple half-wave filter into a series of symmetrical periods, the properties of which can be predicted by finding the equivalent admittance. Take for example the design we have already examined.

$$\text{Ge}|LHLLHLHLHLLHLH|air.$$

This can be split up into the arrangement

$$\text{Ge}|LHL \quad LHLHLHLH \quad LHLH|air.$$

The part of the filter which determines its properties is the central section $LHLHLHLH$ which is a symmetrical assembly. It can therefore be replaced by a single layer having the usual series of high-reflectance zones where the admittance is imaginary, and pass zones where the admittance is real. We are interested in the latter because they represent the pass bands of the final filter. The

symmetrical section must then be matched to the substrate and the surrounding air, and matching layers are added for that purpose on either side. This is the function of the remaining layers of the filter. The condition for perfect matching is easily established because the layers are all of quarter-wave optical thicknesses.

A most useful feature of this design approach is that the central section of the filter can be repeated many times, steepening the edges of the pass band and improving the rejection without affecting the bandwidth to any great extent.

In order to make predictions of performance straightforward, Thelen has computed formulae for the bandwidth of the basic sections. We use Thelen's technique here, with some slight modifications, in order to fit in with the pattern of analysis already carried out for the Fabry-Perot. In order to include filters of order higher than the first, we write the basic period as

$$H^m L H L H L H \dots L H^m \quad \text{or} \quad L^m H L H L H L \dots H L^m$$

where there are $2x + 1$ layers, $x + 1$ of the outermost index and x of the other, and m is the order number. We have already mentioned how Seeley [4], in the course of developing expressions for the Fabry-Perot filter, arrived at an approximate formula for the product of the characteristic matrices of quarter-wave layers of alternating high and low indices. Using an approach similar to Seeley's, we can put the characteristic matrix of a quarter-wave layer in the form:

$$\begin{bmatrix} -\varepsilon & i/n \\ in & -\varepsilon \end{bmatrix} \quad (7.52)$$

where $\varepsilon = (\pi/2)(g - 1)$ and $g = \lambda_0/\lambda$. This expression is valid for wavelengths close to that for which the layer is a quarter-wave. First let us consider m odd, and write m as $2q + 1$. Then, to the same degree of approximation, the matrix for H^m or L^m is

$$(-1)^q \begin{bmatrix} -m\varepsilon & i/n \\ in & -m\varepsilon \end{bmatrix}.$$

Neglecting terms of second and higher order in ε , then the product of the $2x - 1$ layers making up the symmetrical period is

$$\begin{bmatrix} M_{11} & iM_{12} \\ iM_{21} & M_{22} \end{bmatrix} \quad (7.53)$$

where

$$M_{11} = M_{22} = (-1)^{x+2q}(-\varepsilon) \left[m \left(\frac{n_1}{n_2} \right)^x + \left(\frac{n_1}{n_2} \right)^{x+1} + \left(\frac{n_1}{n_2} \right)^{x-2} + \dots + \left(\frac{n_2}{n_1} \right)^{x-1} + m \left(\frac{n_2}{n_1} \right)^x \right]$$

$$iM_{12} = i(-1)^x / [(n_1/n_2)^x n_1]$$

and

$$iM_{21} = i(-1)^x [(n_1/n_2)^x n_1].$$

Now it is not easy from this expression to derive the halfwidth of the final filter analytically. Instead of deriving the halfwidth, therefore, Thelen chose to define the edges of the pass band as those wavelengths for which

$$\frac{1}{2} \left| M_{11} + M_{22} \right| = 1$$

or, since $M_{11} = M_{22}$,

$$\left| M_{11} \right| = 1.$$

These points will not be too far removed from the half peak transmission points, especially if the sides of the pass band are steep. Applying this to equation (7.53), we obtain

$$\left| M_{11} \right| = \varepsilon \left[m \left(\frac{n_1}{n_2} \right)^x + \left(\frac{n_1}{n_2} \right)^{x-1} + \dots + \left(\frac{n_2}{n_1} \right)^{x-1} + m \left(\frac{n_2}{n_1} \right)^x \right]. \quad (7.54)$$

Now, this expression is quite symmetrical in terms of n_1 and n_2 . Then if we replace n_1 and n_2 by n_H and n_L , regardless of which is which, we will obtain the same expression

$$\varepsilon \left[m \left(\frac{n_H}{n_L} \right)^x + \left(\frac{n_H}{n_L} \right)^{x-1} + \left(\frac{n_H}{n_L} \right)^{x-2} + \dots + \left(\frac{n_L}{n_H} \right)^{x-1} + m \left(\frac{n_L}{n_H} \right)^x \right] = 1$$

i.e.

$$\varepsilon \left[(m-1) \left(\frac{n_H}{n_L} \right)^x + (m-1) \left(\frac{n_L}{n_H} \right)^x + \left(\frac{n_H}{n_L} \right)^x \left(\frac{1 - (n_L/n_H)^{x+1}}{1 - (n_L/n_H)} \right) \right] = 1$$

where we have used the formula for the sum of a geometric series just as in the case of the Fabry–Perot. We now neglect terms of power x or higher in (n_L/n_H) to give

$$\varepsilon \left(\frac{n_H}{n_L} \right)^x \left((m-1) + \frac{1}{1 - (n_L/n_H)} \right) = 1$$

i.e.

$$\varepsilon = \left(\frac{n_L}{n_H} \right)^x \frac{[1 - (n_L/n_H)]}{[m - (m-1)(n_L/n_H)]}. \quad (7.55)$$

The bandwidth will be given by

$$\left| \frac{\Delta\lambda_B}{\lambda_0} \right| = \left| \frac{\Delta\nu_B}{\nu_0} \right| = 2(g-1) = \frac{4\varepsilon}{\pi}$$

so that, manipulating equation (7.55) slightly,

$$\left| \frac{\Delta\lambda_B}{\lambda_0} \right| = \frac{4}{m\pi} \left(\frac{n_L}{n_H} \right)^x \frac{(n_H - n_L)}{(n_H - n_L + n_L/m)}. \quad (7.56)$$

The equivalent admittance is given by

$$\eta_E = \left(\frac{M_{21}}{M_{12}} \right)^{1/2} = \left(\frac{n_1}{n_2} \right)^x n_1. \quad (7.57)$$

The case of m even, i.e. $m = 2q$, is arrived at similarly. Here the matrix of H^m or L^m is

$$(-1)^q \begin{bmatrix} 1 & im\varepsilon/n \\ im\varepsilon n & 1 \end{bmatrix}$$

and a similar multiplication, neglecting terms higher than first in ε gives

$$\frac{\Delta\lambda_B}{\lambda_0} = \frac{4}{m\pi} \left(\frac{n_L}{n_H} \right)^x \frac{(n_H - n_L)}{(n_H - n_L + n_L/m)}$$

that is, exactly as equation (7.56), but

$$\eta_E = \left(\frac{n_2}{n_1} \right)^{x-1} n_2 \quad (7.58)$$

for equivalent admittance. This is to be expected since the layers L^m or H^m act as absentees because of the even value of m .

Expression (7.56) should be compared with the Fabry–Perot expressions (7.22) and (7.23). If we consider multiple cavity filters to be a series of Fabry–Perot cavities then the number of layers in each reflector is half that in the basic symmetrical period. Equations (7.22), (7.23) and (7.58) are, therefore, consistent.

In order to complete the design we need to match the basic period to the substrate and the surrounding medium. We first consider the case of first-order filters and the modifications which have to be made in the case of higher order will become obvious. For a first-order filter, then, matching will best be achieved by adding a number of quarter-wave layers to the period. The first layer should have index n_1 , the next n_2 and so on, alternating the indices in the usual manner. The equivalent admittance of the combination of symmetrical period and matching layers will then be

$$\frac{n_1^{2y}}{n_2^{2(y-1)}} \left(\frac{n_2}{n_1} \right)^x \frac{1}{n_2} \quad \text{or} \quad \left(\frac{n_2}{n_1} \right)^{2y} n_2 \left(\frac{n_1}{n_2} \right)^x \quad (7.59)$$

where there are y layers of index n_1 and either $(y - 1)$ or y layers of index n_2 respectively. We have also used the fact that the addition of a quarter-wave of

index n to an assembly of equivalent admittance E alters the admittance of the structure to n^2/E .

This equivalent admittance should be made equal to the index of the substrate on the appropriate side, and to the index of the surrounding medium on the other. The following discussion should make the method clear.

When we try to apply this formula to the design of multiple half-wave filters, we find to our surprise that quite a number of designs which we have looked at previously, and which seemed satisfactory, do not satisfy the conditions. For example, let us consider the design arrived at in the earlier part of this section:

$$\text{Ge}|LH\ LL\ HLH\ L\ H\ LH\ LL\ H\ LH|\text{air}$$

where L indicates zinc sulphide of index 2.35 and H germanium of index 4.0. The central period is $LHLHLHLH$, which has equivalent admittance n_L^5/n_H^4 . The LHL combination alters this equivalent admittance to

$$\frac{n_L^4}{n_L^2} \frac{n_H^4}{n_L^5} = \frac{n_H^2}{n_L}$$

which is a gross mismatch to the germanium substrate. The $LHLH$ combination on the other side alters the admittance to

$$\frac{n_H^4}{n_L^4} \frac{n_L^5}{n_H^4} = n_L$$

which in turn is not a particularly good match to air.

The explanation of this apparent paradox is that in this particular case the total filter, taking the phase thickness of the central symmetrical period into account, has unity transmittance because it satisfies Smith's conditions given in the previous section, but, over a wide range of wavelengths, pronounced transmission fringes would be seen if the bandwidth of the filter were not much narrower than a single fringe. Adding extra periods to the central symmetrical one has the effect of decreasing this fringe width, bringing them closer together. Eventually, given enough symmetrical periods, the width of the fringes becomes less than the filter bandwidth and they appear as a pronounced ripple superimposed on the pass band. This is illustrated clearly in figure 7.19. The triple half-wave version is still acceptable when an extra L layer is added, but this quintuple half-wave version is quite unusable. The presence or absence of an outermost L layer has no effect on the performance, other than inverting the fringes. The simple method of cancelling out half-waves for predicting the pass-band transmission therefore breaks down, because it merely ensures that λ_0 will coincide with a fringe peak.

It is profitable to look at the possible combinations of the two materials which can be made into a filter on germanium and where the centre section can be repeated as many times as required. The combinations for up to 11 layers in the centre section are given in table 7.3.

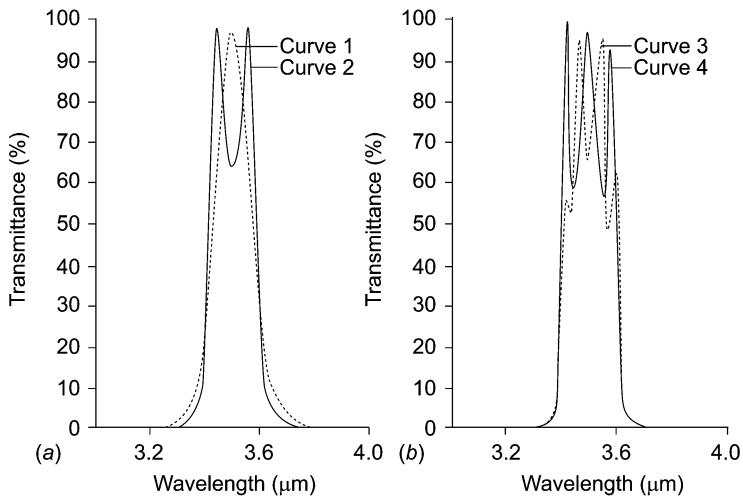


Figure 7.19. (a) Curve 1: Computed transmittance of the triple half-wave filter: Air| $LHLHL(LHLHLHLH)^2 LHL|Ge$. Curve 2: shows the effect of omitting the L layer next to the air in the design of curve 1: Air| $HLHL(LHLHLHLH)^2 LHL|Ge$. (b) Computed transmittance of quintuple half-wave filters. Curve 3: Air| $HLHL(LHLHLHLH)^4 LHL|Ge$. Curve 4: As curve 3 but with an extra L layer: Air| $LHLHL(LHLHLHLH)^4 LHL|Ge$. The presence or absence of the L layer has little effect on the ripple in the pass band. For all curves, H = germanium ($n_H = 4.0$) and L = zinc sulphide ($n_L = 2.35$).

Table 7.3.

Matching combination for germanium	Symmetrical period	Matching combination for air
Ge L	LHL	air (already matched)
Ge LH	$HLHLH$	$H air$
Ge LHL	$LHLHLHL$	$LH air$
Ge $LHLH$	$HLHLHLHLH$	$HLH air$
Ge $LHLHL$	$LHLHLHLHLH$	$LHLH air$

L : ZnS, $n_L = 2.35$ H : Ge, $n_H = 4.0$.

The validity of any of these combinations can easily be tested. Take for example the fourth one, with the nine-layer period in the centre. Here the equivalent admittance of the symmetrical period is $E = n_H^5/n_L^4$. The $LHLH$ section between the germanium substrate and the centre section transforms the

admittance into

$$\frac{n_L^4}{n_H^4} \frac{n_H^5}{n_L^4} = n_H$$

which is a perfect match for germanium. The matching section at the other end is $H L H$ and this transforms the admittance into

$$\frac{n_H^4}{n_L^2} \frac{n_L^4}{n_H^5} = \frac{n_L^2}{n_H}$$

which, because zinc sulphide is a good antireflection material for germanium, gives a good match for air.

For higher-order filters, the method of designing the matching layers is similar. However, we can choose, if we wish, to add half-wave layers to that part of the matching assembly next to the symmetrical period in order to make the resulting cavity of the same order as the others. For example, the period $H H H L H L H L H L H H H$, based on the fourth example of table 7.3, can be matched either by $Ge|L H L H$ and $H L H|air$, as shown, or by $Ge|L H L H H H$ and $H H H L H|air$, making all cavities of identical order regardless of the number of periods.

This method, then, gives the information necessary for the design of multiple half-wave filters. The edge steepness and rejection in the stop bands will determine the number of basic symmetrical periods in any particular case. Usually, because of the approximations which have been used in establishing the various formulae, and also because the definition used for bandwidth is not necessarily the halfwidth, although it would not be too far removed from it, it is advisable to check the design by accurate computation before actually manufacturing the filter. It may also be advisable to make an estimate of the permissible errors which can be tolerated in the manufacture because it is pointless attempting to achieve a performance beyond the capabilities of the process. The result will just be worse than if a less demanding specification had been attempted. The estimation of manufacturing errors is a subject which has not received much attention in the literature on thin-film filters. A brief discussion of permissible errors is given in chapter 11, pp 535–44, with some examples of calculations applied to multiple half-wave filters. Typical multiple half-wave filters are shown in figure 7.20.

7.4 Higher performance in multiple cavity filters

The curve of figure 7.20(b) shows the square shape of the pass band of a multiple cavity filter but also illustrates one of the problems inherent in this type of design, the ‘rabbit’s ears’, or the rather prominent peaks at either side of the pass band. This can become even worse with increasing numbers of periods. Figure 7.21 shows this clearly.

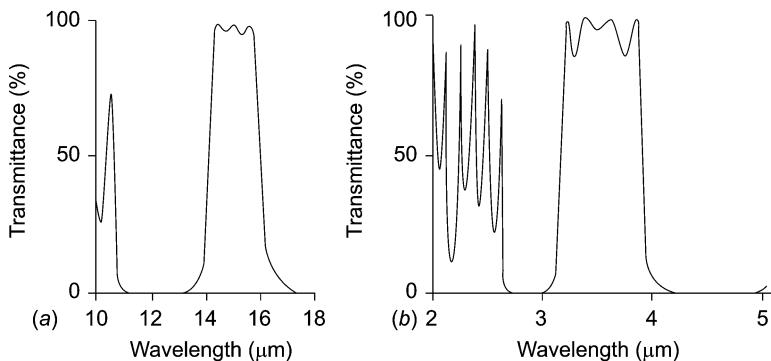


Figure 7.20. (a) Transmittance of a multiple half-wave filter. Design: Air|HHLHLLHLHLLH|Ge with $H = \text{PbTe}$ ($n = 5.0$); $L = \text{ZnS}$ ($n = 2.35$), $\lambda_0 = 15 \mu\text{m}$. (b) Transmittance of a multiple half-wave filter. Design: Air|HHLHLLHLHLLHLLHLLHLLH|silica $H = \text{Ge}$ ($n = 4.0$); $L = \text{ZnS}$ ($n = 2.35$); silica substrate ($n = 1.45$) $\lambda_0 = 3.5 \mu\text{m}$. (Courtesy of Sir Howard Grubb, Parsons & Co. Ltd.)

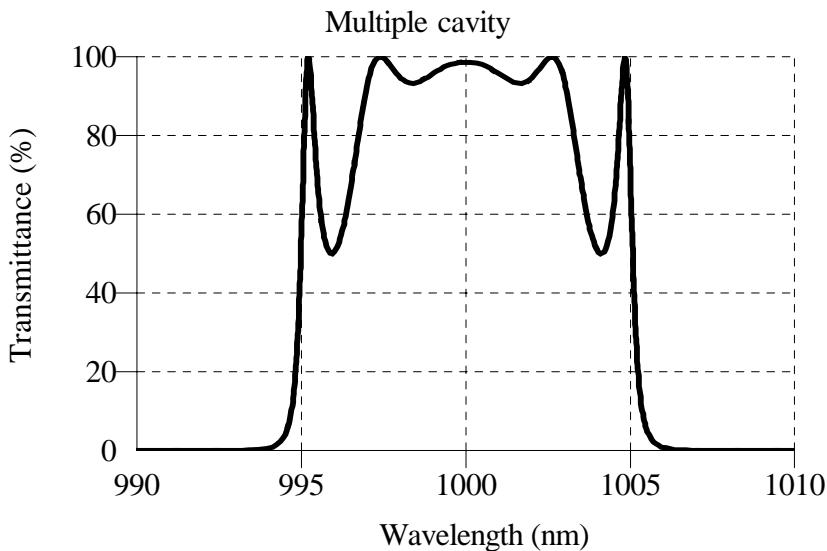


Figure 7.21. A multiple cavity filter with a central core of five symmetrical periods. Design: Glass|HLLHLHLH (HLLHLHLHLHLHLH)⁵ HLLHLHLH|Glass, with $y_H = 2.35$, $y_L = 1.35$, $y_{\text{glass}} = 1.52$, $\lambda_0 = 1000 \text{ nm}$. Note the very prominent peaks at the edges of the pass band sometimes called 'rabbit's ears'.

The reason for this problem feature is the dispersion of the equivalent admittance of the symmetrical period. In the design approach, this is assumed to be a constant across the pass band but in reality it varies considerably, tending to either zero or infinity at the pass-band edges; see figure 7.22. It is, in fact, exactly the same problem as in edge filters where better ripple suppression near the edge demands a matching system that exhibits similar dispersion. Shifted periods are, however, difficult to arrange in the case of band-pass filters because of the need for ripple suppression at both edges of the pass band. However, inspired by the shifted periods technique, we seek a solution, where part of the matching is due to a symmetrical system that has a dispersion of the appropriate form so that its matching remains reasonably good even when the equivalent admittance to be matched to the surrounding media is varying. Any of the symmetrical periods we are dealing with will have an odd number of quarter-waves so that the equivalent phase thickness at $g = 1$ will be an odd number of $\pi/2$. This implies that the period could, itself, be used as a simple matching assembly. Since the pass band in this type of filter is usually narrow, the matching condition will not vary too much over the pass-band width. In order to make use of this possibility, we have to find at least pairs of symmetrical periods that will permit one to be used as a matching assembly for the other. Attempting to find two, or more, periods that have the correct relationship at $g = 1$ for one to match the other to the substrate or incident medium is difficult. If we could find two periods of different width but with the same central admittance, then we could continue to use the straightforward matching illustrated in figure 7.3 which uses a series of quarter-wave layers and is perfectly satisfactory at the centre of the pass band. A solution lies with higher-order periods.

The addition of further half-wave layers to the outside of a symmetrical period does not change its equivalent admittance at the pass-band centre, nor does it change the sense of curvature of the variation of equivalent admittance. Figure 7.23 shows the admittances of *HLHLHLHLHLH*, *HHHLHLHLHLHHH*, *HLLLHLHLHLHLLH* and *HHHLLLH LHLHLHLLLHHH*. All have the value y_H^6/y_L^5 at $g = 1$ and all exhibit a gradually increasing admittance as the value of g moves away from unity. The wider curves have values of admittance intermediate between the narrower curves and the value that all possess at $g = 1$. All represent an odd number of quarter-waves at $g = 1$ but the broader curves remain closer to an odd number of quarter-waves than the narrower as g varies. They could therefore be used to match the narrower ones to a notional medium of constant admittance, y_H^6/y_L^5 . The best one, that is the period that is closest to the ideal values of the required admittance, is chosen. The use of more than one of the wider matching systems does not give very good results because of their differing dispersion curves. Matching of the dispersionless notional medium to the incident and emergent media is then a straightforward matter of a series of quarter-waves, as before.

A simple example uses two of the periods from figure 7.23, *HLHLHLH*

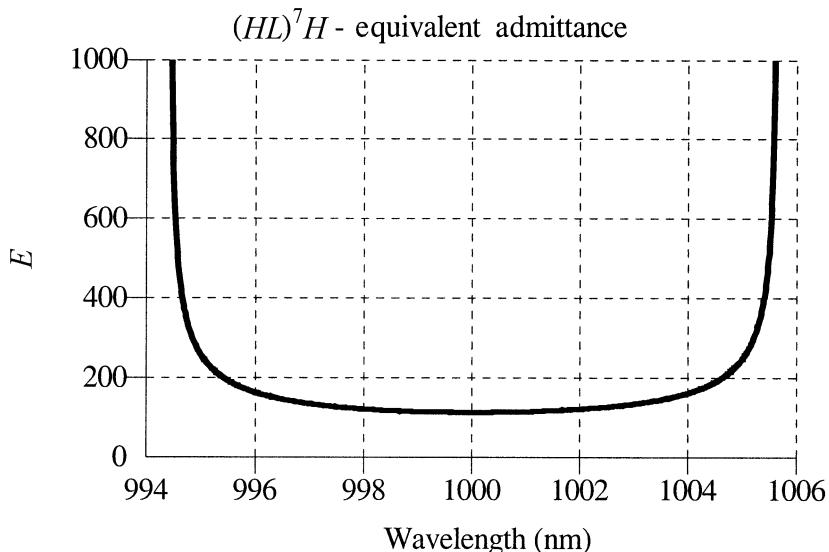


Figure 7.22. The equivalent admittance of $(HL)^7H$ over the potential pass band. Note the rapid change near the edges of the pass band. This dispersion of equivalent admittance is very difficult to match to an essentially dispersionless medium.

LHLH and HHHLHLHLHLHLHH:

$$\text{Glass} | H L H L H H L H L H L H L H (H H H L H L H L H L H L H H H)^q \\ H L H L H L H L H L H L H L H | \text{Glass}.$$

The characteristic curves of two such filters are shown in figures 7.24 and 7.25.

We need an expression for the width of such filters. This is determined principally by the highest order periods. If we write the expression for the highest order period as:

$$mABABA \dots BAmA$$

where there are $2x + 1$ layers including the layers mA , then we can show that the bandwidth, defined in the same way as before, is given by:

$$\frac{\Delta\lambda}{\lambda_0} = \frac{4}{m\pi} \left(\frac{y_L}{y_H} \right)^x \frac{(y_H - y_L)}{(y_H - y_L + y_L/m)}. \quad (7.60)$$

This expression reduces to that already derived if $m = 1$. Using the expression to calculate the bandwidth of the filters of figures 7.24 and 7.25, we find 0.018, implying pass-band edges at 991.1 nm and 1009.1 nm.

Designs arrived at in this way will be satisfactory for a wide range of applications where ripple within the pass band must be small. However, there are

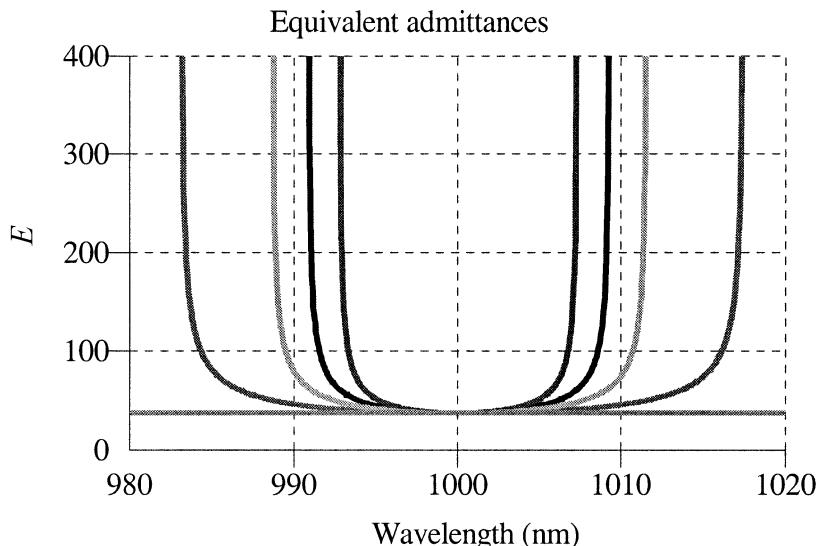


Figure 7.23. The equivalent admittances of symmetrical periods from narrower to broader in order $HHHLLLHLHLLHLLHHH$, $HHHLHLHLLHLLHHH$, $HLLLHLHLLHLLH$ and $HLHLHLHLLHLH$ with H representing characteristic admittance 2.35 and L 1.35. The straight line represents the admittance that all have at $g = 1$.

applications where even the performance in figures 7.24 and 7.25 is inadequate. There are requirements in dense wavelength division multiplexing for peak transmittances in excess of 99%, for example. A useful technique that is somewhat empirical uses additional matching layers. In the following filter H and L indicate admittances of 2.35 and 1.35, and where the substrate is glass of admittance 1.52 and the incident medium is air of admittance 1.00. These correspond to the materials we have been using so far and we prefer not to change at this stage. Recently much use has been made of dense silica and tantalum in the manufacture of narrowband filters, particularly for wavelength division multiplexing, but the design techniques are similar. The filter we use as an example is given by:

$$\text{Air}|L(HL)^3 H (HL)^7 H (HH(HL)^7 HHH)^2 (HL)^7 H H(LH)^3|\text{Glass}.$$

The performance is shown in figure 7.26. The filter is matched to an incident medium of air rather than the glass we have been using and so there is an extra L layer next to the incident medium.

The loss is purely a reflection loss. No absorption is involved. Thus it should be possible by correct matching to reduce the loss to zero and increase the transmittance to 100%. However, we need to accomplish this in as simple a

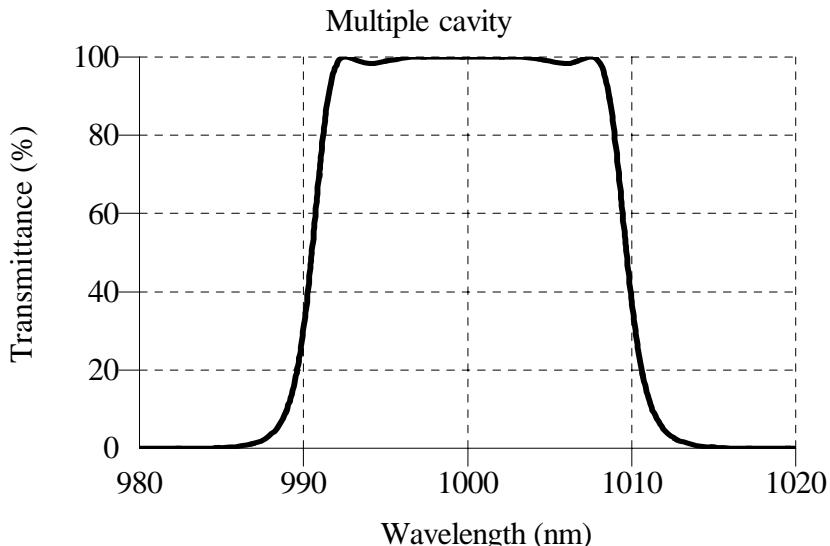


Figure 7.24. A multiple cavity filter similar to that of figure 7.21 but using periods of increasing order to improve the pass-band ripple. Design: Glass| $LHHLH\,HLHLHLHLH(HHHLHLHLHLHH)^2\,HLHLHL$
 $HLHLH\,HLHLH|$ Glass with $y_H = 2.35$, $y_L = 1.35$, $y_{\text{glass}} = 1.52$, $\lambda_0 = 700 \text{ nm}$. Note the much flatter pass-band top compared with figure 7.21.

fashion as possible. We take as our target, therefore, to increase the transmittance so that it is greater than 99% over the entire pass-band top. An analytical approach is unlikely to be profitable and so we use some logic and then rely on automatic methods.

The filter structure is thick and complicated and the matching must be capable of accommodating considerable dispersion of the admittance of the assembly. This implies that a very thin system of layers is unlikely to be of much value. We therefore assume from the start that the matching layer will be fairly thick.

We try two different starting designs for the matching layer, a three-layer LHL and a five-layer $LHLHL$ arrangement to replace the single L matching layer of the original design (the layer next to the air). However, we find single quarter-wave thicknesses insufficient for a good match and we need to make the layer thicker. Some trial and error finds preferred starting designs of $18L18H18L$ and $12L12H12L12H12L$ although the final result is not very sensitive to the exact starting design thicknesses. Some gentle refinement with only the matching layers taking part then yields final matching systems as shown in tables 7.4 and 7.5 and filter characteristics in figures 7.27, 7.29 and 7.30. The admittance locus of the three-layer matching system is shown in figure 7.28.

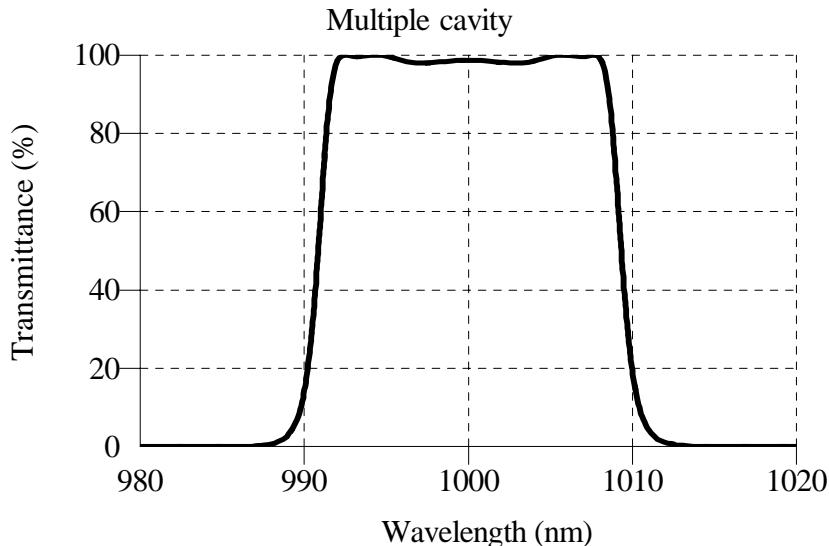


Figure 7.25. A multiple cavity filter similar to that of figure 7.24 but with three central high-order periods rather than two. Design: Glass |HLHLHHLHLHLHLH(HHHLHLHLHLH₃)³ HLHLHL HLHLHHLHLH| Glass with $y_H = 2.35$, $y_L = 1.35$, $y_{\text{glass}} = 1.52$, $\lambda_0 = 700 \text{ nm}$.

Table 7.4.

Three-layer system	
Index	Optical thickness (full waves)
Air	Incident medium
1.3500	4.6377
2.3500	4.4624
1.3500	4.7197
Filter structure	

It is very difficult to carry out this type of design in a completely systematic way. There are other techniques but the quarter-wave thicknesses of the basic filter design help considerably in the thickness control of the deposition process. The final matching layers are much thicker than quarter-waves and are the final layers of the structure and so their monitoring signals are also quite favourable.

The matching can also be conveniently placed between the multiple cavity

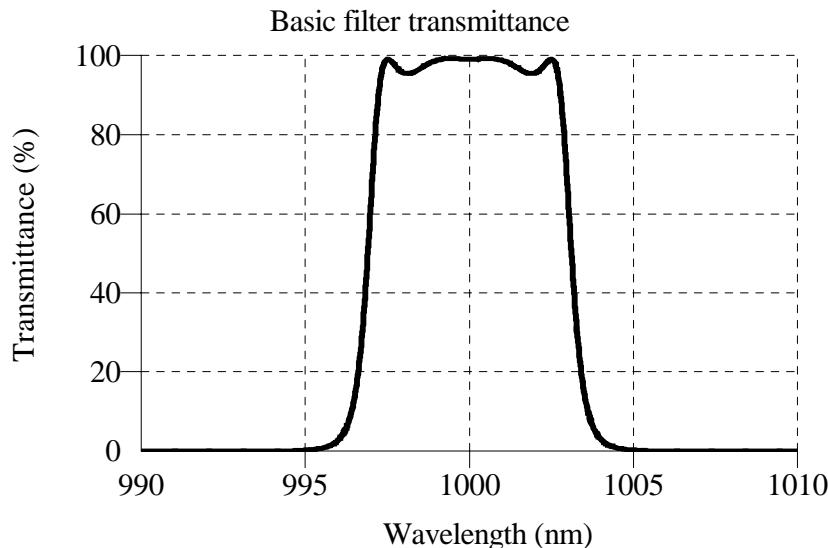


Figure 7.26. Transmittance of filter: Air| $L(HL)^3 H(HL)^7 H(HH(HL)^7 HHH)^2 (HL)^7 H H(LH)^3$ |Glass.

Table 7.5.

Five-layer system	
Index	Optical thickness (full waves)
Air	Incident medium
1.3500	3.0727
2.3500	2.9991
1.3500	3.0674
2.3500	2.9651
1.3500	3.1973

Filter structure	
------------------	--

structure and the substrate. Automatic refinement of the matching layers is again the preferred technique for arriving most easily at the thicknesses required for the layers.

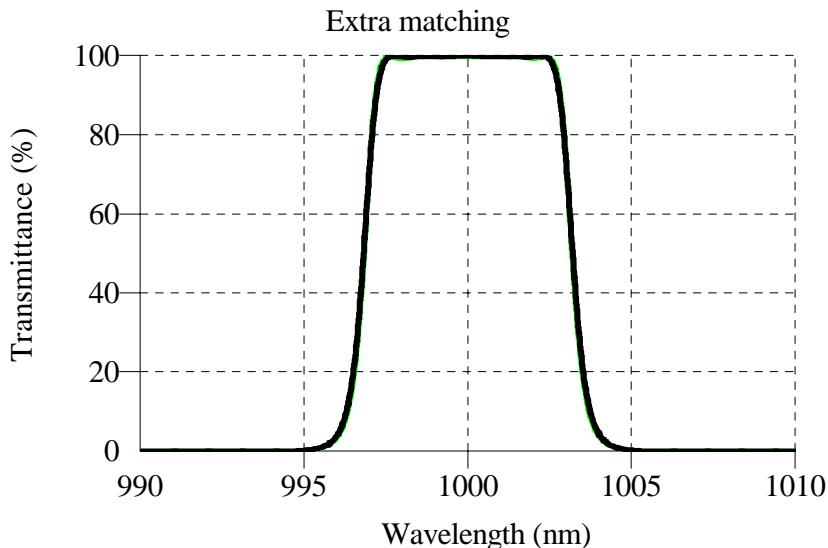


Figure 7.27. The two curves when the additional matching three- and five-layer systems are added. The five-layer system is superimposed over the three-layer, which is almost invisible in the scale of the figure. An expanded transmittance is shown in figure 7.29.

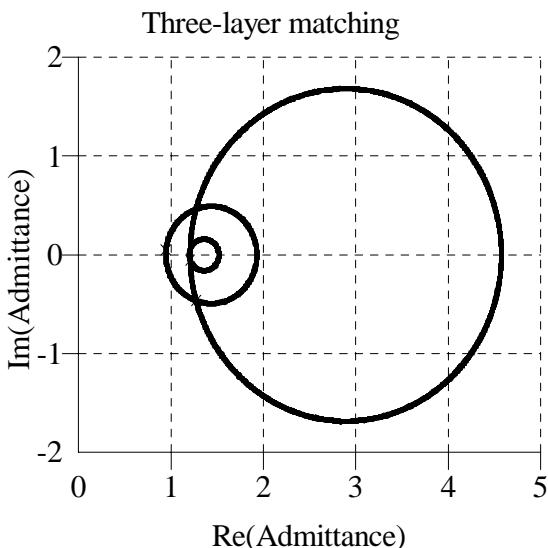


Figure 7.28. The admittance locus of the three-layer matching system plotted at the centre wavelength of the filter.

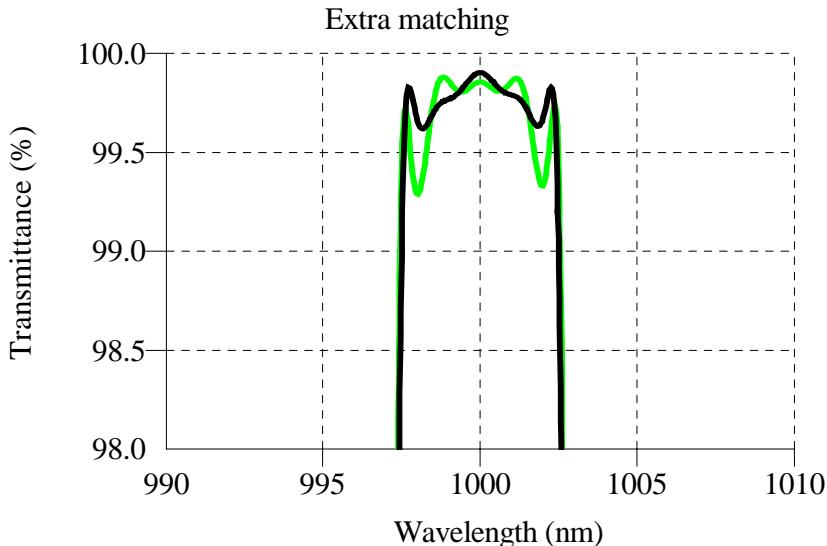


Figure 7.29. The two curves of figure 7.27 plotted on an expanded scale to show the differences. The three-layer system (lighter curve) is slightly inferior to the five-layer system.

7.4.1 Effect of tilting

A feature of the design not so far mentioned is the sensitivity to changes in angle of incidence. Thelen [21] has examined this aspect and for those types which involve symmetrical periods consisting of quarter-waves of alternating high and low index and where the spacers are of the first order, he arrived at exactly the same expressions as those of Pidgeon and Smith for the Fabry–Perot. As far as angular dependence is concerned, the filter behaves as if it were a single layer with an effective index of

$$n^* = (n_1 n_2)^{1/2}$$

where $n_1 > n_2$ or

$$n^* = \frac{n_1}{[1 - (n_1/n_2) + (n_1/n_2)^2]^{1/2}}$$

where $n_2 > n_1$.

For higher-order filters, therefore, we should be safe in making use of expressions (7.33) and (7.35).

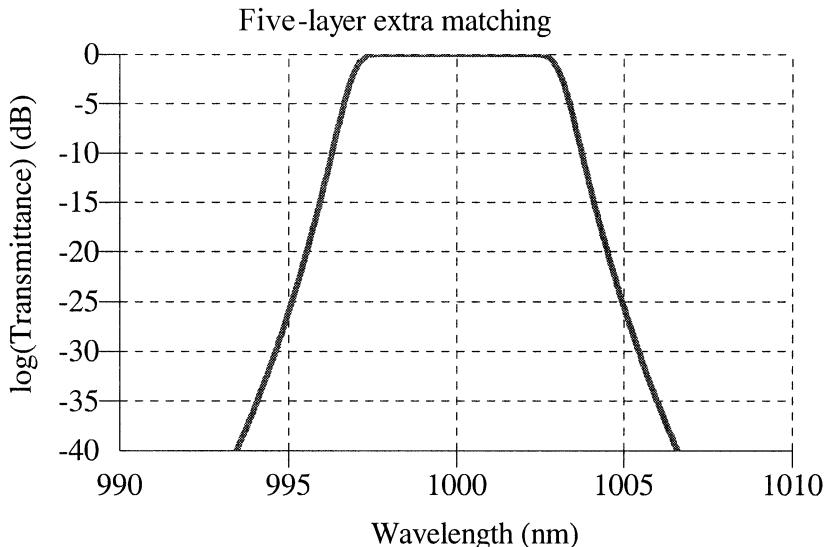


Figure 7.30. The performance of the filter with five-layer matching from figures 7.27 and 7.29 shown on a logarithmic scale.

7.4.2 Losses in multiple cavity filters

Losses in multiple cavity filters can be estimated in the same way as for the Fabry–Perot filter. There are so many possible designs that a completely general approach would be very involved. However, we can begin by assuming that the basic symmetrical unit is perfectly matched at either end. The scheme of admittances through the basic unit will then be as shown in table 7.6.

Then, in the same way as for the Fabry–Perot, we can write

$$\begin{aligned}
 \sum \mathcal{A} &= \beta_1 \left[\left(\frac{n_1}{n_2} \right)^{x-1} + \left(\frac{n_2}{n_1} \right)^{x-1} \right] + \beta_2 \left[\left(\frac{n_2}{n_1} \right)^{x-2} + \left(\frac{n_1}{n_2} \right)^{x-2} \right] \\
 &\quad + \beta_1 \left[\left(\frac{n_1}{n_2} \right)^{x-3} + \left(\frac{n_2}{n_1} \right)^{x-3} \right] + \dots + \beta_2 \left[\left(\frac{n_2}{n_1} \right)^{x-2} + \left(\frac{n_1}{n_2} \right)^{x-2} \right] \\
 &\quad + \beta_1 \left[\left(\frac{n_1}{n_2} \right)^{x-1} + \left(\frac{n_2}{n_1} \right)^{x-1} \right] \\
 &= \beta_1 \left\{ \left[\left(\frac{n_1}{n_2} \right)^{x-1} + \left(\frac{n_1}{n_2} \right)^{x-3} + \dots + \left(\frac{n_2}{n_1} \right)^{x-1} \right] \right. \\
 &\quad \left. + \left[\left(\frac{n_2}{n_1} \right)^{x-1} + \left(\frac{n_2}{n_1} \right)^{x-2} + \dots + \left(\frac{n_1}{n_2} \right)^{x-1} \right] \right\}
 \end{aligned}$$

Table 7.6.

n_1	n_1^x/n_2^{x-1}
n_2	n_2^{x-1}/n_1^{x-2}
n_1	n_1^{x-2}/n_2^{x-3}
n_2	n_2^{x-3}/n_1^{x-4}
\vdots	n_1^{x-2}/n_2^{x-3}
n_2	n_2^{x-1}/n_1^{x-2}
n_1	n_1^x/n_2^{x-1}

x layers of n_1 .
 $(x - 1)$ layers of n_2 .

$$+ \beta_2 \left\{ \left[\left(\frac{n_2}{n_1} \right)^{x-2} + \left(\frac{n_2}{n_1} \right)^{x-4} + \dots + \left(\frac{n_1}{n_2} \right)^{x-2} \right] + \left[\left(\frac{n_1}{n_2} \right)^{x-2} + \left(\frac{n_1}{n_2} \right)^{x-4} + \dots + \left(\frac{n_2}{n_1} \right)^{x-2} \right] \right\}.$$

We note that the second expression of each pair is the same as the first with inverse order.

The layers are quarter-waves and so we can write, as before,

$$\beta_1 = \frac{\pi}{2} \frac{k_1}{n_1} \quad \text{and} \quad \beta_2 = \frac{\pi}{2} \frac{k_2}{n_2}.$$

Once again we divide the cases into high- and low-index cavities.

7.4.3 Case I: high-index cavities

We replace n_1 by n_H , k_1 by k_H , n_2 by n_L and k_2 by k_L . Then, neglecting, as before, terms in $(n_L/n_H)^x$ compared with unity,

$$\begin{aligned} \sum \mathcal{A} &= \frac{\pi(k_H/n_H)(n_H/n_L)^{x-1}}{1 - (n_L/n_H)^2} + \frac{\pi(k_L/n_L)(n_H/n_L)^{x-2}}{1 - (n_L/n_H)^2} \\ &= \pi \left(\frac{n_H}{n_L} \right)^x \frac{n_L(k_H + k_L)}{(n_H^2 - n_L^2)} \end{aligned}$$

or, using (7.56) with $m = 1$,

$$\frac{\Delta\lambda_B}{\lambda_0} = \frac{4}{\pi} \left(\frac{n_L}{n_H} \right)^x \frac{(n_H - n_L)}{n_H}$$

i.e.

$$\sum A = 4 \left(\frac{\lambda_0}{\Delta\lambda_B} \right) \frac{n_L(k_H + k_L)}{n_H(n_H + n_L)}.$$

Now, this is the loss of one basic symmetrical unit. If further basic units are added each will have the same loss. In addition, there are the matching stacks at either end of the filter. We will not be far in error if we assume that they add a further loss equal to one of the basic symmetrical units. The total number of units is then equal to the number of cavities. If we denote this by q then $q = 2$ for a two-cavity (or DHW) filter and so on. We can also assume that $R = 0$ so that the absorption loss becomes

$$A = q\pi \left(\frac{n_H}{n_L} \right)^x \frac{n_L(k_H + k_L)}{(n_H^2 - n_L^2)} \quad (7.61)$$

or

$$A = 4q \left(\frac{\lambda_0}{\Delta\lambda_B} \right) \frac{n_L(k_H + k_L)}{n_H(n_H + n_L)}. \quad (7.62)$$

7.4.4 Case II: low-index cavities

In the same way

$$A = q\pi \left(\frac{n_H}{n_L} \right)^x \frac{(n_H^2 k_L + n_L^2 k_H)}{n_H(n_H^2 - n_L^2)} \quad (7.63)$$

or

$$A = 4q \left(\frac{\lambda_0}{\Delta\lambda_B} \right) \left(\frac{n_L}{n_H} \right) \frac{[k_L(n_H/n_L) + k_H(n_L/n_H)]}{(n_H + n_L)}. \quad (7.64)$$

Expressions (7.62) and (7.64) are approximately q times the absorption of single-cavity, or Fabry–Perot, filters with the same halfwidth, a not surprising result.

7.4.5 Further information

Many of the examples of multiple cavity filters so far described have been for the infrared, but of course they can be designed for any region of the spectrum where suitable thin-film materials exist. An account of filters for the visible and ultraviolet is given by Barr [22]. All-dielectric filters, both of the Fabry–Perot and multiple cavity types for the near ultraviolet are described by Nielson and Ring [23]. They used combinations of cryolite and lead fluoride and of cryolite and antimony trioxide, the former for the region 250–320 nm and the latter for 320–400 nm. Apart from the techniques required for the deposition of these materials, the main difference between such filters and those for the infrared is that the values

of the high and low refractive indices are much closer together, requiring more layers for the same rejection. Nielson and Ring's filters contained basic units of 17 or 19 layers, in most cases, so that complete DHW filters consisted of 31 or 39 layers respectively. Malherbe [24] has described a lanthanum fluoride and magnesium fluoride filter for 205.5 nm in which the basic unit had 51 layers (high-index first-order spacer), the full design being $(HL)^{12}H\ H(LH)^{25}H(LH)^{12}$ with a total number of 99 layers, giving a measured bandwidth of 2.5 nm.

7.5 Phase dispersion filter

The phase dispersion filter represents an attempt to find an approach to the design of narrowband filters which would avoid some of the manufacturing difficulties inherent in Fabry–Perot filters. The Fabry–Perot becomes increasingly difficult to manufacture as halfwidths are reduced below 0.3% of peak wavelength. Attempts to improve the position by using higher-order spacers are not effective when the spacer becomes thicker than the fourth order because of what has been described as increased roughness of the spacer. Much more is now known about the Fabry–Perot filter and the causes of manufacturing difficulties, and those will be dealt with in some detail in a subsequent chapter. Although the phase dispersion filter was not, as it turned out, the solution to the narrowband filter problem, nevertheless it does have very interesting properties and the philosophy behind the design is worth discussing.

The reflecting stack with extended bandwidth which was originally intended for classical Fabry–Perot plates and was described in chapter 5 shows a large dispersion of the phase change on reflection and this suggested to Baumeister and Jenkins [25] that it might form the basis for a new type of filter in which the narrow bandwidth would depend almost entirely on this phase dispersion rather than on the very high reflectances of the reflecting stacks. They called this type of filter a ‘phase dispersion filter’. It consists quite simply of a Fabry–Perot all-dielectric filter which has, instead of the conventional dielectric quarter-wave stacks on either side of the spacer layer, reflectors consisting of the staggered multilayers. The rapid change in phase causes the bandwidth of the filter and the position of the peak to be much less sensitive to the errors in thickness of the spacer layer than would otherwise be the case.

The results which they themselves [25] and also with Jeppesen [26] eventually achieved were good, although they never quite succeeded in attaining the performance possible in theory. This prompted a study [27] of the influence of errors in any of the layers of a filter on the position of the peak. The idea behind this study was that random errors in both thickness and uniformity in layers other than the spacer might be responsible for the discrepancy between theory and practice. If, in a practical filter, the errors were causing the peak to vary in position over the surface of the filter, then the integrated response would exhibit a rather wider bandwidth and lower transmittance than those of any very small portion of

the filter, which might well be attaining the theoretical performance. It seemed possible that there might be a design of filter which could yield the minimum sensitivity to errors and therefore give the minimum possible bandwidth with a given layer ‘roughness’.

Giacomo *et al*'s findings [27] can be summarised as follows (the notation in their paper has been slightly altered to agree with that used throughout this book): the peak of an all-dielectric multilayer filter is given by

$$\frac{\phi_a + \phi_b}{2} - \delta = m\pi \quad (7.65)$$

where

$$\delta = \frac{2\pi n d_s}{\lambda} = 2\pi n d_s \nu$$

the symbols having their usual meanings.

For a change Δd_i in the i th layer, Δd_j in the j th layer and Δd_s in the spacer, the corresponding change in the wavenumber of the peak $\Delta\nu$ is given by

$$\sum_i \frac{\partial \phi_a}{\partial d_i} \Delta d_i + \sum_j \frac{\partial \phi_b}{\partial d_j} \Delta d_j - 2 \frac{\partial \delta}{\partial d_s} \Delta d_s + \left(\frac{\partial \phi_a}{\partial \nu} + \frac{\partial \phi_b}{\partial \nu} - 2 \frac{\partial \delta}{\partial \nu} \right) \Delta \nu = 0. \quad (7.66)$$

Now

$$\frac{\partial \delta}{\partial d_s} = 2\pi n \nu = \frac{\delta}{d_s} \quad (7.67)$$

and

$$\frac{\partial \delta}{\partial \nu} = 2\pi n d_s = \frac{\delta}{\nu} \quad (7.68)$$

and also, since d_i and ν appear in the individual thin-film matrices only in the value of $\delta_i = 2\pi n_i d_i \nu$, then

$$\sum_i \frac{\partial \phi_a}{\partial d_i} \Delta_0 d_i = \frac{\partial \phi_a}{\partial \nu} \Delta_0 \nu$$

and similarly for ϕ_b , where Δ_0 indicates that the changes in d_i are related by

$$\frac{\Delta_0 d_i}{d_i} = \frac{\Delta_0 \nu}{\nu}.$$

This gives

$$\frac{\partial \phi_a}{\partial \nu} = \sum_i \left(\frac{\partial \phi_a}{\partial d_i} \frac{d_i}{\nu} \right) \quad (7.69)$$

which is independent of the particular choice of Δ_0 used to arrive at it. A similar expression holds for ϕ_b . Using equations (7.67), (7.68) and (7.69) in

equation (7.66):

$$\sum_i \frac{\partial \phi_a}{\partial d_i} \Delta d_i + \sum_j \frac{\partial \phi_b}{\partial d_j} \Delta d_j - 2\delta \frac{\Delta d_s}{d_s} + \left(\sum_i \frac{\partial \phi_a}{\partial d_i} d_i + \sum_j \frac{\partial \phi_b}{\partial d_j} d_j - 2\delta \right) \frac{\Delta v}{v} = 0$$

i.e.

$$\begin{aligned} \frac{\Delta v}{v} = & - \left[-2\delta \alpha_s + \sum_i \left(\frac{\partial \phi_a}{\partial d_i} d_i \alpha_i \right) + \sum_j \left(\frac{\partial \phi_b}{\partial d_j} d_j \alpha_j \right) \right] \\ & \times \left[-2\delta + \sum_i \left(\frac{\partial \phi_a}{\partial d_i} d_i \right) + \sum_j \left(\frac{\partial \phi_b}{\partial d_j} d_j \right) \right]^{-1} \end{aligned} \quad (7.70)$$

where

$$\alpha_i = \frac{\Delta d_i}{d_i} \quad \text{etc.}$$

Now, in a real filter, the fluctuations in thickness, or ‘roughness’, will be completely random in character, and in order to deal with the performance of any appreciable area of the filter, we must work in terms of the mean square deviations. Each layer in the assembly can be thought of as being a combination of a large number of thin elementary layers of similar mean thicknesses but which fluctuate in a completely random manner quite independently of each other. The RMS variation in thickness of any layer in the filter can then be considered to be proportional to the square root of its thickness. This can be written:

$$\varepsilon_i = k d_i^{1/2}$$

where k can be assumed to be the same for all layers regardless of thickness. If a_i is the RMS fractional variation of the i th layer, then

$$a_i = \frac{\varepsilon_i}{d_i} = \frac{k}{d_i^{1/2}}$$

where

$$a_i^2 = \overline{\alpha_i^2}.$$

We now define β as being

$$\beta^2 = \overline{\left(\frac{\Delta v}{v} \right)^2}.$$

Then

$$\beta^2 = \left\{ 4\delta^2 a_s^2 + \sum_i \left[\left(\frac{\partial \phi_a}{\partial d_i} d_i \right)^2 a_i^2 \right] + \sum_j \left[\left(\frac{\partial \phi_b}{\partial d_j} d_j \right)^2 a_j^2 \right] \right\} \\ \times \left[-2\delta + \sum_i \left(\frac{\partial \phi_a}{\partial d_i} d_i \right) + \sum_j \left(\frac{\partial \phi_b}{\partial d_j} d_j \right) \right]^{-2}$$

which gives

$$\beta^2 = \left(k^2 \sum_{k=1}^q \frac{1}{d_k} A_k^2 \right) \left(\sum_{k=1}^q A_k \right)^{-2} \quad (7.71)$$

where

$$A_k = \frac{\partial \phi_a}{\partial d_k} d_k \quad \text{or} \quad \frac{\partial \phi_b}{\partial d_k} d_k \quad \text{or} \quad -2\delta$$

whichever is appropriate. q is the number of layers in the filter. The expression will be a minimum when

$$A_k/d_k = A_l/d_l = \dots \quad (7.72)$$

Then

$$\beta^2 = k^2/T \quad (7.73)$$

where T is the total thickness of the filter.

In the general case,

$$\beta \geq k/T^{1/2}$$

and one might hope to attain a limiting resolution of

$$R = T^{1/2}/k. \quad (7.74)$$

The condition written in equation (7.62) can be developed with the aid of equation (7.59) into

$$\frac{\partial \phi_a}{\partial d_k} = \frac{\partial \phi_b}{\partial d_l} = -4\pi n v$$

so that

$$v \left(\frac{\partial \phi_a}{\partial v} \right) = \sum_i \frac{\partial \phi_a}{\partial d_i} d_i = -4\pi n v d_m$$

and likewise for reflector b, where d_m = total thickness of the appropriate reflector and a is the index of the spacer. This gives

$$\frac{\partial \phi_a}{\partial v} = -4\pi n d_m. \quad (7.75)$$

This condition is necessary but not sufficient for the resolution to be a maximum and it can be used as a preliminary test of the suitability of any particular multilayer reflector which may be employed.

The classical quarter-wave stack is very far from satisfying it but the staggered multilayer is much more promising. In their paper, Giacomo *et al* compare a staggered multilayer reflector with a conventional quarter-wave stack. Both reflectors have 15 layers, and the results are quoted for the broadband reflector at $17\ 000\ \text{cm}^{-1}$ and for the conventional reflector at $20\ 000\ \text{cm}^{-1}$.

Equation (7.75) can be written

$$\sum_i \frac{\partial \phi_a}{\partial d_i} d_i = \sum_i \frac{\partial \phi_a}{\partial \alpha_i} = -4\pi n v d_m.$$

Now, from table 7.7,

$$-\sum_i \frac{\partial \phi_a}{\partial \alpha_i} = 30.662$$

and

$$4\pi n v d_m = 34.5$$

so that on the preliminary basis of equation (7.75) the prospects look extremely good. However, this is not a sufficient condition. We must calculate the actual relationship between β and k and compare it with the theoretical condition given by equation (7.73). Now

$$A_i = d_i \frac{\partial \phi}{\partial d_i} = \frac{\partial \phi}{\partial \alpha_i}$$

which is the last column given for each reflector. This can be used in equation (7.71) giving for a filter using the broadband reflector

$$\beta = 1.023k$$

which can be compared with the value obtained in the same way for the conventional quarter-wave stack of table 7.7:

$$\beta = 1.289k.$$

For a total filter thickness of $2.35\ \mu\text{m}$ the theoretical minimum value of β is given by (7.73) as

$$\beta = 0.652k$$

(k having units of $\mu\text{m}^{1/2}$).

Table 7.7.

Layer number	Broadband film				Classical film		
	Thickness d_i (μm)	Index n	$\partial\phi/\partial d_i$ (μm^{-1})	$\partial\phi/\partial\alpha_i$	Thickness d_i (μm)	$\partial\phi/\partial d_i$ (μm^{-1})	$\partial\phi/\partial\alpha_i$
Substrate	—	1.52	—	—	—	—	—
1	0.0751	2.30	0.32	0.024	0.0543	0.01	0.001
2	0.1279	1.35	0.60	0.076	0.0926	0.02	0.002
3	0.0751	2.30	1.97	0.148	0.0543	0.05	0.003
4	0.1235	1.35	1.85	0.229	0.0926	0.06	0.005
5	0.0626	2.30	4.75	0.298	0.0543	0.16	0.009
6	0.1299	1.35	4.60	0.597	0.0926	0.16	0.015
7	0.0681	2.30	11.68	0.795	0.0543	0.48	0.026
8	0.0957	1.35	10.63	1.018	0.0926	0.48	0.044
9	0.0566	2.30	30.85	1.746	0.0543	1.39	0.075
10	0.0859	1.35	30.37	2.608	0.0926	1.39	0.128
11	0.0504	2.30	78.33	3.948	0.0543	4.03	0.219
12	0.0805	1.35	62.33	5.019	0.0926	4.03	0.373
13	0.0450	2.30	121.58	5.471	0.0543	11.69	0.635
14	0.0767	1.35	65.41	5.015	0.0926	11.69	1.082
15	0.0450	2.30	81.59	3.672	0.0543	33.92	1.843
Medium of incidence	—	1.35	—	—	—	—	—
\sum	1.1978	—	506.8	30.662	1.0829	69.53	4.460

After Giacomo *et al* [27].

Thus, although the phase description filter using the reflectors shown in table 7.7 appears to be promising on the basis of the criterion (7.75), in the event its performance is somewhat disappointing. It is, however, certainly better than the straightforward classical filter. So far no design which better meets the condition of equation (7.72) has been proposed.

Some otherwise unpublished results obtained by Ritchie [28] are shown in figure 7.31. This filter used zinc sulphide and cryolite as the materials on glass as substrate. Its design is given in table 7.8. An experimental filter monitored at $1.348 \mu\text{m}$ gave peaks with corresponding bandwidths of

- 1.047 μm , bandwidth 3.0 nm
- 1.159 μm , bandwidth 2.5 nm
- 1.282 μm , bandwidth 4.0 nm.

Theoretically, the bandwidths should have been 0.8 nm, 1.7 nm and 4.6 nm respectively.

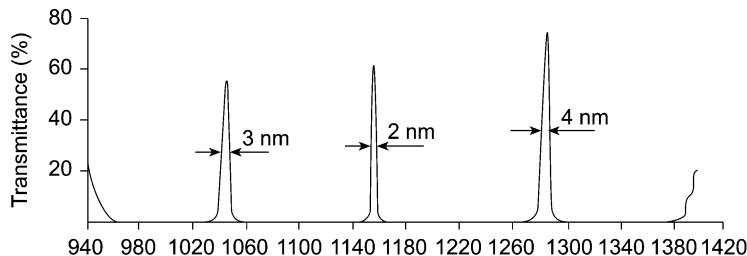


Figure 7.31. The measured transmittance of a 35-layer phase-dispersion filter. The design is given in table 7.7. (After Ritchie [28].)

Table 7.8.

Layer number	Material	Optical thickness as fraction of monitoring wavelength
1	ZnS	0.2375
2	Na ₃ AlF ₆	0.2257
3	ZnS	0.2143
4	Na ₃ AlF ₆	0.2036
5	ZnS	0.1934
6	Na ₃ AlF ₆	0.1838
7	ZnS	0.1746
8	Na ₃ AlF ₆	0.1649
9	ZnS	0.1576
10	Na ₃ AlF ₆	0.1498
11	ZnS	0.1423
12	Na ₃ AlF ₆	0.1352
13	ZnS	0.1285
14	Na ₃ AlF ₆	0.1220
15	ZnS	0.1159
16	Na ₃ AlF ₆	0.1101
17	ZnS	0.1046
Spacer	Na ₃ AlF ₆	0.5000

These 17 layers are followed by another 17 layers which are a mirror image of the first 17.

7.6 Multiple cavity metal-dielectric filters

Metal-dielectric filters are indispensable in suppressing the longwave sidebands of narrowband all-dielectric filters, and as filters in their own right, especially in the extreme shortwave region of the spectrum. Unlike all-dielectric filters, however, they possess the disadvantage of high intrinsic absorption. In single

Fabry–Perot filters this means that the pass bands must be wide in order to achieve reasonable peak transmission and the shape is far from ideal. It is possible to combine metal–dielectric elements into multiple cavity filters which, because of their more rectangular shape, are more satisfactory but, again, losses can be high.

The accurate design procedure for such metal–dielectric filters can be lengthy and tedious and frequently they are simply designed by trial and error as they are manufactured. We have already mentioned the metal–dielectric Fabry–Perot filter. These filters may be coupled together simply by depositing them one on top of the other with no coupling layer in between.

We can illustrate this by choosing silver as our metal, which we can give an index of $0.055 - i3.32$ at 550 nm [29]. The thickness of the spacer layer in the Fabry–Perot filter, as we have already noted, should be rather thinner than a half-wave at the peak wavelength to allow for the phase changes in reflection at the silver/dielectric interfaces. This phase change varies only slowly with silver thickness when it is thick enough to be useful as a reflector and we can assume, as a reasonable approximation, that it is equal to the limiting value for infinitely thick material. We can then use equation (4.5) to calculate the thickness of the spacer layer. Equation (4.5) calculates for us exactly one-half of the filter because it gives the thickness of the dielectric material to yield real admittance with zero phase change at the outer surface of the metal–dielectric combination. Adding a second exactly similar structure with the two dielectric layers facing each other, so that they join to form a single spacer, yields a Fabry–Perot filter in which the phase condition, equation (7.2), is satisfied.

Let us choose a spacer of index 1.35, similar to that of cryolite. Then half the spacer thickness is given by

$$D_f = \frac{1}{4\pi} \tan^{-1} \left(\frac{2\beta n_f}{n_f^2 - \alpha^2 - \beta^2} \right) \quad (7.76)$$

where $\alpha - i\beta$ is the index of the metal and n_f that of the cryolite and the angle is in the first or second quadrant.

With $\alpha - i\beta = 0.055 - i3.32$ and $n_f = 1.35$ we find

$$D_f = 0.18855$$

so that the spacer thickness should be 0.3771 full waves.

We can choose a metal layer thickness of 35 nm, quite arbitrarily, simply for the sake of illustration. Our Fabry–Perot filter is then

Glass	Ag	Cryolite	Ag	Glass
35 nm	0.3771 full waves		35 nm	

(the geometrical thickness being quoted for the silver and the optical thickness for the cryolite) and the DHW filter is exactly double this structure:

Glass	Ag	Cryolite	Ag	Cryolite	Ag	Glass.
35 nm	$D = 0.3771$	70 nm	$D = 0.3771$	35 nm		

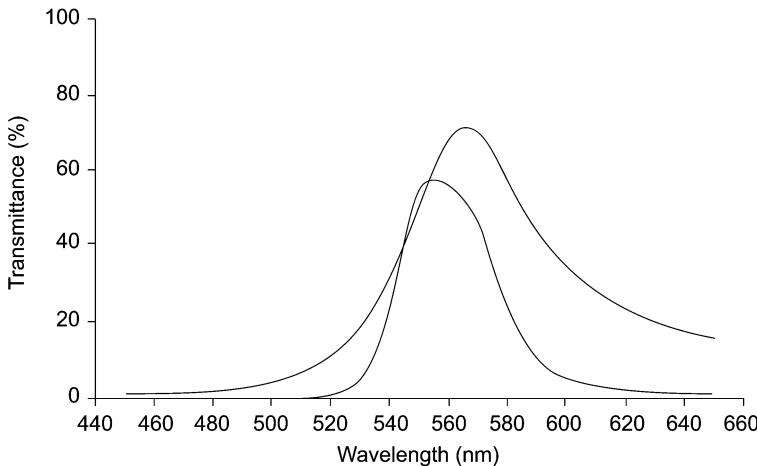


Figure 7.32. The transmittance as a function of wavelength of filters of design:

Glass | Silver 35 nm | Cryolite $0.3771\lambda_0$ | Silver 35 nm | Glass

and

Glass | Silver 35 nm | Cryolite $0.3771\lambda_0$ | Silver 70 nm | Cryolite $0.3771\lambda_0$ | Silver 35 nm | Glass

where $\lambda_0 = 550$ nm, $n - ik = 0.055 - i3.32$ and $n_{\text{cryolite}} = 1.35$. Dispersion in the materials has been neglected.

Curves of these filters are shown in figure 7.32. The peaks are slightly displaced from 550 nm because of the approximations inherent in the design procedure.

The Fabry-Perot has reasonably good peak transmission but its typical triangular shape means that its rejection is quite poor even at wavelengths far from the peak. The DHW filter has better shape but rather poorer peak transmittance. The rejection can be improved by increasing the metal thickness, but at the expense of peak transmission.

The design approach we have described is quite crude and simply concentrates on ensuring that the peak of the filter is centred near the desired wavelength. Peak transmittance and bandwidth are either accepted as they are or a new metal thickness is tried. Performance is in no way optimised.

The unsatisfactory nature of this design procedure led Berning and Turner [30] to develop a new technique for the design of metal–dielectric filters in which the emphasis is on ensuring that maximum transmittance is achieved in the filter pass band. For this purpose they devised the concept of potential transmittance and created a new type of metal–dielectric filter known as the induced-transmission filter.

7.6.1 The induced-transmission filter

Given a certain thickness of metal in a filter, what is the maximum possible peak transmission, and how can the filter be designed to realise this transmission? This is the basic problem tackled and solved by Berning and Turner [30]. The development of the technique as given here is based on their approach, but it has been adjusted and adapted to conform more nearly to the general pattern of this book.

The concept of potential transmittance has already been touched on in chapter 2 and used in the analysis of losses in dielectric multilayers. We recall that the potential transmittance ψ of a layer or assembly of layers is defined as the ratio of the intensity leaving the rear surface to that actually entering at the front surface, and it represents the transmittance which the layer or assembly of layers would have if the reflectance of the front surface were reduced to zero. Thus, once the parameters of the metal layer are fixed, the potential transmittance is determined entirely by the admittance of the structure at the exit face of the layer. Furthermore, it is possible to determine the particular admittance which gives maximum potential transmittance. To achieve this transmittance it is sufficient to add a coating to the front surface to reduce the reflectance to zero. The maximum potential transmittance is a function of the thickness of the metal layer.

The design procedure is then as follows. The optical constants of the metal layer at the peak wavelength are given. Then the metal layer thickness is chosen and the maximum potential transmittance together with the matching admittance at the exit face of the layer, which is required to produce that level of potential transmittance, is found. Often a minimum acceptable figure for the maximum potential transmittance will exist and that will put an upper limit on the metal layer thickness. A dielectric assembly which will give the correct matching admittance when deposited on the substrate must then be designed. The filter is then completed by the addition of a dielectric system to match the front surface of the resulting metal–dielectric assembly to the incident medium. Techniques for each of these steps will be developed. The matching admittances for the metal layer are such that the dielectric stacks are efficient in matching over a limited region only, outside which their performance falls off rapidly. It is this rapid fall in performance that defines the limits of the pass band of the filter.

Before we can proceed further, we require some analytical expressions for the potential transmittance and for the matching admittance. This leads to some lengthy and involved analysis, which is not difficult but rather time-consuming.

(a) Potential transmittance

We limit the analysis to an assembly in which there is only one absorbing layer, the metal. The potential transmittance is then related to the matrix for the

assembly, as shown in chapter 2

$$\begin{bmatrix} B'_i \\ C'_i \end{bmatrix} = [M] \begin{bmatrix} 1 \\ Y_e \end{bmatrix}$$

where $[M]$ is the characteristic matrix of the metal layer and Y_e is the admittance of the terminating structure. Then the potential transmittance ψ is given by

$$\psi = \frac{T}{(1-R)} = \frac{\operatorname{Re}(Y_e)}{\operatorname{Re}(B'_i C'^{*}_i)}. \quad (7.77)$$

Let

$$Y_i = X + iZ.$$

Then

$$\begin{bmatrix} B'_i \\ C'_i \end{bmatrix} = \begin{bmatrix} \cos \delta & (i \sin \delta)/y \\ iy \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} 1 \\ X + iZ \end{bmatrix}$$

where

$$\begin{aligned} \delta &= 2\pi(n - ik)d/\lambda = 2\pi nd/\lambda - i2\pi kd/\lambda \\ &= \alpha - i\beta \\ \alpha &= 2\pi nd/\lambda \\ \beta &= 2\pi kd/\lambda. \end{aligned}$$

If free space units are used, then

$$y = n - ik.$$

Now,

$$\begin{aligned} (B'_i C'^{*}_i) &= [\cos \delta + i(\sin \delta/y)(X + iZ)][iy \sin \delta + \cos \delta(X + iZ)]^* \\ &= [\cos \delta + i(\sin \delta/y)(X + iZ)][-iy^* \sin \delta^* + \cos \delta^*(X - iZ)] \\ &= -iy^* \cos \delta \sin \delta^* + \frac{\sin \delta \sin \delta^* y^{*2} (X + iZ)}{yy^*} \\ &\quad + \cos \delta \cos \delta^*(X - iZ) + \frac{i \sin \delta \cos \delta^* y^* (X - iZ)(X + iZ)}{yy^*}. \end{aligned}$$

We require the real part of this and we take each term in turn.

$$\begin{aligned} -iy^* \cos \delta \sin \delta^* &= -i(n + ik)(\cos \alpha \cosh \beta + i \sin \alpha \sinh \beta)(\sin \alpha \cosh \beta \\ &\quad + i \cos \alpha \sinh \beta) \end{aligned}$$

and the real part of this, after a little manipulation, is

$$\operatorname{Re}(-iy^* \cos \delta \sin \delta^*) = n \sinh \beta \cosh \beta + k \cos \alpha \sin \alpha.$$

Similarly

$$\operatorname{Re}\left(\frac{\sin \delta \sin \delta^* y^{*2} (X + iZ)}{yy^*}\right) = \frac{X(n^2 - k^2) - 2nkZ}{(n^2 + k^2)} (\sin^2 \alpha \cosh^2 \beta + \cos^2 \alpha \sinh^2 \beta)$$

$$\operatorname{Re}[\cos \delta \cos \delta^* (X - iZ)] = X(\cos^2 \alpha \cosh^2 \beta + \sin^2 \alpha \sinh^2 \beta)$$

$$\begin{aligned} \operatorname{Re}\left(\frac{i \sin \delta \cos \delta^* y^* (X - iZ)(X + iZ)}{yy^*}\right) \\ = \frac{X^2 + Z^2}{(n^2 + k^2)} (n \sinh \beta \cosh \beta - k \sin \alpha \cos \alpha). \end{aligned}$$

The potential transmittance is then

$$\begin{aligned} \psi = & \left(\frac{(n^2 - k^2) - 2nk(Z/X)}{(n^2 + k^2)} (\sin^2 \alpha \cosh^2 \beta + \cos^2 \alpha \sinh^2 \beta) \right. \\ & + (\cos^2 \alpha \cosh^2 \beta + \sin^2 \alpha \sinh^2 \beta) \\ & + \frac{1}{X} (n \sinh \beta \cosh \beta + k \cos \alpha \sin \alpha) \\ & \left. + \frac{X^2 + Z^2}{X(n^2 + k^2)} (n \sinh \beta \cosh \beta - k \cos \alpha \sin \alpha) \right)^{-1}. \end{aligned} \quad (7.78)$$

(b) Optimum exit admittance

Next we find the optimum values of X and Z . From equation (7.78)

$$\frac{1}{\psi} = \left(\frac{q[n^2 - k^2 - 2nk(Z/X)]}{[n^2 + k^2]} + r + \frac{p}{X} + \frac{s(X^2 + Z^2)}{X(n^2 + k^2)} \right) \quad (7.79)$$

where p , q , r and s are shorthand for the corresponding expressions in equation (7.78). For an extremum in ψ , we have an extremum in $1/\psi$ and hence

$$\frac{\partial}{\partial X} \left(\frac{1}{\psi} \right) = 0 \quad \text{and} \quad \frac{\partial}{\partial Z} \left(\frac{1}{\psi} \right) = 0$$

i.e.

$$\frac{q2nkZ}{X^2(n^2 + k^2)} - \frac{p}{X^2} + \frac{s}{(n^2 + k^2)} - \frac{sZ^2}{X^2(n^2 + k^2)} = 0 \quad (7.80)$$

and

$$\frac{q(-2nk)}{X(n^2 + k^2)} + \frac{2sZ}{X(n^2 + k^2)} = 0. \quad (7.81)$$

From equation (7.81):

$$Z = nkq/s$$

and, substituting for equation (7.80),

$$X^2 = p(n^2 + k^2)/s - n^2k^2q^2/s^2.$$

Then, inserting the appropriate expressions for p , q and s , from equation (7.79)

$$X = \left(\frac{(n^2 + k^2)(n \sinh \beta \cosh \beta + k \sin \alpha \cos \alpha)}{(n \sinh \beta \cosh \beta - k \sin \alpha \cos \alpha)} - \frac{n^2k^2(\sin^2 \alpha \cosh^2 \beta + \cos^2 \alpha \sinh^2 \beta)^2}{(n \sinh \beta \cosh \beta - k \sin \alpha \cos \alpha)^2} \right)^{1/2} \quad (7.82)$$

$$Z = \frac{nk(\sin^2 \alpha \cosh^2 \beta + \cos^2 \alpha \sinh^2 \beta)}{(n \sinh \beta \cosh \beta - k \sin \alpha \cos \alpha)}. \quad (7.83)$$

We note that for β large $X \rightarrow n$ and $Z \rightarrow k$, that is:

$$Y_e \rightarrow (n + ik) = (n - ik)^*.$$

(c) Maximum potential transmittance

The maximum potential transmittance can then be found by substituting the values of X and Z , calculated by equations (7.82) and (7.83), into equation (7.78). All these calculations are best performed by computer or calculator and so there is little advantage in developing a separate analytical solution for maximum potential transmittance.

(d) Matching stack

We have to device an assembly of dielectric layers which, when deposited on the substrate, will have an equivalent admittance of

$$Y = X + iZ.$$

This is illustrated diagrammatically in figure 7.33 where a substrate of admittance $(n_s - ik_s)$ has an assembly of dielectric layers terminating such that the final equivalent admittance is $(X + iZ)$. Now, the dielectric layer circles are executed in a clockwise direction always. If we therefore reflect the diagram in the x axis and then reverse the direction of the arrows, we get exactly the same set of circles—that is, the layer thicknesses are exactly the same—but the order is reversed (it was ABC and is now CBA) and they match a starting admittance of $X - iZ$, i.e. the complex conjugate of $(X + iZ)$, into a terminal admittance of $(n_s + ik_s)$, i.e. the complex conjugate of the substrate index. In our filters the substrate will have real admittance, i.e. $k_s = 0$, and it is a more straightforward problem to match $(X - iZ)$ into n_s than n_s into $(X + iZ)$.

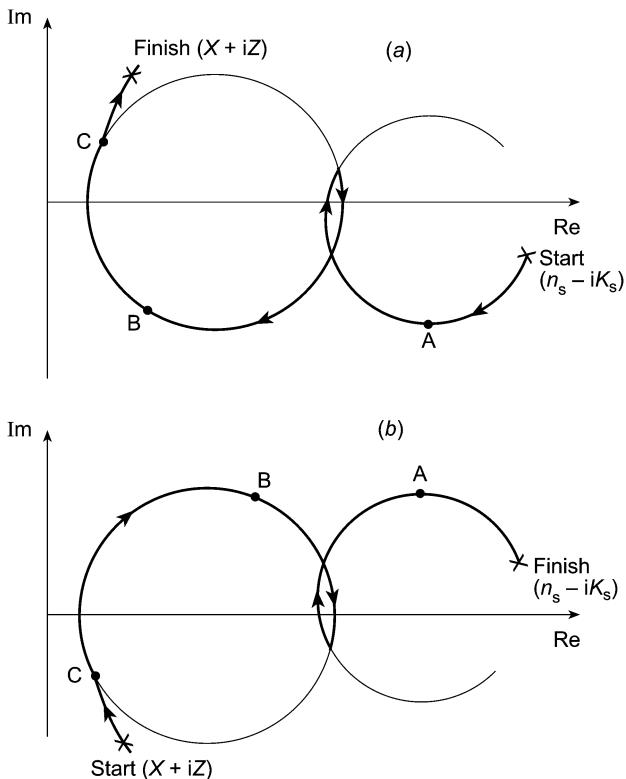


Figure 7.33. (a) A sketch of the admittance diagram of an arbitrary dielectric assembly of layers matching a starting admittance of $(n_s - ik_s)$ to the final admittance of $(X + iZ)$. (b) The curves of figure 7.33(a) reflected in the real axis and with the directions of the arrows reversed. This is now a multilayer identical to (a) but in the opposite order and connecting an admittance of $(X - iZ)$ (i.e. $(X + iZ)^*$) to one of $(n_s + ik_s)$ (i.e. $(n_s - ik_s)^*$).

There is an infinite number of possible solutions, but the simplest involves adding a dielectric layer to change the admittance $(X - iZ)$ into a real value and then to add a series of quarter-waves to match the resultant real admittance into the substrate. We will illustrate the technique shortly with several examples. At the moment we recall that the necessary analysis was carried out in chapter 4. There we showed that a film of optical thickness D given by

$$D = \frac{1}{4\pi} \tan^{-1} \left(\frac{2Zn_f}{(n_f^2 - X^2 - Z^2)} \right) \quad (7.84)$$

(where the tangent is taken in the first or second quadrant) will convert an

admittance ($X - iZ$) into a real admittance of value

$$\mu = \frac{2Xn_f^2}{(X^2 + Z^2 + n_f^2) + [(X^2 + Z^2 + n_f^2)^2 - 4X^2n_f^2]^{1/2}}. \quad (7.85)$$

n_f can be of high or low index, but μ will always be lower than the index of the substrate (except in very unlikely cases) because it is the first intersection of the locus of n_f with the real axis which is given by equations (7.84) and (7.85). Since the substrate will always have an index greater than unity, then the quarter-wave stack to match μ to n_s must start with a quarter-wave of low index. Alternate high- and low-index layers follow, the precise number being found by trial and error.

In order to complete the design, we need to know the equivalent admittance at the front surface of the metal layer and then we construct a matching stack to match it to the incident medium.

(e) Front surface equivalent admittance

If the admittance of the structure at the exit surface of the metal layer is the optimum value ($X + iZ$) given by equations (7.82) and (7.83), then it can be shown that the equivalent admittance which is presented by the front surface of the metal layer is simply the complex conjugate ($X - iZ$). The analytical proof of this requires a great deal of patience, although it is not particularly difficult. Instead, let us use a logical justification.

Consider a filter consisting of a single metal layer matched on either side to the surrounding media by dielectric stacks. Let the transmittance of the assembly be equal to the maximum potential transmittance and let the admittance of the structure at the rear of the metal layer be the optimum admittance ($X + iZ$) given by equations (7.82) and (7.83). Let the equivalent admittance at the front surface be ($\xi + i\eta$) and let this be matched perfectly to the incident medium. Now we know that the transmittance is the same regardless of the direction of incidence. Let us turn the filter around, therefore, so that the transmitted light proceeds in the opposite direction. The transmittance of the assembly must be the maximum potential transmittance once again. The admittance of the structure at what was earlier the input, but is now the new exit face of the metal layer, must therefore be ($X + iZ$). But, since the layers are dielectric and the medium is of real admittance, this must also be the complex conjugate of ($\xi + i\eta$), that is, ($\xi - i\eta$). ($\xi + i\eta$) must therefore be ($X - iZ$), which is what we set out to prove.

The procedure for matching the front surface to the incident medium is therefore exactly the same as that for the rear surface and, indeed, if the incident medium is identical to the rear exit medium, as in a cemented filter assembly, then the front dielectric section can be an exact repetition of the rear.

7.6.2 Examples of filter designs

We can not attempt some filter designs. We choose the same material, silver, as we did for the Fabry–Perot and the DHW filters earlier. Once again, arbitrarily, we select a thickness of 70 nm. The wavelength we retain as 550 nm, at which the optical constants of silver are 0.055 – i3.32.

The filter is to use dielectric materials of indices 1.35 and 2.35 corresponding to cryolite and zinc sulphide respectively. The substrate is glass, $n = 1.52$, and the filter will be protected by a cemented cover slip so that we can also use $n = 1.52$ for the incident medium.

$$\alpha = 2\pi nd/\lambda = 0.04398$$

$$\beta = 2\pi kd/\lambda = 2.6549$$

and from equations (7.82) and (7.83) we find the optical admittance

$$X + iZ = 0.4572 + i3.4693.$$

Substituting this in equation (7.78) gives

$$\psi = 80.50\%.$$

We can choose to have either a high- or a low-index spacer. Let us choose first a low index and from equation (7.84) we obtain an optical thickness for the 1.35 index layer of 0.19174 full waves. Equation (7.85) yields a value of 0.05934 for μ which must be matched to the substrate index of 1.52. We start with a low-index quarter-wave and simply work through the sequence of possible admittances:

$$\frac{n_L^2}{\mu}, \quad \frac{n_H^2 \mu}{n_L^2}, \quad \frac{n_L^4}{n_H^2 \mu}, \quad \frac{n_H^4 \mu}{n_L^4} \quad \text{etc}$$

until we find one sufficiently close to 1.52. The best arrangement in this case involves three layers of each type.

$$\frac{n_H^6 \mu}{n_L^6} = 1.6511$$

equivalent to a loss of 0.2% at the interface with the substrate.

The structure so far is then

$$|\text{Ag}|L''LHLH| \text{Glass} \tag{7.86}$$

with $L'' = 0.19174$ full waves. This can be combined with the following L layer into a single layer $L' = 0.25 + 0.19174 = 0.44174$ full waves, i.e.

$$|\text{Ag}|L'HHLH| \text{Glass}.$$

Since the medium is identical to the substrate then the matching assembly at the front will be exactly the same as that at the rear so that the complete design is

$$\text{Glass}|H L H L H L' \text{Ag} L' H L H L H| \text{Glass}$$

with

$$\begin{aligned} & \text{Ag } 70 \text{ nm (geometrical thickness)} \\ & L' 0.44174 \text{ full waves (optical thickness)} \\ & H, L 0.25 \text{ full waves} \\ & \lambda_0 550 \text{ nm.} \end{aligned}$$

The performance of this design is shown in figure 7.34(a). Dispersion of the silver has not been taken into account to give a clearer idea of the intrinsic characteristics. The peak is indeed centred at 550 nm with transmittance virtually that predicted.

A high-index matching layer can be handled in exactly the same way. For an index of 2.35, equation (7.84) yields an optical thickness of 0.1561 and equation (7.85) gives a value of 0.1426 for μ . Again, the matching quarter-wave stack should start with a low-index layer. There are two possible arrangements, H' representing 0.1561 full waves:

$$(a) \quad \text{Ag} H' L H L H | \text{Glass}$$

with $n_H^4 \mu / n_L^4 = 1.310$, i.e. a loss of 0.6% at the glass interface, or

$$(b) \quad \text{Ag} H' L H L H | \text{Glass}$$

with $n_L^6 / n_H^4 \mu = 1.392$ representing a loss of 0.2% at the glass interface.

We choose alternative (b) and the full design can then be written

$$\text{Glass}|H L H L H' \text{Ag} H' L H L H| \text{Glass}$$

with

$$\begin{aligned} & \text{Ag } 70 \text{ nm (geometrical thickness)} \\ & H' 0.1561 \text{ full waves (optical thickness)} \\ & H, L 0.25 \text{ full waves.} \end{aligned}$$

The performance of this design is shown in figure 7.34(b), where, again the dispersion of silver has not been taken into account. Peak transmission is virtually as predicted.

When, however, we plot the performance of any of these designs, including the metal-dielectric Fabry-Perot and DHW filters over an extended wavelength region, we find that the performance at longer wavelengths appears very

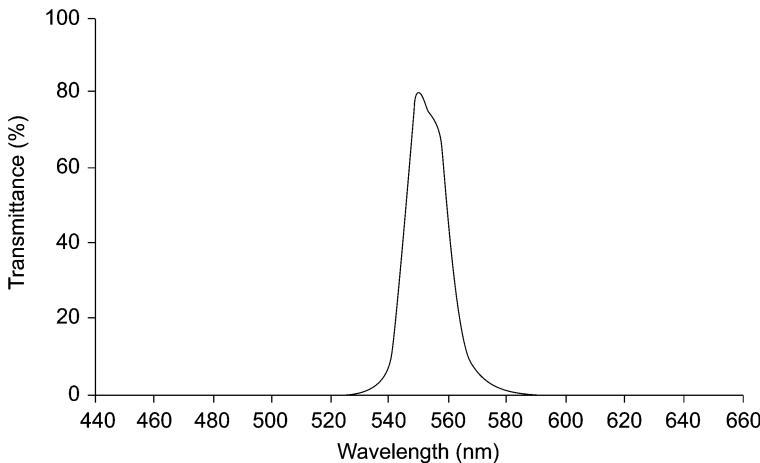


Figure 7.34. (a) Calculated performance of the design:

Glass| H L H L H L'Ag L' H L H L H|Glass

where

$$n_{\text{Glass}} = 1.52$$

Ag = 70 nm (geometrical thickness) of index $0.055 - i3.32$

$H = 0.25\lambda_0$ (optical thickness) of index 2.35

$L = 0.25\lambda_0$ (optical thickness) of index 1.35

$L' = 0.4417\lambda_0$ (optical thickness) of index 1.35

$\lambda_0 = 550$ nm.

Dispersion has been neglected.

disappointing. One example, the low-index matched induced-transmission filter, is shown in figure 7.35(a). In the case of the Fabry–Perot and the DHW, the rise is smoother, but is of a similar order of magnitude. The reason for the rise is, in fact, our assumption of zero dispersion. This means that β is reduced as λ increases. α is always quite small and the performance of the metal layers is determined principally by β . Silver, however, over the visible and near infrared, shows an increase in k which corresponds roughly to the increase in λ so that k/λ is roughly constant (to within around $\pm 20\%$) over the region 400 nm–2.0 μm . This completely alters the picture and is the reason why the first-order metal–dielectric filters do not show longwave sidebands.

Taking dispersion into account, the performance of the induced transmission filter improves considerably and is shown in figure 7.35(b). The rejection is, however, not particularly high, being between 0.01 and 0.1% transmittance over most of the range with an increase to 0.15% in the vicinity of 860 nm. This level

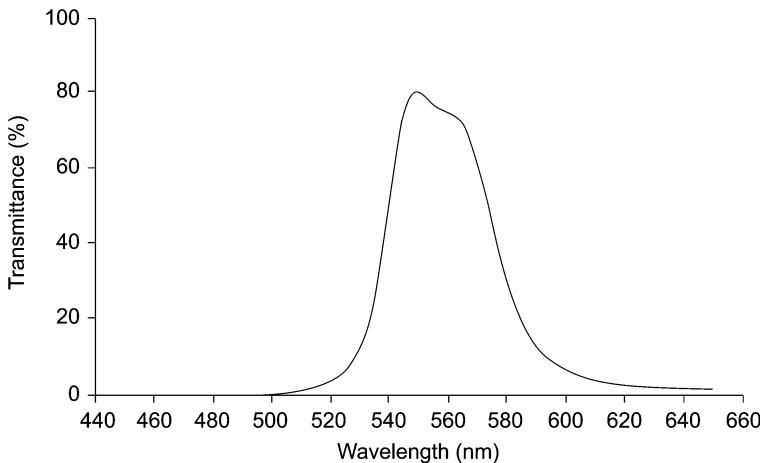


Figure 7.34. (b) Calculated performance of the design:

Glass|H L H L H' Ag H' L H L H|Glass

where

$$n_{\text{Glass}} = 1.52$$

Ag = 70 nm (geometrical thickness) of index $0.055 - i3.32$

$H = 0.25\lambda_0$ (optical thickness) of index 2.35

$L = 0.25\lambda_0$ (optical thickness) of index 1.35

$H' = 0.1561\lambda_0$ (optical thickness) of index 2.35

$\lambda_0 = 550$ nm.

Dispersion has been neglected.

of rejection can be acceptable in some applications and the induced-transmission filter represents a very useful, inexpensive general purpose filter. The dispersion which improves the performance on the longwave side of the peak degrades it on the shortwave side, and to complete the filter it is normal to add a longwave-pass absorption glass filter which is cemented to the induced transmission component.

To improve the rejection of the basic filter it is necessary to add further metal layers. The simplest arrangement is to have these extra metal layers of exactly the same thickness as the first. The potential transmittance of the complete filter will then be the product of the potential transmittances of the individual layers. The terminal admittances for all the metal layers can be arranged to be optimum quite simply, giving optimum performance for the filter. All that is required is a dielectric layer in between the metal layers which is twice the thickness given by equation (7.84) for the first matching layer. We can see why this is by imagining a matching stack on the substrate overcoated with the first metal

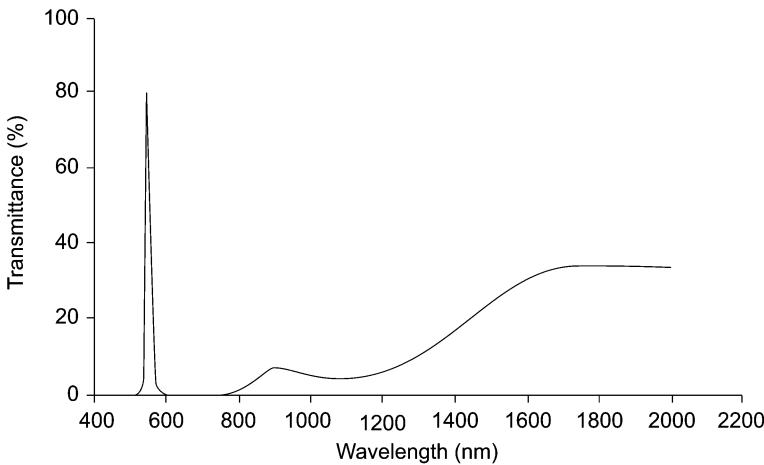


Figure 7.35. (a) The design of figure 7.34(a) computed over a wider spectral region neglecting dispersion.

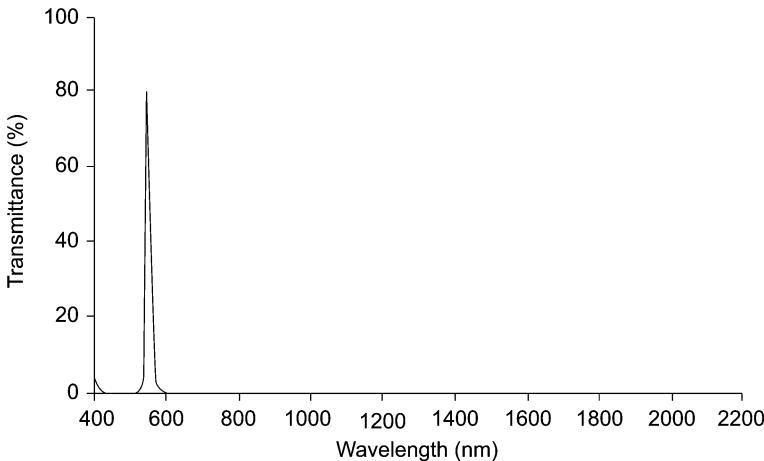


Figure 7.35. (b) The design of figure 7.34(a) computed this time including dispersion. The rise in transmittance at longer wavelengths has vanished but there is now obvious transmittance at 400 nm.

layer. Since its terminal admittance will be optimum, the input admittance will be the complex conjugate, as we have discussed already. Addition of the thickness given by equation (7.84) renders the admittance real, that is, the admittance locus has reached the real axis. Addition of a further identical thickness must give an equivalent input admittance which is the complex conjugate of the metal input admittance and hence is equal to the optimum admittance. This can be repeated

as often as desired.

Returning to our example, a two-metal layer induced-transmission filter will have peak transmission, if perfectly matched, of $\psi = (0.80501)^2$, that is, 64.8%, a three-metal layer should have $\psi = (0.80501)^3$ that is, 52.17%, and so on.

The designs, based on the low-index matching layer version, are then, from equation (7.86)

$$\begin{aligned} \text{Glass} & | H L H L H L L'' \text{Ag} L'' L'' \text{Ag} L'' L H L H L H | \text{Glass} \\ & = \text{Glass} | H L H L H L' \text{Ag} L''' \text{Ag} L' L H L H L H | \text{Glass} \end{aligned} \quad (7.87)$$

where

$$L' = 0.25 + 0.19174 = 0.44174 \text{ full waves}$$

$$L'' = 0.19174 \text{ full waves}$$

$$L''' = 2 \times 0.19174 = 0.38348 \text{ full waves}$$

$$\text{Ag} = 70 \text{ nm}$$

and

$$\text{Glass} | H L H L H L' \text{Ag} L''' \text{Ag} L''' \text{Ag} L' L H L H L H | \text{Glass}. \quad (7.88)$$

Unfortunately, these designs, although they do have the peak transmittance predicted, possess a poor pass-band shape, in that it has a hump on the longwave side. To eliminate this hump, it is necessary to add an extra half-wave layer to each of the layers marked L''' , i.e.

$$\text{Glass} | H L H L H L' \text{Ag} L'''' \text{Ag} L' L H L H L H | \text{Glass} \quad (7.89)$$

and

$$\text{Glass} | H L H L H L' \text{Ag} L'''' \text{Ag} L'''' \text{Ag} L' L H L H L H | \text{Glass} \quad (7.90)$$

where

$$L'''' = 0.5 + 0.38348 = 0.88348 \text{ full waves.}$$

Figure 7.36 shows the form of designs (7.87) and (7.88) and the hump can clearly be seen together with the improved shape of designs (7.89) and (7.90).

Dispersion was not included in the computation of figure 7.36. To examine the rejection over an extended region, we must include the effects of dispersion. Unfortunately, the modified designs (7.89) and (7.90) act as metal-dielectric-metal ($M-D-M$ is a frequently used shorthand notation for such a filter) and metal-dielectric-metal-dielectric-metal ($M-D-M-D-M$) filters at approximately 1100 nm which gives a very narrow leak, rising to around 0.15% in the former and 0.05% in the latter. Elsewhere, the rejection is excellent, of the order of 0.0001% at 900 nm and 0.000015% at $1.05 \mu\text{m}$ for the former and 0.0000001% at 900 nm and $3 \times 10^{-9}\%$ at $1.05 \mu\text{m}$ for the latter.

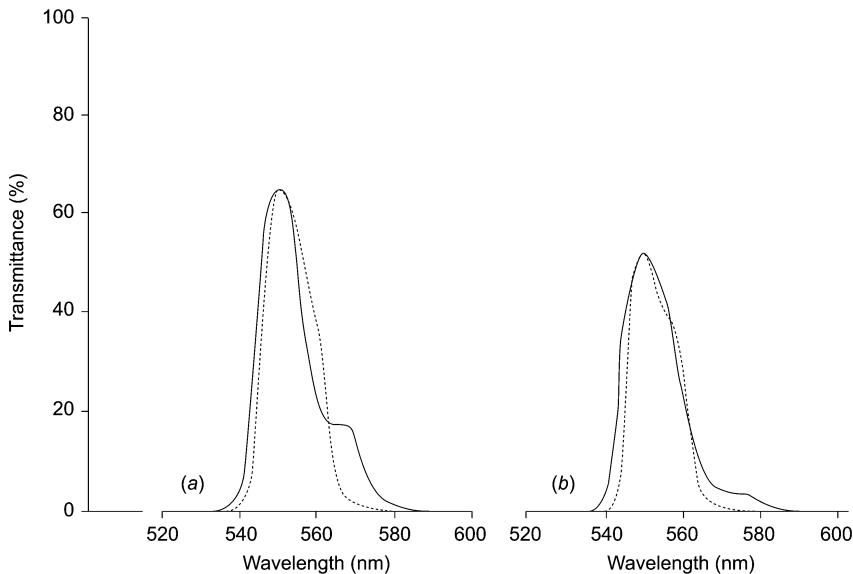


Figure 7.36. Performance, neglecting dispersion, of (a) two-metal-layer designs and (b) three-metal-layer designs of induced-transmission filter. The full curves denote (7.87) and (7.88) and there is a spurious shoulder on the longwave side of the peak in each case. This can be eliminated by the addition of half-wave decoupling layers as the dashed lines show. They are derived from (7.89) and (7.90) respectively.

If the leak is unimportant, then the filter can be used as it is with the addition of a longwave-pass filter of the absorption type as before. For the suppression of all-dielectric filter sidebands, it is better to use filters of type (7.87) and (7.88) since the shape of the sides of the pass band is relatively unimportant. The rejection of these filters is slightly better than that of (7.89) and (7.90) and, of course, the leak is missing (figure 7.37).

The bandwidth of the filters is not an easy quantity to predict analytically and the most straightforward approach is simply to compute the filter profile.

Berning and Turner [30] show that a figure of merit indicating the potential usefulness of a metal is the ratio k/n . The higher this ratio, the better is the performance of the completed filter.

Induced-transmission filters for the visible region having only one single metal layer are relatively straightforward to manufacture. The thickness of the metal layer can be arrived at by trial and error. If the metal layer is less than optimum in thickness, the effect will be a broadening of the pass band and a rise in peak transmission at the expense of an increase in background transmission remote from the peak. A splitting of the pass band will also become noticeable with the appearance eventually, if the thickness is further reduced, of two separate

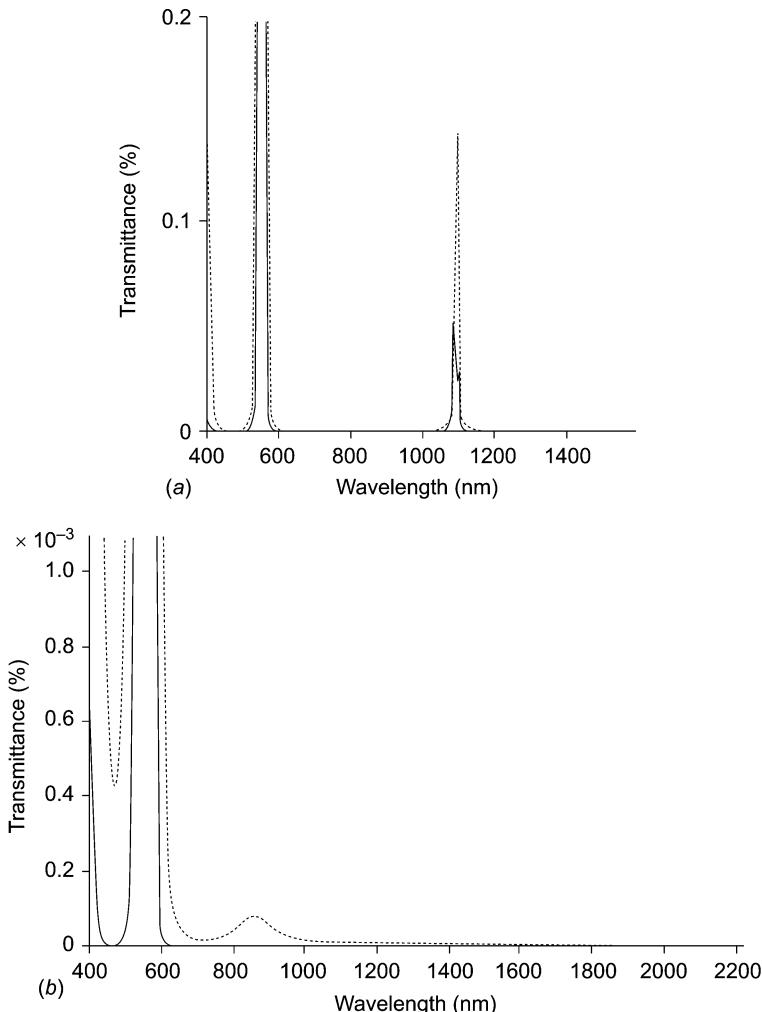


Figure 7.37. (a) Calculation, including dispersion, of the performance of the designs of (7.89) (dashed curve) and (7.90) over an extended spectral range. These designs include the half-wave decoupling layers and the penalty for the improved pass-band shape is the narrow transmission spike near $1.05\ \mu\text{m}$. (b) Calculation, including dispersion, of the original designs (7.87) (dashed curve) and (7.88). The transmission spike is no longer there but the pass-band shape includes the shoulder (off scale).

peaks. If, on the other hand, the silver layer is made too thick, the effect will be a narrowing of the peak with a reduction of peak transmission. The best results are usually obtained with a compromise thickness where the peak is still single

in shape but where any further reduction in silver thickness would cause the splitting to appear. A good approximation in practice, which can be used as a first attempt at a filter, is to deposit the first dielectric stack and to measure the transmission. The silver layer can then be deposited using a fresh monitor glass so that the optical density is twice that of the dielectric stack. The second spacer and stack can then be added on yet another fresh monitor. A measurement of the transmission of the complete filter will quickly indicate which way the thickness of the silver layer should be altered in order to optimise the design. Usually, one or two tests are sufficient to establish the best parameters. If, after this optimising, the background rejection remote from the peak is found to be unsatisfactory, then not enough silver is being used. As the thickness was chosen to be optimum for the two dielectric sections, a pair of quarter-wave layers should be added to each in the design and the trial-and-error optimisation repeated. This will also narrow the bandwidth, but this is usually preferable to high background transmission.

In the ultraviolet the available metals do not have as high a performance as, for instance, silver in the visible, and it is very important, therefore, to ensure that the design of a filter is optimised as far as possible; otherwise a very inferior performance will result. An important paper in this field is that by Baumeister *et al* [31]. Aluminium is the metal commonly used for this region and measured and computed results obtained by these workers for filters with aluminium layers are shown in figure 7.38. The performance which has been achieved is most satisfactory and the agreement between practical and theoretical curves is good.

Induced-transmission filters have been the subject of considerable study by many workers. Metal–dielectric multilayers are reviewed by MacDonald [32]. A useful, recent account of induced-transmission filters is given by Lissberger [33]. Multiple cavity induced-transmission filters have been described by Maier [34]. An alternative design technique for metal–dielectric filters involving symmetrical periods has been published by Macleod [35]. Symmetrical periods for metal–dielectric filter design have also been used by McKenney [36] and by Landau and Lissberger [37].

7.7 Measured filter performance

Not a great deal has been published on the measured performance of actual filters and the main source of information for a prospective user is always the literature issued by manufacturers. Performance of current production filters tends to improve all the time so that inevitably such information does not remain up to date for long. Two papers [38, 39] quote the results of a number of tests on commercial filters, and, although they were written some time ago, they will still be found useful sources of information.

Blifford examined the performance of the products of four different manufacturers, covering the region 300–1000 nm. The variation of peak wavelength with angle of incidence was found to be similar to the relationship

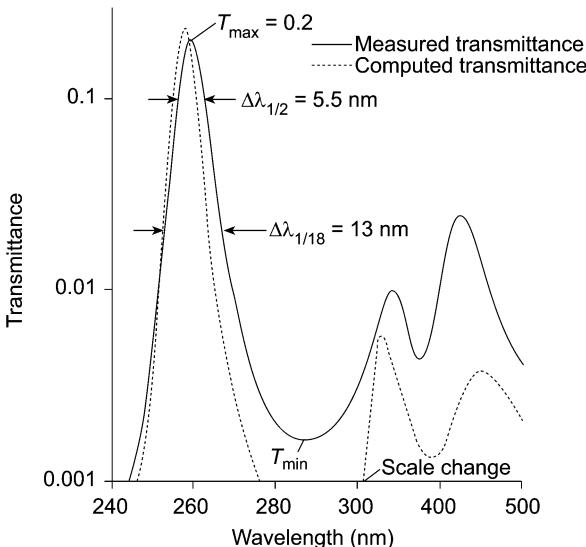


Figure 7.38. Computed and measured transmittance of an induced transmission filter for the ultraviolet. Design:

Air|*H*L*H*L*H*L*H* 1.76*L* Al 1.76*L* *H*L*H*L*H*L*H*|Quartz

where $H = \text{PbF}_2$ ($n_H = 2.0$) and $L = \text{Na}_3\text{AlF}_6$ ($n_L = 1.36$). The physical thickness of the aluminium layer is 40 nm and $\lambda_0 = 253.6$ nm. (After Baumeister *et al* [31].)

already established (see p 283). Unfortunately, information on the design and materials is lacking, so that the expression for the effective index cannot be checked. The sensitivities to tilt varied from $P = 0.22$ to $P = 0.51$, where P corresponds to the quantity $1/n^{*2}$ in equation (7.39). Blifford suggests that an average value of 0.35 for P would probably be the best value to assume in any case where no other data were available. Changes in peak transmittance with angle of incidence were found, but were not constant from one filter to another and apparently must always be measured for each individual filter. Possibly, the effect is due to the absorption filters which are used for sideband suppression and which, because they do not show any shift in edge wavelength with angle of incidence, may cut into the pass band of the interference section at large angles of incidence. In most cases examined, the change in peak transmission was less than 10% for angles of 5° – 10° .

The variation in peak transmittance over the surface of the filter was also measured in a few cases. For a typical filter with a peak wavelength of 500 nm and a bandwidth not explicitly mentioned, but probably 2.1 nm (from information

given elsewhere in the paper), the extremes of peak transmission were 54% and 60%. This is, in fact, one aspect of a variation of peak wavelength, bandwidth and peak transmittance which frequently occurs, although the magnitude can range from very small to very large. The cause is principally the adsorption of water vapour from the atmosphere before a cover slip can be cemented over the layers and it is dealt with in greater detail in chapter 9. Infrared filters appear to suffer less from this defect than visible and near infrared filters.

Another parameter measured by Blifford was the variation of peak wavelength with temperature. Variation of the temperature from $-60\text{ }^{\circ}\text{C}$ to $+60\text{ }^{\circ}\text{C}$ resulted in changes of peak wavelength from $+0.01\text{ nm }^{\circ}\text{C}^{-1}$ to $+0.03\text{ nm }^{\circ}\text{C}^{-1}$. The relationship was found to be linear over the whole of this temperature range with little, if any, change in the pass-band shape and peak transmittance. In most cases, the temperature coefficients of bandwidth and peak transmittance were found to be less than $0.01\text{ nm }^{\circ}\text{C}^{-1}$. Filters for the visible region have also been the subject of a detailed study by Pelletier and his colleagues [40]. The shift with temperature for any filter is a function of the coefficients of optical thickness change with temperature, depending on the design of the filter and especially on the material used for the spacers. Measurements made on different filter designs yielded the following coefficients of optical thickness for the individual layer materials:

$$\begin{aligned} \text{zinc sulphide} & \quad (4.8 \pm 1.0) \times 10^{-5} \text{ }^{\circ}\text{C}^{-1} \\ \text{cryolite} & \quad (3.1 \pm 0.7) \times 10^{-5} \text{ }^{\circ}\text{C}^{-1}. \end{aligned}$$

Hysteresis is frequently found with temperature cycling narrowband filters over an extended temperature range. The hysteresis is particularly pronounced when the filters are uncemented and when they are heated towards $100\text{ }^{\circ}\text{C}$. It is usually confined to the first cycle of temperature, takes the form of a shift of peak wavelength towards shorter wavelengths and is caused by the desorption of water which is discussed again in chapter 9.

An effect of a different kind, although related, is the subject of a contribution by Title and his colleagues [41, 42]. A permanent shift of a filter characteristic towards shorter wavelengths amounting to a few tenths of nanometres accompanied by a distortion of pass-band shape was produced by a high level of illumination. The filters were for the H_{α} wavelength, 656.3 nm, and the changes were interpreted as due to a shift in the properties of the zinc sulphide material, the fundamental nature of the shift being unknown. Zinc sulphide can be transformed into zinc oxide by the action of ultraviolet light, especially in the presence of moisture, and the shifts that were observed could probably have been caused by such a mechanism.

The possibility of variations in filter properties both over the surface of the filter and as a function of time, temperature and illumination level should clearly be borne in mind in the designing of apparatus incorporating filters.

A useful survey which compares the performance achievable from different types of narrowband filters was the subject of a report by Baumeister [43].

A study was carried out by Baker and Yen on infrared filters. The effects studied were those of variation in angle of incidence and temperature, and both theoretical and experimental results were quoted.

Accurate calculation of the effects of changes in the angle of incidence yielded a variation of peak wavelength of the expected form, but no significant variation of bandwidth for angles of incidence up to 50°. They also calculated that the peak transmittance and the shape of the pass band should remain unchanged for angles up to 45°. For angles above 50°, both the shape and the peak transmittance gradually deteriorated. The calculations were confirmed by measurements on real filters.

The effects of varying temperatures were also investigated both theoretically and practically. As in the case of the shorter wavelength filters examined by Blifford, they measured a shift towards longer wavelengths with increasing temperature. For temperatures down to liquid helium the filters show little loss of peak transmittance or variation of characteristic pass-band shape. However, serious losses in transmittance occurred above 50 °C. Although not mentioned in the paper, this is probably due to the use of germanium, either as substrate or one of the layer materials, which always exhibits a marked fall in transmittance at elevated temperatures above 50 °C. Baker and Yen make the point that filters designed to be least sensitive to variations in the angle of incidence are usually most sensitive to temperature and vice versa. The temperature coefficients of peak wavelength which they quote vary from $+0.0035\% \text{ }^{\circ}\text{C}^{-1}$ to $+0.0125\% \text{ }^{\circ}\text{C}^{-1}$. Unfortunately, neither the materials used in the filters nor the designs are quoted in the paper, but it is likely that the figures will apply to most interference filters for the infrared.

Similar measurements of the temperature shift of infrared filters were made at Grubb Parsons. The materials used were zinc sulphide and lead telluride, and the filters which had first-order high-index spacers gave temperature coefficients of peak wavelength of $-0.0135\% \text{ }^{\circ}\text{C}^{-1}$. These filters were of the type used in the selective chopper radiometer described in chapter 12. The negative temperature coefficient is usual with filters having lead telluride as one of the layer materials. This negative coefficient in lead telluride is especially useful as it tends to compensate for the positive coefficient in zinc sulphide, and Seeley *et al* [44] have succeeded in designing and constructing filters using lead telluride which have zero temperature coefficient.

References

- [1] Epstein L 1952 The design of optical filters *J. Opt. Soc. Am.* **42** 806–10
- [2] Turner A F 1950 Some current developments in multilayer optical films *J. Phys. Radium* **11** 443–60
- [3] Bates B and Bradley D J 1966 Interference filters for the far ultraviolet (1700 to 2400 Å) *Appl. Opt.* **5** 971–5
- [4] Seeley J S 1964 Resolving power of multilayer filters *J. Opt. Soc. Am.* **54** 342–6

- [5] Hemingway D J and Lissberger P H 1973 Properties of weakly absorbing multilayer systems in terms of the concept of potential transmittance *Opt. Acta* **20** 85–96
- [6] Dobrowolski J A 1959 Mica interference filters with transmission bands of very narrow half-widths *J. Opt. Soc. Am.* **49** 794–806 and 1963 Further developments in mica interference filters *J. Opt. Soc. Am.* **53** 1332 (summary only)
- [7] Austin R R 1972 The use of solid etalon devices as narrowband interference filters *Opt. Eng.* **11** 65–9
- [8] Candille M and Saurel J M 1974 Ralisation de filtres ‘double onde’ a bandes passantes tres etroites sur supports en matire plastique (mylar) *Opt. Acta* **21** 947–62
- [9] Smith S D and Pidgeon C R 1963 Application of multiple beam interferometric methods to the study of CO₂ emission at 15 μm *Mem. Soc. R. Sci. Liege 5ieme serie* **9** 336–49
- [10] Roche A E and Title A M 1974 Tilt tunable ultra narrow-band filters for high resolution photometry *Appl. Opt.* **14** 765–70
- [11] Dufour C and Herpin A 1954 Applications des methodes matricielles au calcul d’ensembles complexes de couches minces alternees *Opt. Acta* **1** 1–8
- [12] Lissberger P H 1959 Properties of all-dielectric filters. I—A new method of calculation *J. Opt. Soc. Am.* **49** 121–5
- [13] Lissberger P H and Wilcock W L 1959 Properties of all-dielectric filters. II—Filters in parallel beams of light incident obliquely and in convergent beams *J. Opt. Soc. Am.* **49** 126–38
- [14] Pidgeon C R and Smith S D 1964 Resolving power of multilayer filters in non-parallel light *J. Opt. Soc. Am.* **54** 1459–66
- [15] Hernandez G 1974 Analytical description of a Fabry–Perot spectrometer, 3. Off-axis behaviour and interference filters *Appl. Opt.* **13** 2654–61
- [16] For example, Reports 4, 5 and 6 of Contract DA-44-009-eng-1113 covering the period January–October 1953
- [17] Turner A F 1952 Wide pass band multilayer filters *J. Opt. Soc. Am.* **42** 878(a)
- [18] Smith S D 1958 Design of multilayer filters by considering two effective interfaces *J. Opt. Soc. Am.* **48** 43–50
- [19] Knittl Z 1965 Dielektrische Interferenzfilter mit rechteckigen Maximum *Proc. Coll. Thin Films (Budapest)* pp 153–61 (The method is described in detail in reference 20 also)
- [20] Knittl Z 1976 *Optics of Thin Films* (London: Wiley)
- [21] Thelen A 1966 Equivalent layers in multilayer filters *J. Opt. Soc. Am.* **56** 1533–8
- [22] Barr E E 1974 Visible and ultraviolet bandpass filters *Optical Coatings, Applications and Utilization* ed G W DeBell and D H Harrison *Proc. SPIE* **50** 87–118
- [23] Neilson R G T and Ring J 1967 Interference filters for the near ultra-violet *J. Phys.* **28** C2–270–5 (supplement to no 3–4 March–April)
- [24] Malherbe A 1974 Interference filters for the far ultraviolet *Appl. Opt.* **13** 1275–6
- [25] Baumeister P W and Jenkins F A 1957 Dispersion of the phase change for dielectric multilayers. Application to the interference filter *J. Opt. Soc. Am.* **47** 57–61
- [26] Baumeister P W, Jenkins F A and Jeppesen M A 1959 Characteristics of the phase-dispersion interference filter *J. Opt. Soc. Am.* **49** 1188–90
- [27] Giacomo P, Baumeister P W and Jenkins F A 1959 On the limiting band width of interference filters *Proc. Phys. Soc.* **73** 480–9
- [28] Ritchie F S Unpublished work on Ministry of Technology Contract

KX/LSO/C.B.70(a)

- [29] Hass G and Hadley L 1972 Optical constants of metals *American Institute of Physics Handbook* ed D E Gray (New York: McGraw-Hill) pp 6-124-56
- [30] Berning P H and Turner A F 1957 Induced transmission in absorbing films applied to band pass filter design *J. Opt. Soc. Am.* **47** 230-9
- [31] Baumeister P W, Costich V R and Pieper S C 1965 Bandpass filters for the ultraviolet *Appl. Opt.* **4** 911-13
- [32] MacDonald J 1971 *Metal-Dielectric Multilayers* (London: Adam Hilger)
- [33] Lissberger P H 1981 Coatings with induced transmission *Appl. Opt.* **20** 95-104
- [34] Maier R L 1967 2M interference filters for the ultraviolet *Thin Solid Films* **1** 31-7
- [35] Macleod H A 1978 A new approach to the design of metal-dielectric thin-film optical coatings *Opt. Acta* **25** 93-106
- [36] McKenney D B 1969 Ultraviolet interference filters with metal-dielectric stacks *PhD Dissertation* (Optical Services Center, University of Arizona)
- [37] Landau B V and Lissberger P H 1972 Theory of induced transmission filters in terms of concept of equivalent layers *J. Opt. Soc. Am.* **62** 1258-64
- [38] Blifford I H Jr 1966 Factors affecting the performance of commercial interference filters *Appl. Opt.* **5** 105-11
- [39] Baker M L and Yen V L 1967 The effect of the variation of angle of incidence and temperature on infrared filter characteristics *Appl. Opt.* **6** 1343-51
- [40] Pelletier F, Roche P and Bertrand L 1974 On the limiting bandwidth of interference filters: influence of temperature during production *Opt. Acta* **21** 927-46
- [41] Title A M, Pope T P and Andelin J P 1974 Drift in interference filters. 1 *Appl. Opt.* **13** 2675-9
- [42] Title A M 1974 Drift in interference filters. 2: radiation effects *Appl. Opt.* **13** 2680-4
- [43] Baumeister P W 1973 Thin films and interferometry *Appl. Opt.* **12** 1993-4
- [44] Seeley J S, Evans C S, Hunneman R and Whatley A 1976 Filters for the ν_2 band of CO₂; monitoring and control of layer deposition *Appl. Opt.* **15** 2736-45

Chapter 8

Tilted coatings

8.1 Introduction

We have already seen in chapter 2 that the characteristics of coatings change when they are tilted with respect to the incident illumination, and the particular way in which they change depends on the angle of incidence. We have studied the shifts that are induced in narrowband filters. Narrowband filters are a simple case because the tilt angle is usually small and we can assume that the major effect is in the phase thickness of the layers, which is affected equally for each plane of polarisation. For larger tilts, however, the admittances are also affected and then the performance for each plane of polarisation differs. Some important applications involve the difference in performance between one plane and the other, which can be controlled to some extent, making possible the construction of phase retarders and polarisers. On the other hand, the differences in performance can create problems, and although it is impossible to cancel the effects completely, there are ways of modifying it so that a more acceptable performance may be achieved. Then there are some, at first sight, strange effects which occur with dielectric-coated reflectors. Under certain conditions and at reasonably high angles of incidence, sharp absorption bands can exist for one plane of polarisation. This can create difficulties with dielectric-overcoated reflectors such as protected silver. The chapter begins with the addition of tilting effects to the admittance diagram, which allows us to explain qualitatively the behaviour of many different types of tilted coatings including overcoated reflectors and which involves a slight modification to the traditional form of the tilted admittances. Next there is a description of polarisers followed by an account of phase retarders. Some coatings where the polarisation splitting is undesirable, such as dichroic filters, are described with ways of reducing this splitting. Finally some antireflection coatings at high angles of incidence are described.

Some of the material in this chapter has already been mentioned and discussed in earlier chapters but here we attempt to introduce a consistent and

connected account and so there are some advantages in repeating what has been said before in the present context.

8.2 Modified admittances and the tilted admittance diagram

The form of the admittances and the phase thickness of a film which is illuminated at oblique incidence are given in chapter 2 and have already been used in considering the performance of some coatings including narrowband filters. They are:

$$\delta = 2\pi d(n^2 - k^2 - n_0^2 \sin^2 \theta_0 - 2ink)^{1/2}/\lambda \quad (8.1)$$

where the fourth quadrant solution is correct, and then

$$\eta_s = (n^2 - k^2 - n_0^2 \sin^2 \theta_0 - 2ink)^{1/2} \mathcal{Y} \quad (8.2)$$

again in the fourth quadrant, and

$$\eta_p = y^2/\eta_s \quad (8.3)$$

where n, k refer to the film and n_0, θ_0 etc to the incident medium. When the layers are purely dielectric then this is in the simpler form

$$\delta = (2\pi nd \cos \theta)/\lambda \quad (8.4)$$

$$\eta_s = y \cos \theta \quad (8.5)$$

and

$$\eta_p = y / \cos \theta \quad (8.6)$$

where $n \sin \theta = n_0 \sin \theta_0$. Expressions (8.4)–(8.6) can be used instead of expressions (8.1)–(8.3) if the $\cos \theta$ is permitted to become complex.

The calculation of multilayer properties at angles of incidence other than normal simply involves the use of the above expressions instead of those for normal incidence. It should be emphasised that the appropriate tilted values are to be adopted for incident medium and substrate as well as for the films. The use of the admittance diagram is rendered much more complicated because of the change in the incident admittance. The isoreflectance and isophase contours depend on the admittance of the incident medium and we therefore need one set for s-polarisation and one quite different set for p-polarisation, as well as completely new sets each time the angle of incidence is changed. Fortunately, there is a way round this problem, which carries some other advantages as well.

It has been shown by Thelen [1] that the properties of a multilayer are unaffected if all the admittances are multiplied or divided by a constant factor, and indeed it is usual to divide the admittances by \mathcal{Y} , the admittance of free space, so that the normal incidence admittance is numerically equal to the refractive index. We now propose an additional correction to the admittances, the dividing

of the s-polarised admittances, and the multiplying of the p-polarised admittances, by $\cos \theta_0$. This has the effect of preserving, for both s- and p-polarisation, the admittance of the incident medium at its normal incidence value, regardless of the angle of incidence, and means that the isoreflectance and isophase contours of the admittance diagram retain their normal incidence values whatever the angle of incidence or plane of polarisation. We can call these admittances simply the modified admittances, and the expressions for them become

$$\eta_s = (n^2 - k^2 - n_0^2 \sin^2 \theta_0 - 2ink)^{1/2} / \cos \theta_0 \quad (8.7)$$

again in the fourth quadrant, and

$$\eta_p = y^2 / \eta_s. \quad (8.8)$$

Or, when the layers are dielectric, the simpler forms are

$$\eta_s = (y \cos \theta) / \cos \theta_0 \quad (8.9)$$

and

$$\eta_p = (y \cos \theta_0) / \cos \theta. \quad (8.10)$$

The values of reflectance, transmittance, absorptance and phase changes on either transmission or reflection are completely unchanged by the adoption of these values for the admittances. Since the expressions involve $\cos \theta_0$ and $\cos \theta$, which are connected by the admittance of the incident medium, then the dependence of the modified admittances on the index of the incident medium will be somewhat different from the unmodified, traditional ones. Nevertheless, we shall see that this does carry some advantages.

We consider first of all purely dielectric materials. In this case, provided that $n_0 \sin \theta_0$ is less than n , the film index, then the two values for the modified admittances are real and positive. If, however, n_0 is greater than n , then there is a real value of θ_0 at which $n_0 \sin \theta_0$ is equal to n . This angle is known as the critical angle, and, for angles of incidence greater than this value, the admittances are imaginary. We will consider what happens for angles of incidence beyond critical later. First we will limit ourselves to angles less than critical where the admittances are real.

First of all, let us consider air of index unity as the incident medium. We recall that all transparent thin-film materials have refractive index greater than unity. In figure 8.1 the modified admittance is shown for a number of thin-film materials as a function of angle of incidence. The p-admittances of all materials cross the line $n = 1$ at the value known as the Brewster angle for which the single-surface p-reflectance is zero. The s-admittances all increase away from the line $n = 1$, so that the single-surface s-reflectance simply increases with angle of incidence. Since all these materials are dielectric, their modified optical thickness is real and therefore, although a correction has to be made for the effect of angle of incidence, quarter- and half-wave layers can be produced at non-normal incidence

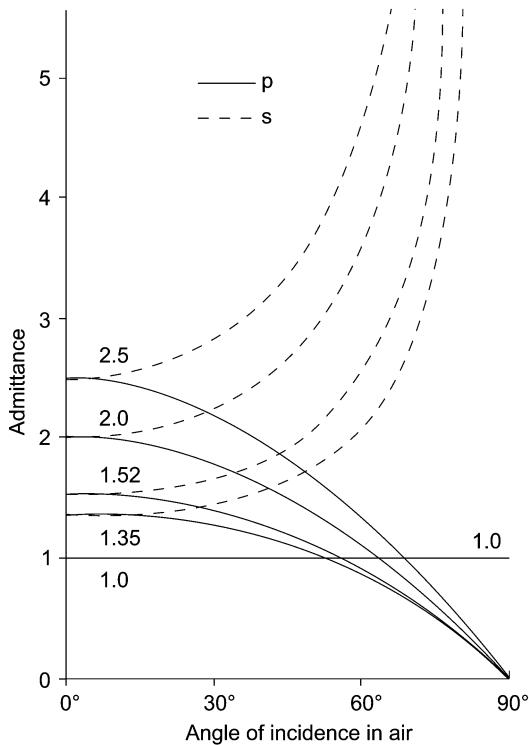


Figure 8.1. Modified p- and s-admittances (i.e. including the extra factor of $\cos \theta_0$) of materials of indices 1.0, 1.35, 1.52, 2.0 and 2.5 for an incident medium of index 1.0.

just as readily as at normal and it cannot be too greatly emphasised that although the optical thickness changes with angle of incidence, it does not vary with the plane of polarisation.

It is possible to make several deductions directly from figure 8.1. The first is that, for any given pair of indices, the ratio of the s-admittances increases with angle of incidence, while that for p-admittances reduces. Since the width of the high-reflectance zone of a quarter-wave stack decreases with decreasing ratio of these admittances, the width will be less for p-polarised light than for s-polarised. As we shall shortly see, this effect is used in a useful type of polariser. The splitting of the admittance of dielectric layers means also that there is a relative phase shift between p- and s-polarised light reflected from a high-reflectance coating when the layers depart from quarter-waves. This effect can be used in the design of phase retarders and we will give a brief account of this. The diagram also helps us to consider the implications of antireflection coatings for high angles of incidence. A frequent requirement is an antireflection coating for a crown glass of index around 1.52. For a perfect single-layer coating we should have a quarter-

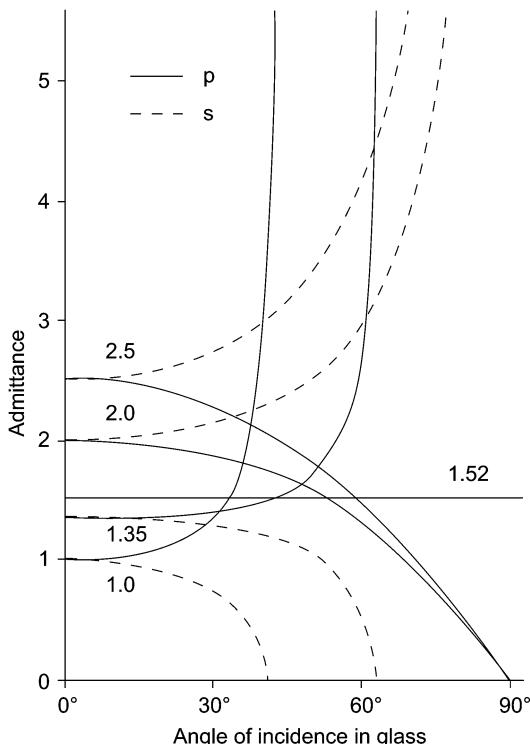


Figure 8.2. Modified p- and s-admittances (i.e. including the extra factor of $\cos \theta_0$) of materials of indices 1.0, 1.35, 1.52, 2.0 and 2.5 for an incident medium of index 1.52.

wave of material of optical admittance equal to the square root of the product of the admittances of the glass and the incident medium. At normal incidence in air there is, of course, no sufficiently robust material with index as low as 1.23. For greater angles of incidence, the s-polarised reflectance increases still further from its normal incidence value and the admittance required for a perfect single-layer antireflection coating remains outside the range of practical materials, corresponding to still lower indices of refraction. The p-polarised behaviour is, however, completely different, and in the range from approximately 50° – 70° the admittance required for the antireflection coating is within the range of what is possible. No coating is required, of course, at the Brewster angle. For angles greater than the Brewster angle, the index required is *greater* than that of the glass. Antireflection coatings for high angles of incidence will also be discussed shortly.

The behaviour of dielectric materials when the incident medium is of a higher index (one that is within the range of available thin-film materials) is somewhat more complicated. Figure 8.2 shows the way in which the admittances

vary when the incident medium is glass of index 1.52. There is the familiar splitting of the s- and p-polarised admittances which, as before, increases with angle of incidence. For indices which are lower than that of the glass it is possible to reach the critical angle, and at that point the admittances reach either zero or infinity and disappear from the diagram. Their behaviour beyond the critical angle will be discussed shortly. A further very important feature is that, while for indices higher than that of the incident medium the p-polarised admittance falls with angle of incidence, for indices lower than the incident medium the p-polarised admittance rises. All cut the incident medium admittance at the Brewster angle, but now a new phenomenon is apparent. The p-admittance curves for materials of index lower than that of the incident medium intersect the curves corresponding to higher indices. An immediate deduction is that a quarter-wave stack, composed of such pairs of materials, will simply behave, at the angle of incidence corresponding to the point of intersection, as a thick slab of material. Provided the admittances of substrate, thin films and incident medium are not too greatly different, the p-reflectance will be low. The ratio of the s-admittances is large, because their splitting increases with angle of incidence, and so the corresponding s-reflectance is high and the width of the high-reflectance zone is large. This is the basic principle of the MacNeille polarising beam splitter that we will return to in a later section. The range of useful angles of incidence will depend partly on the rate at which the curves of p-polarised admittance diverge on either side of the intersection, and this can be estimated from the diagram.

Apart from the polarisation-splitting of the admittance, the behaviour of dielectric layers at angles of incidence less than critical is reasonably straightforward and does not involve difficulties of a more severe order than exist at normal incidence. When metal films are introduced, however, the difficulties increase and the behaviour becomes still stranger when combined with dielectric materials, especially when used beyond the critical angle. The aim in the remainder of this section is to discuss, in a qualitative fashion, such behaviour and to suggest techniques which can be used for visualisation and prediction. The use of admittance loci will be emphasised.

We know already that the admittance locus of a dielectric layer at normal incidence is a circle centred on the real axis. Tilted dielectric layers at angles of incidence less than critical still have circular loci which can be calculated from the tilted admittances in exactly the same way. Provided the modified admittances are used in constructing the loci then the isoreflectance and isophase circles on the admittance diagram will remain exactly the same as at normal incidence for both p- and s-polarisation.

The admittance of a metal layer is a little more complicated than a dielectric. For a lossless metal in which the refractive index, and hence the optical admittance, is purely imaginary, and given by $-ik$, the loci are a set of circles with centres on the real axis and passing through the points ik and $-ik$, which are on the imaginary axis. Figure 8.3 shows the typical form. The circles are like the dielectric ones, traced out clockwise so that they start on ik and end on

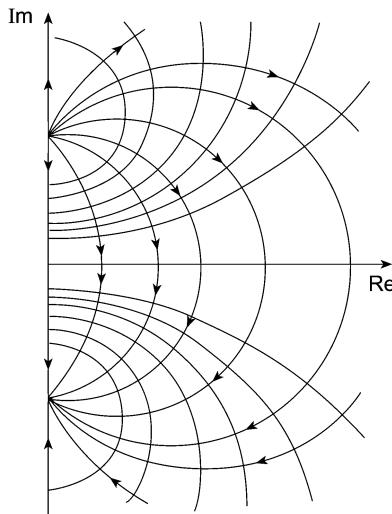


Figure 8.3. Admittance loci for an ideal metal with admittance $-ik$. The loci begin at the point ik and terminate on $-ik$. Equi-thickness contours are also shown at no fixed intervals. Similar loci are obtained for s-polarised frustrated total reflectance (FTR) layers. For p-polarised FTR layers, the shape of the loci is similar but they are traced in the opposite direction.

$-ik$. Real metallic layers depart somewhat from this ideal model but if the metal is of high performance, i.e. if the ratio k/n is high, then the loci are similar to the perfect case. It is as if the diagram were rotated slightly about the origin so that the points where all circles intersect are $(n, -k)$ and $(-n, k)$ respectively, although the circles can never reach the point $(-n, k)$ since admittance loci are constrained to the first and second quadrants of the Argand diagram. Figure 8.4 shows a set of optical admittance loci calculated for silver, $n - ik = 0.075 - i3.41$ [2] demonstrating this typical behaviour. The direction of the loci is now better described as terminating on $(n, -k)$, although most are still described in a clockwise direction. We will omit from the discussion in this chapter metals which are not of high optical quality and for which the loci resemble a set of spirals terminating at $(n, -k)$. What happens at oblique incidence?

The optical phase factor at normal incidence is

$$2\pi(n - ik)d/\lambda \quad (8.11)$$

dominated by the imaginary part. At oblique incidence, it becomes

$$2\pi(n^2 - k^2 - n_0^2 \sin^2 \theta_0 - 2ink)^{1/2}d/\lambda \quad (8.12)$$

still in the fourth quadrant. Since $n_0 \sin \theta_0$ is normally small compared with k , it has little effect on the phase factor. It reduces the real part slightly and increases

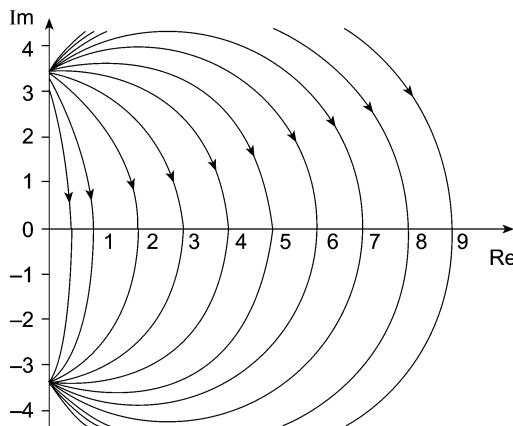


Figure 8.4. Admittance loci for silver at normal incidence in the visible region. The value assumed for the optical constants is $0.075 - i3.51$ [2].

the imaginary part, but the effect is small, and the behaviour is essentially similar to that at normal incidence. At an angle of incidence of 80° in air, for example, the phase factor of silver changes from $2\pi(0.075 - i3.41)d/\lambda$ to $2\pi(0.00721 - i3.549)d/\lambda$. The change in the modified admittance, therefore, is mainly due to the $\cos \theta_0$ term. The ratio of real to imaginary parts remains virtually the same, and the p-admittance simply moves towards the origin (both real and imaginary parts reduced) and the s-admittance away from the origin. Thus the principal effect for high-performance metal layers with tilt is an expansion of the circular loci for s-polarisation and a contraction for p-polarisation. The basic form remains the same.

The shift in the modified optical admittance does mean that the phase shift on reflection from a massive metal will vary. For silver at normal incidence, the phase shift will be in the second quadrant. As the angle of incidence increases, the movement of the p-polarised admittance towards the origin implies that the p-polarised phase shift moves towards the first quadrant, entering it at an angle of incidence of just above 70° (i.e. roughly $\cos^{-1} \frac{1}{3}$) while the s-polarised phase shift moves further towards 180° . The reflectance for s-polarised light increases, while for p-polarised light it shows a very slight drop initially to a shallow minimum, but rising thereafter.

Now we examine what happens when a metal layer is overcoated with a dielectric layer. The arrangement is sketched schematically in figure 8.5. Provided the admittance η_f of the dielectric layer is less than $(\eta_m \eta_m^*)^{1/2}$, where η_m is the admittance of the metal layer, the admittance locus will loop outside the line joining the origin to the starting point, as in the diagram. For dielectric layers having admittance greater than that of the incident medium, the reflectance falls while the locus is in the fourth quadrant of the Argand diagram. As the thickness

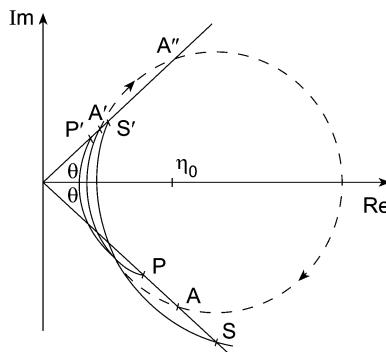


Figure 8.5. Schematic diagram of a dielectric overcoat on a metal surface. At normal incidence the metal admittance is at point A. A' represents a quarter-wave thickness of material, while A'' represents the point at which the reflectance returns to the starting value. The lowest reflectance is given by the intersection with the real axis between the points A and A'. When tilted, the p-locus is given by PP' and the s-locus by SS'.

of the dielectric layer increases, the reflectance is reduced until the intersection with the real axis. It then begins to rise, but, at the quarter-wave point A' given by η_f^2/η_m , it is still below the reflectance of the bare metal. Only at point A'' does the reflectance return to its initial level. The drop in reflectance for silver is slight, but for aluminium it is catastrophic. Silver is therefore usually overcoated with a quarter-wave, but aluminium with a half-wave that limits its useful spectral range somewhat.

As the metal–dielectric combination is tilted, the p-admittance of the metal slides towards the origin, the reflectance dropping, while the s-admittance moves away from the origin with a rise in reflectance. The dielectric layer shows a drop in admittance for p-polarised light and an increase for s-polarised. For dielectric coatings that are a quarter-wave or less these changes tend to compensate, and indeed, in silver, slightly overcompensate, the changes in reflectance of the bare metal. The p-reflectance of the overcoated metal tends to be slightly higher than the s-reflectance.

Eventually, for very high angles of incidence, the p-polarised admittance of the dielectric layer falls below the admittance of the incident medium, and now the fourth quadrant portion of the locus represents increasing reflectance. This means that the dielectric overcoating, when thin, instead of reducing the reflectance of the metal, actually enhances it. Thus, depending on the final thickness of the dielectric layer, the reflectance will tend to be high. For s-polarised light, the admittance of the dielectric layer tends to infinity as the angle of incidence tends to 90° . The locus of the dielectric overcoat, therefore, tends more and more towards a vertical line. As the admittance of the metal moves away from the origin, its projection in the real axis moves further to the right, eventually crossing

the incident medium admittance and continuing towards infinity. There must, therefore, be an angle of incidence, very high, where the locus of the dielectric overcoat will intersect the real axis at the admittance of the incident medium. If the thickness is chosen so that the locus terminates at this point, then the reflectance of the metal–dielectric combination will be zero. This will occur for one particular value of angle of incidence and for a precise value of the dielectric layer thickness, and the dip in reflectance will show a rapid variation with angle of incidence. Such behaviour, for s-polarised light, of a metal overcoated with a thin dielectric layer was predicted by Nevière and Vincent [3] from a quite different analysis based on a Brewster absorption phenomenon in a lossy waveguide used just under its cutoff thickness. Since the modified admittance for s-polarised light increases with angle of incidence only in the case where its refractive index is greater than that of the incident medium, this is a necessary condition for the observation of the effect. The increased flexibility given by two dielectric layers deposited on a metal has been used to advantage in the design of reflection polarisers [4].

A different phenomenon was observed by Cox *et al* [5] in connection with an infrared mirror of aluminium with a protective overcoat of silicon dioxide. The silicon dioxide is heavily absorbing in the region beyond $8 \mu\text{m}$. At a wavelength of just over $8 \mu\text{m}$, n and k have values around 0.4 and 0.3 respectively. At normal incidence, the admittance loci of the silicon dioxide are spirals which end on the admittance of the silicon dioxide and are described in a clockwise manner in much the same way as the silver loci already discussed. At non-normal incidence, the s-polarised admittance and the phase factor for the layer remain in the fourth quadrant, and so the behaviour of the silicon oxide is similar to that at normal incidence. The p-polarised admittance, however, moves towards the first quadrant, and enters it at an angle of incidence around 40° . The behaviour of such a material, where the phase thickness is in the fourth quadrant but the optical admittance is in the first, is different from normal materials in that the spirals are now traced out anticlockwise, rather than clockwise. The admittance of aluminium at $8.1 \mu\text{m}$ is around $18.35 - i55.75$ and, for p-polarised light at an angle of incidence of 60° , the modified admittance becomes $9.176 - i27.87$. The dielectric locus sweeps down towards the real axis, as in figure 8.6, and, in a thickness of 150 nm, terminates in the vicinity of the point $(1, 0)$, so that the reflectance is near zero.

This behaviour is quite unlike the normal behaviour to be expected with lossless dielectric overcoats which have refractive index greater than that of the incident medium. However, we shall see that it does have a certain similarity with one of the techniques for generating surface electromagnetic waves, which we shall be dealing with shortly, where the coupling medium is a dielectric layer of index lower than that of the incident medium, and where the angle of incidence is beyond the critical angle.

We now turn back to dielectric materials and investigate what happens when angles of incidence exceed the critical angle. Equations (8.7), (8.8) and (8.12) are

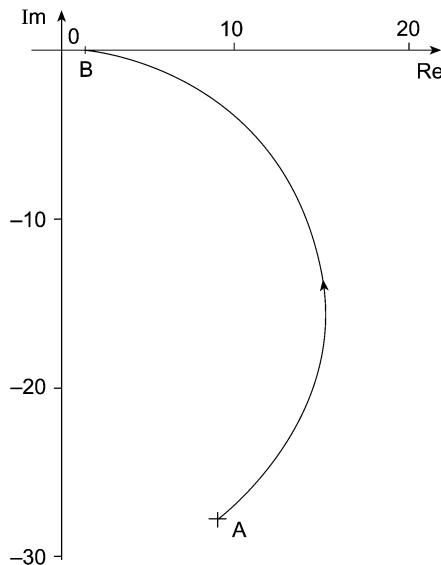


Figure 8.6. p-polarised admittance locus for 150 nm thickness of SiO_2 , $0.39 - i0.29$, on aluminium, $18.35 - i55.75$, at an angle of incidence of 60° . A is the point corresponding to the modified admittance of aluminium and the anticlockwise curvature of the spiral locus carries it into the region of low reflectance.

the relevant equations and we have $k = 0$ and $n_0 \sin \theta_0 > n$. The phase thickness at normal incidence, $2\pi n d / \lambda$, becomes, from equation (8.12),

$$2\pi(n^2 - n_0^2 \sin^2 \theta_0)^{1/2} d / \lambda$$

i.e.

$$-i2\pi(n_0^2 \sin^2 \theta_0 - n^2)^{1/2} d / \lambda \quad (8.13)$$

at oblique incidence, where, again, the fourth rather than second quadrant solution is correct. The modified admittances are then

$$\begin{aligned} \eta_s &= -i(n_0^2 \sin^2 \theta_0 - n^2)^{1/2} / \cos \theta_0 && \text{(fourth quadrant)} \\ \eta_p &= n^2 / \eta_s. \end{aligned} \quad (8.14)$$

Since η_s is negative imaginary, η_p must be positive imaginary. The behaviour of the modified admittance is shown diagrammatically in figure 8.7. For a thin film of material used beyond the critical angle, then, the s-polarised behaviour is indistinguishable from that of an ideal metal. We have a set of circles centred on the real axis, described clockwise and ending on the point η_s which is on the negative imaginary axis. For p-polarised light, the behaviour is, in one important respect, different. Here, the combination of negative imaginary phase

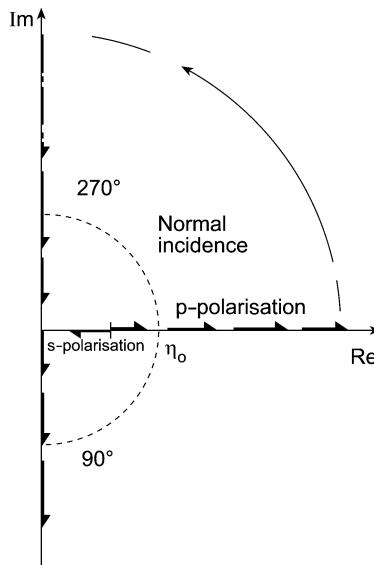


Figure 8.7. The variation of the s-polarised and p-polarised modified admittances of free space with respect to an incident medium of higher index. η_0 is the incident admittance. The s-admittance falls along the real axis until zero at the critical angle and then it turns along the negative direction of the imaginary axis tending to negative imaginary infinity as the angle of incidence tends to 90° . The p-admittance rises along the real axis, passing the point η_0 at the Brewster angle, becoming infinite at the critical angle, switching over to positive imaginary infinity and then sliding down the imaginary axis tending to zero as the angle of incidence tends to 90° .

thickness and positive imaginary admittance inverts the way in which the circles are described, so that although they are still centred on the origin, they are anticlockwise and terminate at η_p on the positive imaginary axis. This behaviour plays a significant part in what follows. We assume a beam of light incident on the hypotenuse of a prism beyond the critical angle. Simply for plotting some of the following figures, we assume a value for the index of the incident medium of 1.52.

For an uncoated hypotenuse, the second medium is air of refractive index unity. The modified admittance for p-polarised light is positive imaginary and, as θ_0 increases, falls down the imaginary axis towards the origin. The reflectance is unity and figure 8.7 shows that the phase shift varies from 180° through the third and fourth quadrants towards 0° . The s-polarised reflectance is likewise unity, but the admittance is negative imaginary, and falls from zero to infinity along the imaginary axis so that the s-polarised phase shift increases with θ_0 from zero, through the first and second quadrants towards 180° . Since the incident medium

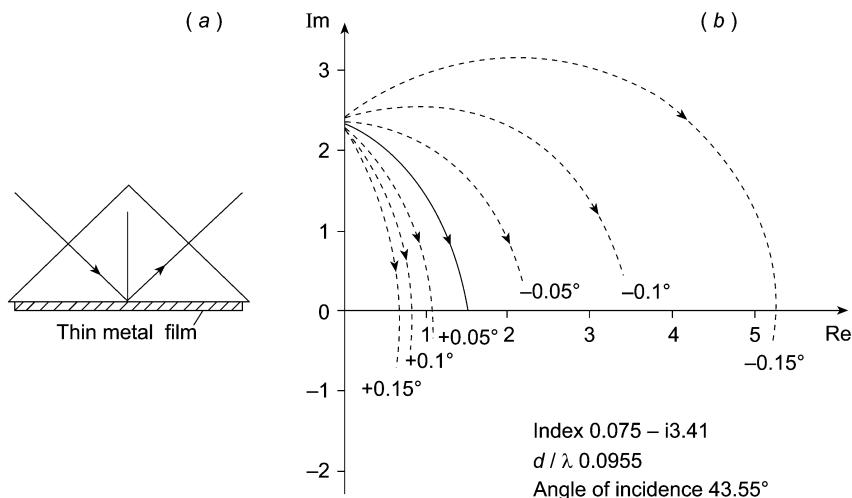


Figure 8.8. (a) Coupling to a surface plasma wave. (After Kretschmann and Raether [8].) (b) p-polarised admittance locus corresponding to the arrangement in (a). The solid curve corresponds to the optimum angle of incidence and thickness of metal (silver) film. The dashed curves correspond to changes in the angle of incidence as marked on each curve.

has admittance 1.52, the circle separating the first and second quadrants and the third and fourth quadrants, which has centre the origin, has radius 1.52.

Now let a thin film be added to the hypotenuse. Since we are treating our glass prism as the incident medium, we should treat the surrounding air as the substrate. Thus the starting admittance for the film is on the imaginary axis. Provided the thin film has no losses, then the admittance of the film–substrate combination must remain on the imaginary axis. If the film admittance is imaginary, the combination admittance will simply move towards the film admittance. If, however, the film admittance is real, the admittance of the combination will move along the imaginary axis in a positive direction, returning to the starting point every half-wave. The lower the modified admittance, the slower the locus moves in the vicinity of the origin and the faster at points far removed from the origin. The variation of phase change between the fourth quadrant and the start of the first quadrant is, therefore, slower, while that between the third and second quadrants is faster than for a higher admittance. Thus there is a wide range of possibilities for varying the relative phase shifts for p- and s-polarisations by choosing an overcoat of higher or lower index and varying the thickness [6, 7].

Given that the starting point is on the axis, then the only way in which the admittance can be made to leave it is by an absorbing layer. We turn to the set of metal loci (figure 8.4) and we can see that for a range of values of starting admittance on the imaginary axis, the metal loci loop around, away from the axis,

to cut the real axis. Although figure 8.4 shows the behaviour of metal layers for an incident medium of unity at normal incidence, the tilted behaviour for an incident admittance of 1.52 is quite similar. Figure 8.8 shows the illuminating arrangement and the loci. For a very narrow range of starting values, the metal locus cuts the real axis in the vicinity of the incident admittance, and, if the metal thickness is such that the locus terminates there, then the reflectance of the combination will be low. For one particular angle of incidence and metal thickness the reflectance will be zero. It should not be too much of a surprise to find that the condition is very sensitive to angle of incidence. Since the admittance of the metal varies much more slowly than the air substrate, the zero reflectance condition will no longer hold, even for quite small tilts. This very narrow drop in reflectance to a very low value, which has all the hallmarks of a sharp resonance, can be interpreted as the generation of a surface plasma wave, or plasmon, on the metal film. This coupling arrangement, devised by Kretschmann and Raether [8], cannot operate for s-polarised light without modification. The admittance of the substrate for s-polarisation is now on the negative part of the real axis and, therefore, any metal which is deposited will simply move the admittance of the combination towards the admittance of the bulk metal.

An alternative coupling arrangement, devised by Otto [9], involves the excitation of surface waves through an evanescent wave in an FTR layer (frustrated total reflectance). We recall that the admittance locus for p-polarisation of a layer used beyond the critical angle is a circle which is described in an anticlockwise direction. This means that such a layer can be used to couple into a massive metal. Here the metal acts as the substrate, with a starting admittance in the fourth quadrant of the Argand diagram. For p-polarised light, the dielectric FTR layer has a circular locus which cuts the real axis. Clearly, then, for the correct angle of incidence and dielectric layer thickness, the reflectance can be made zero. Surface plasma oscillations and their applications are extensively reviewed by Raether [10]. Abelès [11] includes an account of the optical features of such effects in his review of the optical properties of very thin films.

Now let us return to the first case of coupling and let us examine what happens when a thin layer is deposited over the metal next to the surrounding air. The starting admittance is, as before, on the imaginary axis, but now the dielectric layer modifies that position, so that the starting point for the metal locus is changed. Because the metal loci at the imaginary axis are clustered closely together, almost intersecting, a small change in starting point produces an enormous change in the locus, and hence in the point at which it cuts the real axis, leading to a substantial change in reflectance (figure 8.9). This very large change which a thin external dielectric film makes to the internal reflectance of the metal film has been used in the study of contaminant films adsorbed on metal surfaces. Film thicknesses of a few ångstroms have been detected in this way. Provided that the film is very thin, then an additional tilt of the system will be sufficient to pull the intersection of the metal locus with the real axis back to the incident admittance, and so the effect can be interpreted as a shift in the resonance

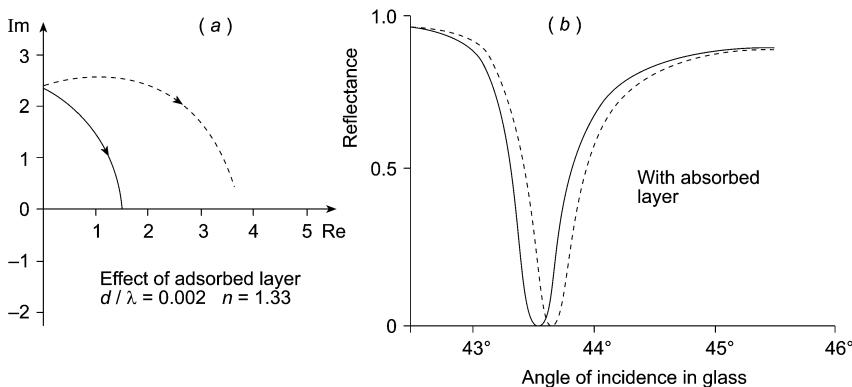


Figure 8.9. (a) The effect of a thin adsorbed layer on the surface of the silver in figure 8.8. The solid line is the optimum while the dashed line is the change in the metal locus due to the adsorbed layer. (b) Calculated reflectance as a function of angle of incidence with and without the adsorbed layer.

rather than a damping.

This result helps us to devise a method for exciting a similar resonance with s-polarised light. The essential problem is the starting point on the negative imaginary axis, which means that the subsequent metal locus remains within the fourth quadrant, never crossing the real axis to make it possible to have zero reflectance. The addition of a dielectric layer between the metal surface and the surrounding air can move the starting point for the metal on to the positive part of the imaginary axis so that the coated metal locus can cut the real axis for s-polarised light in just the same way as the uncoated metal in p-polarised light. Moreover, for both p- and s-polarised light, the low reflectance will be repeated for each additional half-wave dielectric layer which is added. This behaviour was used by Greenland and Billington [12] for the monitoring of optical layers intended as spacer layers for metal–dielectric interference filters. The operation of the cavities for inducing absorption devised by Harrick and Turner [13], although designed on the basis of a different approach, can also be explained this way.

8.3 Polarisers

8.3.1 The Brewster angle polarising beam splitter

This type of beam splitter was first constructed by Mary Banning [14] at the request of S M MacNeille, the inventor of the device [15] which is frequently known as a MacNeille polariser.

The principle of the device is that it is always possible to find an angle of incidence so that the Brewster condition for an interface between two materials

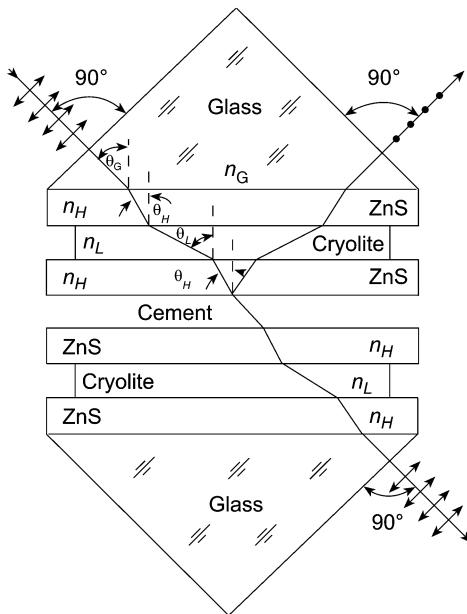


Figure 8.10. Schematic diagram of a polarising beam splitter. (After Banning [14].)

of differing refractive index is satisfied. When this is so, the reflectance for the p-plane of polarisation vanishes. The s-polarised light is partially reflected and transmitted. To increase the s-reflectance, retaining the p-transmittance at or very near unity, the two materials may then be made into a multilayer stack. The layer thickness should be quarter-wave optical thicknesses at the appropriate angle of incidence.

When the Brewster angle for normal thin-film materials is calculated, it is found to be greater than 90° referred to air as the incident medium. In other words, it is beyond the critical angle for the materials. This presents a problem which is solved by building the multilayer filter into a glass prism so that the light can be incident on the multilayer at an angle greater than critical. The type of arrangement is shown in figure 8.10.

The calculation of the design is quite straightforward. Consider two materials with refractive indices n_H and n_L (where H and L refer to high and low relative indices respectively). The Brewster condition is satisfied when the angle of incidence is such that

$$n_H / \cos \theta_H = n_L / \cos \theta_L \quad (8.15)$$

where

$$n_H \sin \theta_H = n_L \sin \theta_L = n_G \sin \theta_G. \quad (8.16)$$

G refers to the glass of the prism. These equations can be solved easily for θ_H

$$\sin^2 \theta_H = \frac{n_L^2}{n_H^2 + n_L^2} \quad (8.17)$$

the form in which we shall use the result. (A more familiar form is $\tan^2 \theta_H = n_L^2/n_H^2$.)

Given the layer indices there are two possible approaches to the design. Either we can decide on the refractive index of the glass and then calculate the angle at which the prism must be set, or we can decide on the prism angle, 45° being a convenient figure, and calculate the necessary refractive index of the glass. The approach which was used by Banning was the latter.

First suppose that the condition $\theta_G = 45^\circ$ must be met. Using equations (8.16) and (8.17) we obtain

$$\sin^2 \theta_H = \frac{n_G^2 \sin^2 \theta_G}{n_H^2} = \frac{1}{2} \frac{n_G^2}{n_H^2} \quad \text{for } \theta_G = 45^\circ$$

i.e.

$$n_G^2 = \frac{2n_H^2 n_L^2}{n_H^2 + n_L^2} \quad (8.18)$$

the condition obtained by Banning.

If, however, n_G is fixed, then equations (8.16) and (8.17) give

$$\frac{n_G^2 \sin^2 \theta_G}{n_H^2} = \sin^2 \theta_H = \frac{n_L^2}{n_H^2 + n_L^2}$$

i.e.

$$\sin^2 \theta_G = \frac{n_H^2 n_L^2}{n_G^2 (n_H^2 + n_L^2)}. \quad (8.19)$$

Banning used zinc sulphide with an index of 2.30 and cryolite evaporated at a pressure of 10^{-3} Torr to give a porous layer of index around 1.25. With these indices it is necessary to have an index of 1.55 for the glass if the prism angle is to be 45° . For an index of 1.35, a more usual figure for cryolite, together with zinc sulphide with an index of 2.35, the glass index should be 1.66. Alternatively, for glass of index 1.52, the angle of incidence using the second pair of materials should be 50.5.

The degree of polarisation at the centre wavelength can also be calculated.

$$R = \left(\frac{\eta_G - (\eta_H^2/\eta_G)(\eta_H/\eta_L)^{n-1}}{\eta_G + (\eta_H^2/\eta_G)(\eta_H/\eta_L)^{n-1}} \right)^2 \quad (8.20)$$

where n is the number of layers and we are assuming n to be odd.

For s-waves	For p-waves :
$\eta_G = n_G \cos \theta_G$	$\eta_G = n_G / \cos \theta_G$
$\eta_H = n_H \cos \theta_H$	$\eta_H = n_H / \cos \theta_H$
$\eta_L = n_L \cos \theta_L$	$\eta_L = n_L / \cos \theta_L.$

Now, for p-waves, by the condition we have imposed, $\eta_H = \eta_L$ and

$$\begin{aligned} R_p &= \left(\frac{\eta_G - (\eta_H^2 / \eta_G)}{\eta_G + (\eta_H^2 / \eta_G)} \right)^2 \\ &= \left[\left(\frac{n_G^2 \cos^2 \theta_H}{n_H^2 \cos^2 \theta_G} - 1 \right) \left(\frac{n_G^2 \cos^2 \theta_H}{n_H^2 \cos^2 \theta_G} + 1 \right)^{-1} \right]^2. \end{aligned} \quad (8.21)$$

Similarly,

$$R_s = \left(\frac{n_G^2 \cos^2 \theta_G - n_H^2 \cos^2 \theta_H (n_H \cos \theta_H / n_L \cos \theta_L)^{n-1}}{n_G^2 \cos^2 \theta_G + n_H^2 \cos^2 \theta_H (n_H \cos \theta_H / n_L \cos \theta_L)^{n-1}} \right)^2. \quad (8.22)$$

Now

$$\frac{n_H \cos \theta_L}{n_L \cos \theta_H} = 1$$

so that

$$\frac{n_H \cos \theta_H}{n_L \cos \theta_L} = \frac{n_H^2}{n_L^2}$$

and

$$R_s = \left(\frac{n_G^2 \cos^2 \theta_G - n_H^2 \cos^2 \theta_H (n_H / n_L)^{2(n-1)}}{n_G^2 \cos^2 \theta_G + n_H^2 \cos^2 \theta_H (n_H / n_L)^{2(n-1)}} \right)^2. \quad (8.23)$$

The degree of polarisation in transmission is given by

$$P_T = \frac{T_p - T_s}{T_p + T_s} = \frac{1 - R_p - 1 + R_s}{1 - R_p + 1 - R_s} = \frac{R_s - R_p}{1 - R_p - R_s} \quad (8.24)$$

and in reflection by

$$P_R = \frac{R_s - R_p}{R_s + R_p}. \quad (8.25)$$

It can be seen that in general, for a small number of layers, the polarisation in reflection is better than the polarisation in transmission, but for a large number of layers it is inferior to that in transmission.

The construction of the beam splitter is similar to the cube beam splitter which was considered in chapter 4. Any number of layers can be used in the stack. Banning's original stack consisted of three layers, probably because of practical difficulties at that time. Two stacks were therefore prepared, one on the hypotenuse of each prism making up the cube, as shown in figure 8.10. The two prisms were then cemented together. Nowadays there is little difficulty in depositing 21 layers or more if need be and this can be conveniently deposited on just one prism and the other untreated prism simply cemented to it.

The very great advantage which this type of polarising beam splitter has over the other polarisers such as the pile-of-plates is its wide spectral range coupled with a large physical aperture. Unfortunately, it does suffer from a limited angular field, particularly at the centre of its range, simply because the Brewster condition is met exactly only at the design angle. As the angle of incidence moves away from this condition, a residual reflectance peak for p-polarisation gradually appears in the centre of the range. The performance well away from the centre remains high even for quite large tilts away from optimum. As an example, we can consider a seven-layer ZnS and cryolite beam splitter in glass of index 1.52 designed so that a wavelength of 510 nm corresponds to the centre of the range. At the design angle of 50.4° and at 510 nm the residual p-reflectance is 1.6%, due to the mismatch between the materials of the stack and the glass prism. (The Brewster angle condition cannot be met for both film materials and the substrate simultaneously—see figure 8.2.) A tilt in the plane of incidence to 55° in glass (that is a tilt to 7° in air) raises the reflectance to 25% at 510 nm and over 30% at 440 nm, since the band centre moves to shorter wavelengths. The reflectance at 650 nm, on the other hand, shows little change. Skew rays present a further difficulty. Polarisation performance is measured with reference to the s- and p-directions associated with the principal plane of incidence containing the axial ray. A skew ray possesses a plane of incidence that is rotated with respect to the principal plane. Thus the s- and p-planes for skew rays are not quite those of the axial ray and although the s-polarised transmittance can be very low there can be a component of the p-polarised light, which is parallel to the axial s-direction and which can represent an appreciable leakage.

A detailed study of the polarising prism has been carried out by Clapham [16].

8.3.2 Plate polariser

The width of the high-reflectance zone of a quarter-wave stack is a function of the ratio of the admittances of the two materials involved. This ratio varies with the angle of incidence and is different for s- and p-polarisations. We recall that

$$\eta_s = n \cos \theta \quad \text{while} \quad \eta_p = n / \cos \theta$$

so that

$$\eta_{Hs}/\eta_{Ls} = \cos \theta_H / \cos \theta_L$$

and

$$\eta_{H_p}/\eta_{L_p} = \cos \theta_L / \cos \theta_H$$

whence

$$\frac{(\eta_H/\eta_L)_s}{(\eta_H/\eta_L)_p} = \frac{(\cos \theta_H)^2}{(\cos \theta_L)^2}. \quad (8.26)$$

The factor $(\cos \theta_H)^2/(\cos \theta_L)^2$ is always less than unity so that the width of the high-reflectance zone for p-polarised light is always less than that for s-polarised light. Within the region outside the p-polarised but inside the s-polarised high-reflectance zone, the transmittance is low for s-polarised light but high for p-polarised so that the component acts as a polariser. The region is quite narrow, so that such a polariser will not operate over a wide wavelength range; but for single wavelengths, such as a laser line, it can be very effective. To complete the design of the component it is necessary to reduce the ripple in transmission for p-polarised light and this can be performed using any of the techniques of chapter 6, probably the most useful being Thelen's shifted-period method because it is the performance right at the edge of the pass region which is important. It is normal to use the component as a longwave-pass filter because this involves thinner layers and less material than would a shortwave-pass filter. The rear surface of the component requires an antireflection coating for p-polarised light. We can omit this altogether if the component is used at the Brewster angle. The design of such a polariser is described by Songer [17] who gives the design shown in figure 8.11. Plate polarisers are used in preference to the prism or MacNeille type when high powers are concerned

Virtually any coating which possesses a sharp edge between transmission and reflection can potentially be used as a polariser. It has been suggested that narrowband filters have advantages over simple quarter-wave stacks as the basis of plate polariser coatings, because the monitoring of the component during deposition is a more straightforward procedure [18].

8.3.3 Cube polarisers

An advantage of the polariser immersed in a prism is that the effective angle of incidence can be very high—much higher than if the incident medium were air. This enhances the polarisation splitting and gives broader regions of high degree of polarisation than could be the case with air as the incident medium. Even if the Brewster angle condition cannot be reached, there is an advantage in using an immersed design, provided the incident power is not too high. Netterfield [19] has considered the design of such polarisers in some detail and his paper should be considered for further information.

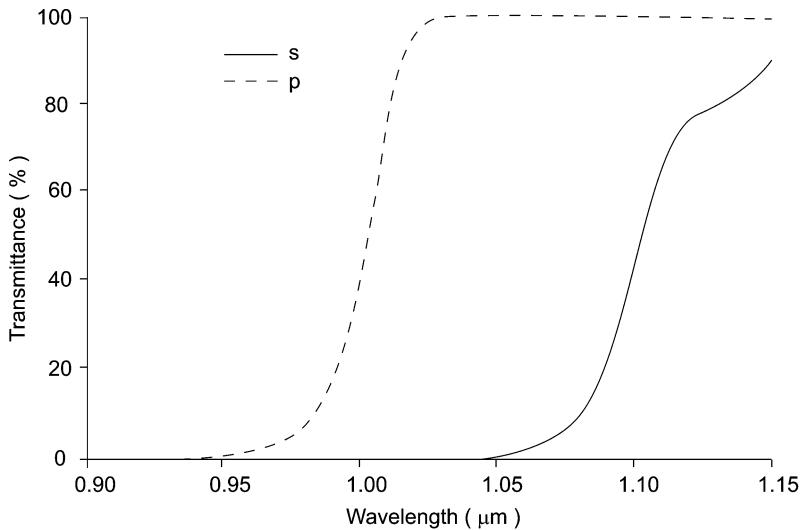


Figure 8.11. Characteristic curve of a plate polariser for $1.06 \mu\text{m}$. Design:

$$\text{Air}|(0.5H'L'0.5H')^3(0.5H''L''0.5H'')^80.5H'L'0.5H'| \text{Glass}$$

where $H' = 1.010H$, $L' = 1.146L$, $H'' = 1.076H$, $L'' = 1.220L$ and with $n_H = 2.25$, $n_L = 1.45$, $\lambda_0 = 0.9 \mu\text{m}$ and $\theta_0 = 56.5^\circ$. The solid line indicates s-polarisation and the dashed line p-polarisation. (After Songer [17].)

8.4 Nonpolarising coatings

The design of coatings which avoid polarisation problems is a much more difficult task than that of polariser design and there is no completely effective method. The changes in the phase thickness of the layers and in the optical admittances are fundamental and cannot be avoided. The best we can hope to do, therefore, is to arrange the sequence of layers so that they give the same performance for p as for s-polarisation. Clearly, the wider the range of either angle of incidence or of wavelength, the more difficult the task. The techniques which are currently available operate only over very restricted ranges of wavelength and angle of incidence (effectively over a very narrow range of angles). There is a small body of published work but the principal techniques we shall use here rely heavily on techniques devised by Thelen [20, 21].

8.4.1 Edge filters at intermediate angle of incidence

This section is based entirely on an important paper by Thelen [20]. However, the expressions found in the original paper have been altered in order to make

the notation consistent with the remainder of this book. Care should be taken, therefore, in reading the original paper. In particular, the x found in the original is defined in a slightly different way.

At angles of incidence which are not so severe that the p-reflectance suffers, the principal effect of operating edge filters at oblique incidence is the splitting between the two planes of polarisation. This limits the edge steepness which can be achieved for light which is unpolarised. Edge filters which have pass regions which are quite limited can be constructed from band-pass filters, but, because band-pass filters are also affected in much the same way, the bandwidth for s-polarised light shrinking and for p-polarised light expanding, they still suffer from the same problem. However, there is a technique which can be used for displacing the pass bands of a band-pass filter to make one pair of edges coincide, resulting in an edge filter of rather limited extent, which for a given angle of incidence has no polarisation splitting. The position of the peak of a band-pass filter can be considered to be a function of both the spacer thickness and the phase shift of the reflecting stacks on either side. At oblique incidence, the relative phase shift between s- and p-polarised light from the reflecting stacks can be adjusted by adding or removing material. This alters the relative positions of the peaks of the pass bands for the two planes of polarisation and, if the adjustment is correctly made, it can make a pair of edges coincide. This, of course, is for one angle of incidence only. As the angle of incidence moves away from the design value, the splitting will reappear.

Rather than apply this technique exactly as we have just described it, we instead adapt the techniques for the design of multiple cavity filters based on symmetrical periods. Let us take a typical multiple cavity filter design:

Incident medium|matching (symmetrical stack)^q matching|substrate.

The symmetrical stack which forms the basis of this filter can be represented as a single matrix which has the same form as that of a single film, as we have already seen in chapter 7. The limits of the pass band are given by those wavelengths for which the diagonal terms of the matrix are unity and the off-diagonal terms are zero. That is, if the matrix is given by

$$\begin{bmatrix} N_{11} & iN_{12} \\ iN_{21} & N_{22} \end{bmatrix}$$

then the edges of the pass band are given by

$$N_{11} = N_{22} = \pm 1.$$

The design procedure simply ensures that this condition is satisfied for the appropriate angle of incidence.

We can consider the symmetrical period as a quarter-wave stack of $2x + 1$ layers which has two additional layers added, one on either side:

$$fB\ A\ B\ A\ B\ \dots\ A\ fB$$

where A and B indicate quarter-wave layers and f is a correction factor which is to be applied to the quarter-wave thickness to yield the thicknesses of the detuned outer layers. We can write the overall matrix as $f B M f B$ where $M = ABAB \dots A$, giving the product:

$$\begin{bmatrix} \cos \alpha & i \sin \alpha / \eta_B \\ i \eta_B \sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} M_{11} & i M_{11} \\ i M_{21} & M_{11} \end{bmatrix} \begin{bmatrix} \cos \alpha & i \sin \alpha / \eta_B \\ i \eta_B \sin \alpha & \cos \alpha \end{bmatrix}.$$

Then N_{11} is given by

$$N_{11} = N_{22} = M_{11} \cos 2\alpha - 0.5(M_{12}\eta_B - M_{21}/\eta_B) \sin 2\alpha = \pm 1 \quad (8.27)$$

for the edge of the zone for each plane of polarisation. This must be satisfied for both planes of polarisation simultaneously for the edges of the pass bands to coincide. In fact, symmetrical periods which are made up of thicknesses other than quarter-waves can be used, when some trial and error will be required to satisfy equation (8.27). A computer can be of considerable help. For quarter-wave stacks we seek assistance in the expressions derived in chapter 7 for narrowband filter design. We use expression (7.53), with $m = 1$ and $q = 0$, giving

$$\begin{aligned} M_{11} = M_{22} &= (-1)^x(-\varepsilon)[(\eta_A/\eta_B)^x + \dots + (\eta_B/\eta_A)^x] \\ iM_{12} &= i(-1)^x/[(\eta_A/\eta_B)^x \eta_A] \\ iM_{21} &= i(-1)^x[(\eta_A/\eta_B)^x \eta_A]. \end{aligned} \quad (8.28)$$

Note that $2x + 1$ is now the number of layers in the inner stack. The total number of layers, including the detuned ones, is $2x + 3$. Now, using exactly the same procedure as in chapter 7, we can write expressions for the coefficients in equation (8.27) as

$$\begin{aligned} M_{11} &= \frac{(-1)^x(-\varepsilon)(n_H/n_L)^x}{(1 - n_L/n_H)} \\ &= (-1)^x(-\varepsilon)P \end{aligned}$$

and

$$\begin{aligned} 0.5(M_{12}\eta_B + M_{21}/\eta_B) &= 0.5(-1)^x[(\eta_B/\eta_A)^{x+1} + (\eta_A/\eta_B)^{x+1}] \\ &= (-1)^xQ \end{aligned}$$

where

$$P = (\eta_H/\eta_L)^x/(1 - \eta_L/\eta_H) \quad \text{and} \quad Q = 0.5(\eta_H/\eta_L)^{x+1}.$$

Then the two equations become

$$\begin{aligned} \pm 1 &= \varepsilon P_p \cos 2\alpha + Q_p \sin 2\alpha \\ \pm 1 &= \varepsilon P_s \cos 2\alpha + Q_s \sin 2\alpha \end{aligned} \quad (8.29)$$

which give for α and ε :

$$\sin 2\alpha = \pm \frac{P_s - P_p}{(P_s Q_p - P_p Q_s)} \quad (8.30)$$

$$\varepsilon = \frac{\pm 1 - Q_p \sin 2\alpha}{P_p \cos 2\alpha}. \quad (8.31)$$

Now,

$$\begin{aligned} \varepsilon &= (\pi/2)(1-g) \quad \text{where} \quad g = \lambda_0/\lambda \\ \alpha &= (\pi/2)(\lambda_R/\lambda) = (\pi/2)(\lambda_R/\lambda_0)g = (\pi/2)fg \end{aligned}$$

so that

$$f = \alpha/(\pi g/2) = \alpha/(\pi/2 - \varepsilon). \quad (8.32)$$

Two values for f will be obtained. Usually, the larger corresponds to a shortwave-pass and the smaller to a longwave-pass filter.

There are some important points about the particular values of α and ε , which are best discussed within the framework of a numerical example. Let us attempt the design of a longwave-pass filter at 45° in air having a symmetrical period of

$$fL \text{ } HL \text{ } HL \text{ } HL \text{ } fL$$

where H represents an index of 2.35 and L of 1.35. The inner stack has seven layers, which corresponds to $2x + 1$, so that x in this example is 3. We will use the modified admittances that for this combination are (the subscripts S and A referring to the substrate and to air, respectively):

$$\begin{array}{ll} \eta_{Hs} = 3.1694 & \eta_{Ls} = 1.6264 \\ \eta_{Ss} = 1.9028 & \eta_{As} = 1.000 \\ \eta_{Hp} = 1.7425 & \eta_{Lp} = 1.1206 \\ \eta_{Sp} = 1.2142 & \eta_{Ap} = 1.000. \end{array}$$

Then

$$\begin{aligned} P_s &= 15.201 & P_p &= 10.535 \\ Q_s &= 7.211 & Q_p &= 2.923 \end{aligned}$$

giving $\sin \alpha = \pm 0.1480$.

Now, the outer tuning layers in their unperturbed state will be quarter-waves and so the two solutions we look for will be near $2\alpha = \pi$, that is, in the second and third quadrants. We continue to keep the results in the correct order and find

$$2\alpha = \pi \pm 0.1485 = 3.2901 \quad \text{or} \quad 2.9931.$$

Then, in both cases, $\cos 2\alpha = 0.9890$ and so

$$\varepsilon = \pm(1 + 2.923 \times 0.148)/(-10.535 \times 0.9890) = \pm(-0.1375)$$

whence

$$f = (3.2901/2)/[(\pi/2) - 0.1375] = 1.148$$

with

$$g = 1 - 2 \times 0.1375/\pi = 0.9125$$

and

$$f = (2.9931/2)/[(\pi/2) + 0.1375] = 0.876$$

with

$$g = 1 + \times 0.1375/\pi = 1.088.$$

We take the second of these which will correspond to a longwave-pass filter. We now need to consider the matching requirements. Since we are attempting to obtain coincident edges for both planes of polarisation in an edge filter of limited pass band extent, we will interest ourselves in having good performance right at the edge of the pass band with little regard for performance further away. We use the symmetrical period method. The basic period is

$$0.876L \text{ } HLHLHLH \text{ } 0.876L$$

with H and L quarter-waves of indices 2.35 and 1.35 respectively, and tuned for 45° . Calculation of the equivalent admittances for the symmetrical period gives the values for s- and p-polarisation shown in table 8.1. (Again they are modified admittances.) We will arrange matching at $g = 1.08$. Adding a $HLHL$ combination to the period with the L layer next to it yields admittances of 0.9625 for p-polarisation and 1.416 for s. The media we have to match have modified admittances of 1.0 for air and 1.214 for glass for p-polarisation and 1.0 and 1.903 respectively for s. As an initial attempt, therefore, this matching is probably adequate. Since the matching is to be at $g = 1.08$, the thicknesses of the four layers in the matching assemblies must be corrected by the factor 1.0/1.08. To complete the design we need to make sure all layers are tuned for 45° which means multiplying their effective thicknesses for 45° by the factor $1/\cos \theta$. The final design with all thicknesses quoted as their normal incidence values is then

$$\text{Air} | (0.971 H \text{ } 1.087 L)^2 | (1.028 L (1.049 H \text{ } 1.174 L)^3 | 1.049 H \text{ } 1.028 L)^q | (1.087 L \text{ } 0.971 H)^2 | \text{Glass.}$$

Table 8.1. Equivalent admittances and phase thicknesses of the symmetrical period ($0.876L H L H L H 0.876L$) where L and H indicate quarter-waves at 45° angle of incidence of index 1.35 and 2.35 respectively.

g	s-polarisation		p-polarisation	
	E (modified)	γ/π	E (modified)	γ/π
1.04	Imaginary values		0.1946	4.4372
1.05	0.0949	4.2955	0.2018	4.4372
1.06	0.1190	4.4454	0.1993	4.5884
1.07	0.1202	4.5786	0.1861	4.6652
1.08	0.0982	4.7211	0.1588	4.7486
1.09	Imaginary values		0.1049	4.8530
1.10	Imaginary values		Imaginary values	

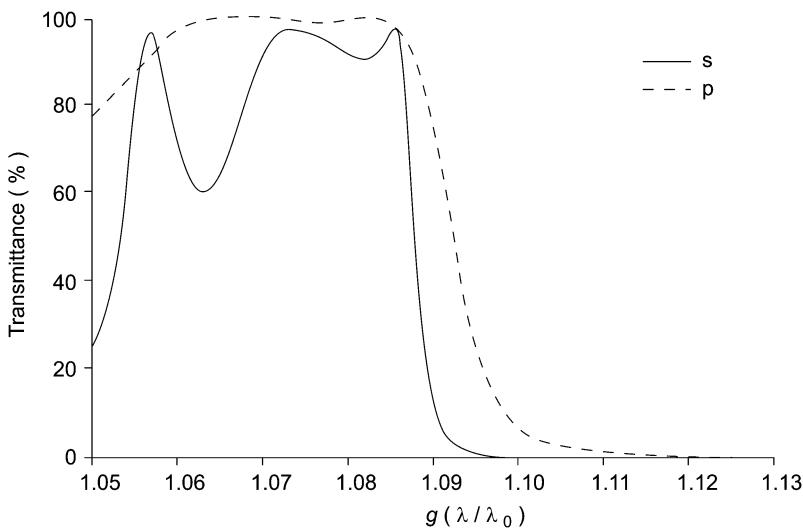


Figure 8.12. Calculated performance of a polarisation-free edge filter designed for use at 45° in air using the method of Thelen [20]. The multilayer structure is given in the text. The solid curve indicates s-polarisation and the dashed curve p-polarisation.

The performance with $q = 4$ is shown in figure 8.12 along with the performance of a band-pass filter of similar design using unaltered quarter-waves to demonstrate the difference. Since the p-admittances are less effective than the s in achieving high reflectance, the steepness of the edge for s-polarisation is somewhat greater and so the two edges coincide at their upper ends. Adjustment

of the factor f can move this point of coincidence up and down the edges. Thelen gives many examples of designs including some which are based on symmetrical periods containing thicknesses other than quarter-waves.

8.4.2 Reflecting coatings at very high angles of incidence

Reflecting coatings at very high angles of incidence suffer catastrophic reductions in reflectance for p-polarisation. This is especially true for coatings that are embedded in glass such as cube beam splitters and we have already seen how they can make good polarisers. The admittances for p-polarised light are not favourable for high reflectance and so to increase the p-reflectance we must use a large number of layers—many more than is usual at normal incidence. The s-reflectance must also at the same time be considerably reduced, otherwise it will vastly exceed what is possible for p-polarisation. The technique we use here is based on yet another method originated by Thelen [21]. A number of authors have studied the problem. For a detailed account of the use of symmetrical periods in the design of reflecting coatings for oblique incidence, the paper by Knittl and Houserkova [22] should be consulted.

We consider a quarter-wave stack. The admittance of such a stack is given at normal incidence by

$$Y = \frac{y_1^2 y_3^2 y_5^2 \dots y_{\text{sub}}^2}{y_2^2 y_4^2 y_6^2 \dots} \quad (8.33)$$

with y_{sub} in the numerator, as shown, if the number of layers is even or in the denominator if odd. The reflectance is

$$R = \left(\frac{y_0 - Y}{y_0 + Y} \right)^2$$

in the normal way. Now, if the stack of quarter-waves is considered to be tilted, with the thicknesses tuned to the particular angle of incidence, the expression for reflectance will be similar except that the appropriate tilted admittances must be used. Here we will use the modified admittances so that y_0 will remain the same. Then Y becomes

$$Y = \frac{\eta_1^2 \eta_3^2 \eta_5^2 \dots \eta_{\text{sub}}^2}{\eta_2^2 \eta_4^2 \eta_6^2 \dots} \quad (8.34)$$

and in order for the reflectances for p- and s-polarisations to be equal, the modified admittances for p- and s-polarisation must be equal. If we write Δ_1 for (η_{1p}/η_{1s}) and so on, then this condition is

$$\frac{\Delta_1^2 \Delta_3^2 \Delta_5^2 \dots \Delta_{\text{sub}}^2}{\Delta_2^2 \Delta_4^2 \Delta_6^2 \dots} = 1. \quad (8.35)$$

(Note that Thelen's paper does not use modified admittances and so includes the incident medium in the formula.) The procedure then is to attempt to find

Table 8.2.

n_f	$1/\cos\theta$	η_p	η_s	$\Delta (= \eta_p/\eta_s)$
1.35	1.6526	1.5776	1.1553	1.3656
1.38	1.5943	1.5558	1.2241	1.2710
1.45	1.4898	1.5275	1.3765	1.1097
1.52	1.4142	1.5200	1.5200	1.0000
1.57	1.3719	1.5230	1.6185	0.9410
1.65	1.3180	1.5377	1.7705	0.8685
1.70	1.2907	1.5515	1.8627	0.8330
1.75	1.2672	1.5680	1.9531	0.8028
1.80	1.2466	1.5867	2.0419	0.7771
1.85	1.2286	1.6072	2.1295	0.7548
1.90	1.2127	1.6292	2.2158	0.7353
1.95	1.1985	1.6525	2.3010	0.7182
2.00	1.1858	1.6770	2.3853	0.7030
2.05	1.1744	1.7023	2.4687	0.6895
2.10	1.1640	1.7285	2.5514	0.6775
2.15	1.1546	1.7554	2.6334	0.6666
2.20	1.1461	1.7829	2.7147	0.6568
2.25	1.1383	1.8110	2.7955	0.6478
2.30	1.1311	1.8396	2.8757	0.6397
2.35	1.1245	1.8686	2.9554	0.6323
2.40	1.1184	1.8980	3.0347	0.6254

Modified admittances
Incident medium index = 1.52
Angle of incidence = 45°

a combination of materials such that condition (8.35) is satisfied and the value of admittance is such that the required reflectance is achieved. This is a matter of trial and error.

An example will help to make the method clear. Table 8.2 gives some figures for modified admittances in glass ($n = 1.52$) and at an angle of incidence of 45°. There is a number of possible arrangements but the most straightforward is to find three materials H , L and M , M being of intermediate index, such that

$$\Delta_H \Delta_L = \Delta_M^2. \quad (8.36)$$

Then the multilayer structure can be ... $H M L M H M L M H M L M$... so that the form of admittance is

$$Y = \frac{\eta_H^2 \eta_L^2 \eta_H^2 \dots}{\eta_M^2 \eta_M^2 \eta_M^2 \dots} \quad (8.37)$$

and the number of layers chosen so that adequate reflectance is achieved. The

substrate does not appear in (8.37) because it is assumed to be of the same material as the incident medium and so Δ_{sub} is unity. Where the substrate is of a different material there may be a slight residual mismatch but practical difficulties will usually make achievement of an exact match difficult. A set of layers giving an approximate match at 45° has indices 1.35, 2.25 and 1.57. For this combination

$$\frac{\Delta_H \Delta_L}{\Delta_M^2} = \frac{1.3656 \times 0.6478}{0.941^2} = 0.999.$$

The p-admittance increase due to one four-layer period of that type is

$$\frac{\eta_{H\text{p}}^2 \eta_{L\text{p}}^2}{\eta_{M\text{p}}^4} = \frac{1.811^2 \times 1.578^2}{1.523^4} = 1.518.$$

Eight periods give a value of 28.2, that is a reflectance of 87% for 32 layers. The particular arrangement of H , L and M layers is flexible as long as H or L are odd and M is even. The performance of a coating to this design is shown in figure 8.13. The basic period is four quarter-waves thick. High-reflectance zones exist wherever the basic period is an integral number of half-waves thick. Since in this case we have four quarter-waves we expect extra-high-reflectance zones at $g = 0.5$ and $g = 1.5$. The peak at $g = 0.5$ (i.e. $\lambda = 2 \times 510 = 1020$ nm) is visible at the long wavelength end of the diagram.

Examination of the modified admittances for the materials shows how the coating does yield the desired performance. Each second pair of layers tends to reduce the s-reflectance of the preceding pair but slightly to increase the p-reflectance. To achieve high reflectance large numbers of layers are needed. Angular sensitivity is quite high and there is little that can be done to improve it.

8.4.3 Edge filters at very high angles of incidence

It is possible to adapt the treatment of the previous section to design edge filters for use at high angles of incidence. Let us illustrate the method by using the example we have just calculated. Figure 8.13 shows the performance. We wish to use this component as a longwave-pass filter and hence to eliminate the ripple on the longwave side of the peak. The ripple is principally confined to s-polarisation and so we concentrate our efforts there. We will use a symmetrical period approach.

The basic symmetrical period can be either

$$(0.5H \text{ } MLM \text{ } 0.5H) \quad \text{or} \quad (0.5L \text{ } MHM \text{ } 0.5L).$$

We use the modified s-admittances that we have already calculated in the previous section and we compute the equivalent admittances as shown in table 8.3. The surrounding material has admittance 1.52 and it appears as though a simple match would be obtained with the $(0.5L \text{ } MHM \text{ } 0.5L)$ combination. We match at $g =$

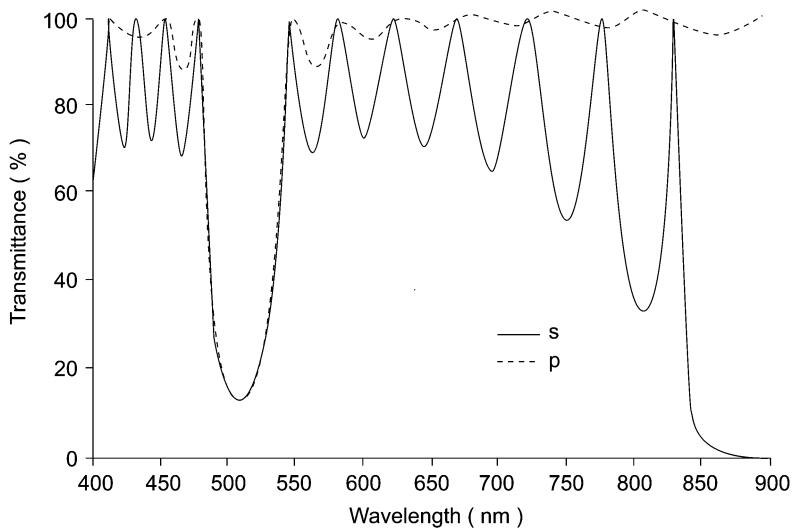


Figure 8.13. Calculated performance of a polarisation-free reflector at an angle of incidence of 45° in glass. The coating was designed using the method of Thelen [21]. Design: Glass|(1.38H 1.372M 1.653L 1.372M)⁸|Glass with $n_H = 2.25$, $n_M = 1.57$, $n_L = 1.35$, $n_{\text{Glass}} = 1.52$ and $\lambda_0 = 510 \text{ nm}$. The solid line indicates s-polarisation and the dashed line p-polarisation.

0.88 where the equivalent admittance is 0.802. To match to 1.52, a quarter-wave of admittance $(0.802 \times 1.52)^{1/2}$ is required. This is 1.104 and corresponds fairly well with the 1.155 admittance of the 1.35 low-index material. A quarter-wave at $g = 0.88$ and 45° has a normal incidence thickness of $(1.0/0.88) \times 1.653 \times 0.25$ full waves, that is, 1.877 quarter-waves or 0.470 full waves. The full design is then

$$\text{Glass}|1.877L(0.826L 1.372M 1.138H 1.372M 0.826L)^q|1.877|\text{Glass}.$$

The performance of a coating with $q = 10$ is shown in figure 8.14. Shortwave-pass filters or filters with different materials can be designed in the same way. The design is fairly sensitive to materials and to angle of incidence.

8.5 Antireflection coatings

Antireflection coatings at high angles of incidence are a stage more difficult than the design of coatings for normal incidence. Some simplification occurs when only one plane of polarisation has to be considered. Then it is a case of taking the tables for modified optical admittance at the appropriate angle of incidence and designing coatings in much the same way as for normal incidence. The

Table 8.3. Equivalent admittances and phase thicknesses of the symmetrical periods ($0.5LMHM0.5L$) and ($0.5HMLM0.5H$) calculated for 45° angle of incidence in glass of index 1.52. $n_H = 2.35$, $n_L = 1.35$ and $n_M = 1.57$.

g	$E_{\text{mod.s}}$		
	($0.5HMLM0.5H$)	($0.5LMHM0.5L$)	γ/π
0.58	Imaginary values		
0.60	18.8985	0.1442	1.0473
0.62	6.8181	0.3965	1.1438
0.64	5.1698	0.5184	1.2061
0.68	4.4178	0.6007	1.2600
0.70	3.9680	0.6613	1.3100
0.72	3.6599	0.7443	1.3577
0.74	3.4300	0.7728	1.4040
0.76	3.2471	0.7949	1.4494
0.78	2.9594	0.8114	1.5382
0.80	2.8362	0.8225	1.5820
0.82	2.7180	0.8281	1.6256
0.84	2.5994	0.8276	1.6691
0.86	2.4741	0.8199	1.7126
0.88	2.3340	0.8024	1.7564
0.90	2.1662	0.7705	1.8005
0.92	1.9467	0.7151	1.8456
0.94	1.6195	0.6135	1.8930
0.96	0.9761	0.3808	1.9489
0.98	Imaginary values		

complication is that the range of admittances available is different from the range at normal incidence and also different for the two planes of polarisation. We therefore consider briefly the problem of antireflection coatings for one plane of polarisation first. In order to simplify the discussion of design we assume an angle of incidence of 60° in air with a substrate of index 1.5 and possible film indices of 1.3, 1.4, 1.5, ..., 2.5. Real designs will be based on available indices and will therefore be more constrained and may require more layers. The modified admittances with values of $\Delta (= \eta_p/\eta_s)$ are given in table 8.4.

8.5.1 p-polarisation only

At 60° the modified p-admittance of the substrate is only 0.9186 giving a single-surface reflectance for p-polarised light of less than 0.2%, acceptable for most purposes. The angle of incidence of 60° is only just greater than the Brewster angle. If still lower reflectance is required then a single quarter-wave

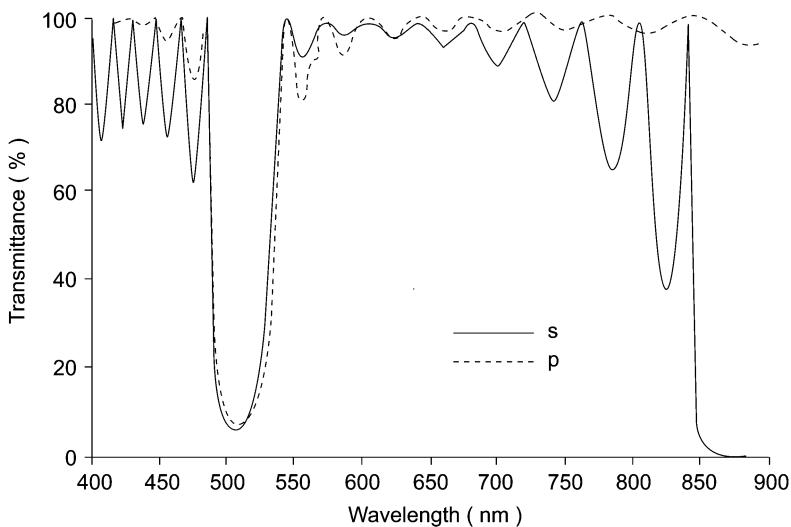


Figure 8.14. Calculated performance of a polarisation-free edge filter at an angle of incidence of 45° in glass. Design: Glass| $1.877L$ ($0.826L$) $1.372M$ $1.138H$ $1.372M$ $0.826L$)¹⁰ 1.877 |Glass with $n_H = 2.25$, $n_M = 1.57$, $n_L = 1.35$, $n_{\text{Glass}} = 1.52$ and $\lambda_0 = 510$ nm. The solid line indicates s-polarisation and the dashed line p-polarisation.

of admittance given by $(0.9186 \times 1.0000)^{1/2}$, that is 0.9584, is required. This corresponds from table 8.3 to an index of 1.6, that is *greater* than the index of the substrate. As the angle of incidence increases still further from 60° the required index will become still greater. Eventually, at very high angles of incidence indeed, the required single layer index will be greater than the highest index available and at that stage designs based on combinations such as Air| HL |Glass will be required with quarter-wave thicknesses at the appropriate angle of incidence. Such coatings operate over a very small range of angles of incidence only and are very difficult to produce with any reasonable degree of success. If at all possible it is better to avoid such designs altogether by redesigning the optical system.

8.5.2 s-polarisation only

The modified s-admittance for the substrate is 2.449 and the required single-layer admittance for perfect antireflection is $(2.4495 \times 1.0000)^{1/2}$ or 1.5650, well below the available range. The problem is akin to that at normal incidence where we do not have materials of sufficiently low index and the solution is similar. We begin by raising the admittance of the substrate to an acceptable level by adding a quarter-wave of higher admittance. In this case a layer of

Table 8.4.

n_f	$1/\cos\theta$	η_p	η_s	$\Delta (= \eta_p/\eta_s)$
1.00	2.0000	1.0000	1.0000	1.0000
1.30	1.3409	0.8716	1.9391	0.4495
1.40	1.2727	0.8909	2.2000	0.4050
1.50	1.2247	0.9186	2.4495	0.3750
1.60	1.1893	0.9514	2.6907	0.3536
1.70	1.1621	0.9878	2.9258	0.3376
1.80	1.1407	1.0266	3.1560	0.3253
1.90	1.1235	1.0673	3.3823	0.3156
2.00	1.1094	1.1094	3.6056	0.3077
2.10	1.0977	1.1526	3.8262	0.3012
2.20	1.0878	1.1966	4.0448	0.2958
2.30	1.0794	1.2414	4.2615	0.2913
2.40	1.0722	1.2867	4.4766	0.2874
2.50	1.0660	1.3325	4.6904	0.2841

Modified admittances
Incident medium index = 1.00
Angle of incidence = 60°

index 1.9 or admittance 3.3823 is convenient and gives a resultant admittance of $3.3823^2/2.449$ or 4.6713 that requires a quarter-wave of admittance $(4.6713 \times 1.0000)^{1/2}$ or 2.1613 to complete the design. This corresponds most nearly to an index of 1.4, admittance 2.2000, and the residual reflectance with such a combination is 0.03%, a considerable improvement over the 17.7% reflectance of the uncoated substrate. We cannot expect that such a coating will have a broad characteristic and figure 8.15 confirms it. A small improvement can be made by adding a high-admittance half-wave layer between the two quarter-waves or a low-admittance half-wave next to the substrate. The latter is also shown in the figure. In terms of normal incidence thicknesses the two designs are:

$$\text{Air}|1.273L\ 1.123H|\text{Glass}$$

and

$$\text{Air}|1.273L\ 1.123H\ 2.682A|\text{Glass}$$

where L , H and A indicate quarter-waves at normal incidence of films of index 1.4, 1.9 and 1.3 respectively. The p-reflectance of these designs is very high and they are definitely suitable for s-polarisation only.

Again it is better wherever possible to avoid the necessity for such antireflection coatings by rearranging the optical design of the instrument so that s-polarised light is reflected and p-polarised light is transmitted.

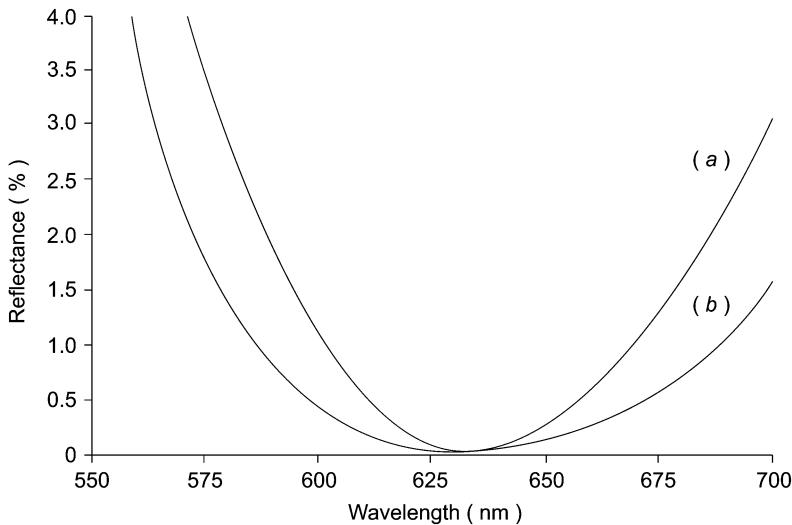


Figure 8.15. Antireflection coatings for s-polarised light at an angle of incidence of 60° in air. (a) Air|1.273L 1.123H|Glass, (b) Air|1.273L 1.123H 2.682A|Glass with $n_L = 1.4$, $n_H = 1.9$, $n_A = 1.3$, $n_{\text{Glass}} = 1.5$ and $\lambda_0 = 632.8 \text{ nm}$.

8.5.3 s- and p-polarisation together

The task of assuring low reflectance for both s- and p-polarised light is almost impossible and should only be attempted as a last and very expensive resort. It is possible to arrive at designs that are effective over a narrow wavelength region and one such technique is included here. Again we use the range of indices given in table 8.4 and design a coating to give low s- and p-reflectance on a substrate of index 1.5 in air.

We use quarter-wave layer thicknesses only and a design technique similar to the procedure we have already used for high-reflectance coatings but with an additional condition that the admittance of both substrate and coating for both p- and s-polarisations should be unity to match the incident medium. This implies

$$\frac{\Delta_1^2 \Delta_3^2 \Delta_5^2 \dots \Delta_{\text{sub}}}{\Delta_2^2 \Delta_4^2 \Delta_6^2 \dots} = 1 \quad (8.38)$$

and

$$Y = \frac{\eta_{1s}^2 \eta_{3s}^2 \eta_{5s}^2 \dots \eta_{\text{sub},s}}{\eta_{2s}^2 \eta_{4s}^2 \eta_{6s}^2 \dots} = 1. \quad (8.39)$$

Equation (8.39) ensures that the reflectance for s-polarised light is zero and equation (8.38) that the p-reflectance equals the s-reflectance. From table 8.4, the starting values are $\Delta_{\text{sub}} = 0.3750$ and $\eta_{\text{sub}} = 2.4495$. Trial and error

shows that with the addition of one single quarter-wave layer, the best result corresponds to an index of 1.3 for which $\Delta_1^2/\Delta_{\text{sub}} = 0.4495^2/0.3750 = 0.5387$ and $\eta_{1s}^2/\eta_{\text{sub}} = 1.9391^2/2.4495 = 1.5350$. Other combinations give values that are further from unity in each case. Adopting a quarter-wave of index 1.3 as the first layer of the coating we need a further combination of layers that will provide a correction factor of 1.3624 in Δ and of 0.8071 in η_s . An additional single layer will not do, but two-layer combinations of a high- followed by a low-index layer can be found that will correct Δ but which are inadequate in terms of η_s . The two-layer combination that comes nearest to satisfying the requirements is a layer of index 1.8 followed by one of index 1.3 making the design so far:

$$|n = 1.3|n = 1.8|n = 1.3|\text{Glass}.$$

This has an overall Δ of $(0.4495^2 \times 0.4495^2)/(0.3253^2 \times 0.375) = 1.0288$ and a η_s of $(1.9391^2 \times 0.9391^2)/(3.1560^2 \times 2.4495) = 0.5795$. But the combination of index 2.5 followed by 1.4 gives approximately the same correction for Δ but a different correction for η_s . This gives the opportunity of using both combinations in a four-layer arrangement to adjust the value of η_s without altering Δ . The correction factor for Δ is given by $(0.4495^2 \times 0.2841^2)/0.4050^2 \times 0.3253^2) = 0.9396$ and for η_s by $(1.9391^2 \times 4.6904^2)/(2.2000^2 \times 3.1560^2) = 1.7159$. This then yields an overall value for Δ of $0.9396 \times 1.0288 = 0.9667$ and for η_s of $1.7159 \times 0.5795 = 0.9944$. The seven layers can be put in various orders without altering the reflectance at the reference wavelength. All that is required is that the 1.3 and 2.5 indices should be odd and the 1.4 and 1.8 indices even. Here we put them in descending value of index from the substrate so that the final design is:

$$\text{Air}|1.3409L\ 1.2727A\ 1.3409L\ 1.1407B\ 1.3409L\ 1.1407B\ 1.066H|\text{Glass}$$

with $n_L = 1.30$, $n_A = 1.40$, $n_B = 1.80$ and $n_H = 2.50$.

The calculated performance of this coating for a reference wavelength of 632.8 nm is shown in figure 8.16. As we might have suspected, the width of the zone of low reflectance is narrow. An alternative design arrived at in the same way but for a substrate of index 1.52 and a range of film indices from 1.35 to 2.40 uses ten layers:

$$\begin{aligned} &\text{Air}|1.3036L\ 1.1748A\ 1.3036L\ 1.1748A\ 1.3036L\ 1.1407B \\ &\quad 1.0722H\ 1.1235C\ 1.0722H\ 1.1235C|\text{Glass} \end{aligned}$$

with $n_L = 1.35$, $n_A = 1.65$, $n_B = 1.80$, $n_C = 1.90$, $n_H = 2.40$, $n_{\text{Glass}} = 1.52$ and $n_{\text{air}} = 1.00$. The performance is similar to that of figure 8.16.

8.6 Retarders

8.6.1 Achromatic quarter- and half-wave retardation plates

As well as being used in the construction of polarisers, optical thin films can find application in the production of achromatic quarter- and half-wave plates.

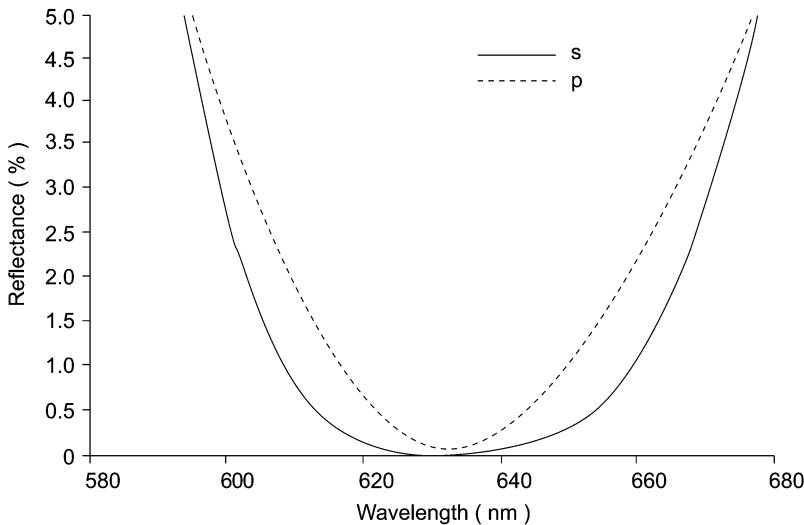


Figure 8.16. Calculated performance of an antireflection coating for glass to have low reflectance for both p- and s-polarisation at an angle of incidence of 60° in air. The solid line indicates s-polarisation and the dashed line p-polarisation. $\lambda_0 = 632.8 \text{ nm}$ and the design is given in the text.

A quarter-wave plate by definition produces between the two principal planes of polarisation a phase shift of 90° , which corresponds to an optical path difference of a quarter of a wavelength, while a half-wave plate produces a phase shift of 180° corresponding to a half wavelength. These components are generally made from mica, or some other similar birefringent material, cut to such a thickness that the difference in optical pathlength for each plane of polarisation is either a quarter or a half wavelength. A considerable disadvantage of such retardation plates is the rapid variation of the performance of the device with wavelength.

The case of the half-wave plate has been considered by Lostis [7], who has used a thin film to alter the phase shift on total internal reflection to make it exactly 180° . The arrangement is shown in figure 8.17. The notation for the various refractive indices and thicknesses is shown also in the figure. Let Y indicate the optical admittance with regard to the s-plane of polarisation and Z with respect to the p-plane. Then $Y_r = n_r \cos \phi_r$, $Z_r = n_r / \cos \phi_r$. Once the notation is established the calculation of the reflectances for the two planes of polarisation is an easy matter. The reflectance will be total for both but their phase shifts will depend on the parameters of the thin film. The condition that the relative phase difference between the two planes of polarisation should be 180° can then be asserted and the necessary condition derived for this to be so. Lostis found this condition to be

$$A \tan \beta + B \tan \beta + C = 0 \quad (8.40)$$

where

$$\begin{aligned}\beta &= \frac{2\pi}{\lambda} n_1 d \cos \phi_1 \\ A &= n_1^2 - \left(\frac{n_0 n_2}{n_1} \right)^2 \\ B &= \frac{\gamma}{n_1 \cos \phi_1} (n_1^2 - n_0^2) + \frac{n_1 \cos \phi_1}{\gamma} \left[\left(\frac{n_0 n_2}{n_1} \right)^2 - n_2^2 \right] \\ C &= n_0 - n_2^2\end{aligned}$$

and

$$Y_2 = n_2 \cos \phi_2 = i(n_0^2 \sin^2 \phi_0 - n_2^2)^{1/2} = i\gamma.$$

In the case where the surrounding medium is air, of index 1.0, the necessary condition for the above equation to have a real root is

$$n_0 \leq 1.46 \quad \text{and} \quad n_1 \geq 2.6.$$

When the limiting values are inserted in equation (8.40), the optical thickness of the film is found to be $\lambda/11$. Having arrived at this value the retardation can be calculated for the rest of the visible spectrum and it is found that the retardation does not vary by more than $\pm\lambda/50$ from 400–700 nm. Lostis constructed such a system using a prism of fused silica and a layer of titanium dioxide as the thin film.

The quarter-wave plate made from mica suffers from the same disability as the half-wave plate. It is correct for only one wavelength. Results derived in chapter 2 show that the phase change on total internal reflection varies with the angle of incidence and the plane of polarisation, and the difference in phase between the two principal planes also varies as the angle of incidence varies. With the materials available in the visible region it is not possible with a single reflection to obtain a retardation of 90° , but, with glass of refractive index 1.51, a retardation of 45° is obtained with an angle of incidence of either $48^\circ 37'$ or $54^\circ 37'$, and with two successive internal reflections the value of 90° can be obtained [23]. This is achieved in a device known as a Fresnel rhomb, shown in figure 8.18. The Fresnel rhomb is almost achromatic in performance, but the dispersion of the glass causes the retardation to increase gradually with decrease in wavelength. A further disadvantage of the Fresnel rhomb is its sensitivity to angle of incidence changes. The performance of the Fresnel rhomb can be considerably improved in both these directions by the addition of a thin-film coating to both surfaces of the rhomb. King [24] has manufactured Fresnel rhombs which show a phase retardation which varies by less than 0.4° over the wavelength range 330–600 nm. These were made from hard crown glass with one surface coated with magnesium fluoride 20 nm thick.

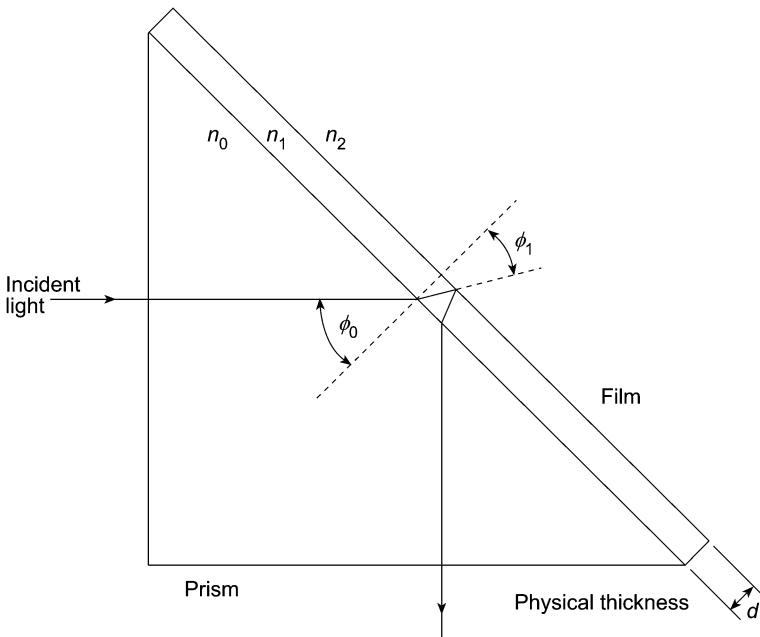


Figure 8.17. A half-wave retardation prism. (After Lostis [7].)

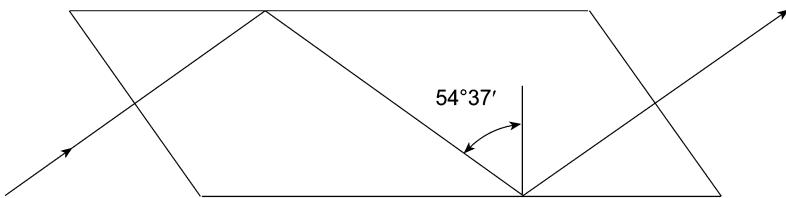


Figure 8.18. A Fresnel rhomb.

8.6.2 Multilayer phase retarders

In recent years there has been a number of applications where reflecting coatings have been required which introduced specified phase retardances between s- and p-polarisation. In particular there is a need in certain types of high-power laser resonators for coatings that introduce a 90° phase shift between s- and p-polarisation at an angle of incidence of 45° . The coatings that have been designed and manufactured for this purpose have been tuned for wavelengths in the infrared and have taken the form of silver films with a multilayer dielectric overcoat. The first published designs were due to Southwell [25, 26] who used a computer synthesis technique. Then Apfel [27] devised an analytical approach that we follow here. The principle of operation of the coatings is that an added dielectric

layer will not affect the reflectance of a system that already has a reflectance of unity. It will simply alter the phase change on reflection. When the component is used at oblique incidence, the alteration in phase will be different for each plane of polarisation. By adding layers in the correct sequence, eventually any desired phase difference between p- and s-polarisation for a single specified angle of incidence and wavelength can be achieved. In practice a silver layer is used as the basic reflecting coating and, although this has reflectance slightly less than unity, in the infrared it is high enough for it to be possible to neglect any error that might otherwise be introduced. It is of course not necessary to use a metal layer as starting reflector. A dielectric stack would be equally effective but would simply have more layers.

The basis of Apfel's method is a plot of phase retardance, denoted by Apfel as D , against the average phase shift A as a function of thickness of added layer of a given index. For simplicity, we retain this notation but in the rest of what follows we alter both notation and derivation to agree with the remainder of the book.

The starting point of the treatment is a reflector with a reflectance of unity, that is, a surface with imaginary admittance. Let this imaginary admittance be $i\beta$. Then

$$\rho e^{i\phi} = e^{i\phi} = (\eta_0 - i\beta)/(\eta_0 + i\beta) \quad (8.41)$$

i.e.

$$\tan(\phi_{\text{sub}}/2) = -\beta/\eta_0. \quad (8.42)$$

Should the incident medium be changed to η_1 then the phase shift becomes

$$\tan(\phi_1/2) = (-\beta/\eta_1) = (\eta_0/\eta_1) \tan(\phi_{\text{sub}}/2). \quad (8.43)$$

Now we add a film of admittance η_1 and phase thickness $\delta_1 = (2\pi/\lambda)n_1 d_1$ to the substrate.

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_1 & i(\sin \delta_1)/\eta_1 \\ i\eta_1 \sin \delta_1 & \cos \delta_1 \end{bmatrix} \begin{bmatrix} 1 \\ i\beta \end{bmatrix}$$

$$= \begin{bmatrix} \cos \delta_1 - (\beta/\eta_1) \sin \delta_1 \\ i(\eta_1 \sin \delta_1 + \beta \cos \delta_1) \end{bmatrix}. \quad (8.44)$$

The phase shift is now given, from equation (8.44), as

$$\tan(\phi_0/2) = \frac{-(\eta_1 \sin \delta_1 + \beta \cos \delta_1)}{\eta_0 [\cos \delta_1 - (\beta/\eta_1) \sin \delta_1]} = (\eta_1/\eta_0) \frac{[(-\beta/\eta_1) - \tan \delta_1]}{[1 + (-\beta/\eta_1) \tan \delta_1]}.$$

The second factor has the form of the tangent of the differences of two angles. Using this and expression (8.43) we have

$$\tan(\phi_0/2) = (\eta_1/\eta_0) \tan(\phi_1/2 - \delta_1). \quad (8.45)$$

This expression is valid for either plane of polarisation simply by inserting the appropriate values of η and δ .

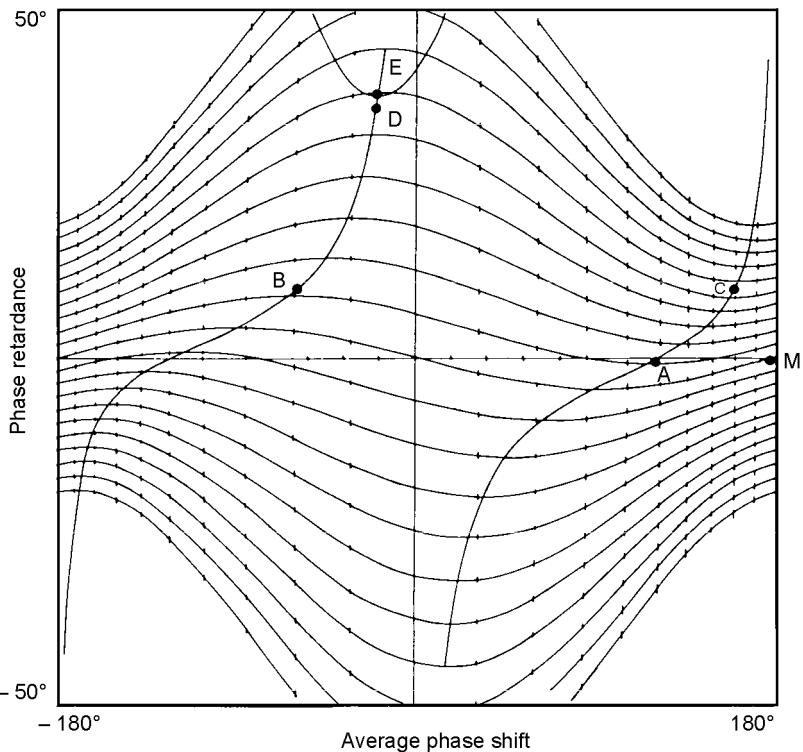


Figure 8.19. Immersed D - A plot for a film of index 4.0 in an incident medium of index 2.2 at an angle of incidence of 45° in air. The two S -shaped vertical curves mark the extrema of the D - A curves. The target retardation of 90° in air is denoted by the U -shaped curve at the top of the figure. The letters M, A, B, C, D and E are explained in the text. (After Apfel [27].)

To draw a D - A curve, we choose a starting point given by $D = 2\psi$ and $A = 0$, equivalent to $\phi_{\text{sub},p} = \psi$ and $\phi_{\text{sub},s} = -\psi$, and plot the difference in phase against the average phase all calculated from (8.45). Different values of ψ yield a family of curves. This family of curves can have a scale of thickness marked along them, in the manner of figure 8.19. Note that as curves disappear off the left-hand side of the diagram they reappear at the right-hand side. The relationships for the various quantities may be written

p-polarisation:

$$\begin{aligned}\tan(\phi_{0,p}/2) &= [(y_1 \cos \theta_0)/(y_0 \cos \theta_1)] \tan[(\phi_{1,p}/2) - \delta_1] \\ \tan(\phi_{1,p}/2) &= [(y_0 \cos \theta_1)/(y_1 \cos \theta_0)] \tan(\psi/2)\end{aligned}\quad (8.46)$$

s-polarisation:

$$\tan(\phi_{0,s}/2) = [(y_1 \cos \theta_1)/(y_0 \cos \theta_0)] \tan[(\phi_{1,s}/2) - \delta_1] \quad (8.47)$$

$$\tan(\phi_{1,s}/2) = [(y_0 \cos \theta_0)/(y_1 \cos \theta_1)] \tan(-\psi/2)$$

where δ_1 is calculated for the appropriate angle of incidence. Then

$$D = \phi_{0,p} - \phi_{0,s} \quad A = (\phi_{0,p} + \phi_{0,s})/2.$$

The curves now make it possible to determine the phase retardation produced by any thickness of the dielectric material added to any substrate of unity reflectance. To complete the design we need to construct similar diagrams for each dielectric material that is to be used. Since these sets of curves will not coincide, it is possible to reach any point of the diagram simply by moving from one set of curves to the other in succession. Only two dielectric materials are necessary and in that case Apfel shows that a technique of immersion simplifies the diagram. If we imagine that the structure is immersed in a medium of admittance equal to y_1 then

$$n_0 = n_1 \quad y_0 = y_1$$

and

$$\tan(\phi_0/2) = \tan[(\phi_1/2) - \delta_1]$$

for both planes of polarisation. Then D is a constant and $A = -2\delta_1$, since $\phi_{1,s} = -\phi_{1,p}$.

This result implies that the curves corresponding to the addition of material of index equal to that of the incident medium are horizontal lines on the diagram and can easily be visualised. The only problem we have now is that the target retardation is specified in a medium that will, in general, be different from that of the layer material. We therefore must add to the diagram the specification for retardation in the dummy immersion medium that will give the correct retardation when the dummy medium is removed and replaced by the correct medium. Let the phase retardation required in the correct incident medium be D_f . Then we can write

$$D_f = \phi_{fp} - \phi_{fs} \quad 2A_f = \phi_{fp} + \phi_{fs}$$

i.e.

$$\phi_{fp} = [(D_f/2) + A_f] \quad \phi_{fs} = [-(D_f/2) + A_f].$$

Converting ϕ_{fp} and ϕ_{fs} to ϕ_{0p} and ϕ_{0s} , the immersed values are

$$\tan(\phi_{0p}/2) = \frac{n_0 \cos \theta_1}{n_1 \cos \theta_0} \tan(\phi_{fp}/2) = \frac{n_0 \cos \theta_1}{n_1 \cos \theta_0} \tan[(D_f/4) + (A_f/2)] \quad (8.48)$$

$$\tan(\phi_{0s}/2) = \frac{n_0 \cos \theta_1}{n_1 \cos \theta_0} \tan(\phi_{fs}/2) = \frac{n_0 \cos \theta_1}{n_1 \cos \theta_0} \tan[(D_f/4) + (A_f/2)].$$

Then varying A_f gives the curves. Note that equations (8.48) are similar to (8.46) and (8.47) but with n_0 and n_1 interchanged.

The method is illustrated by figure 8.19, taken from Apfel [27] and showing the design curves for a retarder constructed of films of germanium, index 4.0, and zinc sulphide, index 2.2, to have a retardance of 90° in air at an angle of incidence of 45° . The curves of figure 8.19 are D - A curves for germanium immersed in a medium of index 2.2. The U -shaped curve in the upper region is the retardation target of 90° in air referred to the dummy medium of 2.2. The S -shaped curves running top to bottom mark the maxima of the D - A curves while the tick marks are made at intervals of one-tenth of a quarter-wave optical thickness. The four-layer design: $0.864H\ 0.778L\ 0.674H\ 0.319L$ Ag gives a retardance of 86.8° at the design wavelength and is represented by the trajectory MABCD. Two extra layers would be required to reach exactly 90° . The diagram could be made into a design aid for any desired retardance by adding a family of target curves.

8.7 Optical tunnel filters

At an earlier stage in the development of narrowband filters a main barrier to their construction was the fabrication of reflecting stacks of sufficiently low loss, and it appeared that the phenomenon of frustrated total internal reflection might offer some hope as a possible solution. This phenomenon has been known for some time. If light is incident on a boundary beyond the critical angle, it will normally be completely reflected. However, the incident light does in fact penetrate a short distance into the second medium, where it decays exponentially. Provided the second medium is somewhat thicker than a wavelength or so, the decay will be more or less complete and the reflectance unity. If, on the other hand, the second medium is made extremely thin, then the decay may not be complete when the wave meets the boundary with the third medium and, if the angle of propagation is then no longer greater than critical, a proportion of the incident light will appear in the third medium and the reflectance at the first boundary will be something short of total. This, as Baumeister [28] has pointed out, is very similar to the behaviour of fundamental particles in tunnelling through a potential barrier, and he has used the term ‘optical tunnelling’ to describe the phenomenon. The most important feature of the effect, as far as the thin-film filter is concerned, is that the frustrated total reflection can be adjusted to any desired value, simply by varying the thickness of the frustrating layer between the first and third media.

The method of constructing a filter using this effect is very similar to the polarising beam splitter (p 362). The hypotenuse of a prism is first coated with a frustrating layer of lower index so that the light will be incident at an angle greater than critical. This is a function of the prism angle, refractive index, and the refractive index of the frustrating layer. Next follows the spacer layer which must necessarily be of higher index so that a real angle of propagation will exist. This in turn is followed by yet another frustrating layer. The whole is then cemented

into a prism block by adding a second prism. The angle at which light is incident on the diagonal face must be greater than the angle ψ given by

$$\sin \psi = n_F/n_G$$

where n_F is the index of the frustrating layer and n_G is the index of the glass of the prism. For $n_F = 1.35$ and $n_G = 1.52$, we find $\psi = 63^\circ$, which is quite an appreciable angle. Usually glass of rather higher index, nearer 1.7, is used to reduce the angle as far as possible.

Although at first sight the optical tunnel or frustrated total reflectance (FTR) filter appears most attractive and simple, there are some tremendous theoretical disadvantages. First there is an enormous shift in peak wavelength between the two planes of polarisation. Typical figures quoted are of the order of 100 nm in the visible region, the peak corresponding to the p-plane of polarisation being at a shorter wavelength. This large polarisation splitting is due to the large angle of incidence at which the device must be used. Another effect of this large angle is that the angle sensitivity of the filter is extremely large. Shifts of 5 nm/degree of arc have been calculated [28].

Added to these disadvantages is the fact that the attempts which have been made to produce FTR filters have been very disappointing in their results, the performance appearing to fall far short of what was expected theoretically. It seems that the difficulties inherent in the construction of the FTR filter are at least as great as those involved in the conventional Fabry–Perot filter. Because of this, interest in the FTR filter has been mainly theoretical and the filter does not appear to be in commercial production.

The theory of the FTR filter has been written up in great detail by Baumeister [28]. Not only has he covered the FTR filter but he has also pointed out that, as far as the theory is concerned, the frustrating layer or, as he has renamed it, the tunnel layer, behaves exactly as a loss-free metal layer. This means that all sorts of filters including induced-transmission filters are possible using tunnel layers. Designs for a number of these are included in the paper. One conclusion which Baumeister reaches is that there appears to be no practical application for the tunnel-layer filter of the induced-transmission and FTR Fabry–Perot types. However he does mention the possibility of a longwave-pass filter constructed from an assembly of many tunnel layers separated by spacer layers and which has the advantage of a limitless rejection zone on the shortwave side of the edge. Even with this type of filter there are some disadvantages which could be serious. The characteristics of the filter near the edge suffer from strong polarisation splitting. This could be overcome by adding a conventional edge filter to the assembly at the front face of the prism. However, the second disadvantage is rather more serious: the appearance of pass bands in the stop region when the filter is tilted in the direction so as to make the angle of incidence more nearly normal. Curves given by Baumeister show a small transmission spike appearing even with a tilt of only 1° internal or 2.7° external with respect to the design value.

References

- [1] Thelen A 1966 Equivalent layers in multilayer filters *J. Opt. Soc. Am.* **50** 1533–8
- [2] Berning P H and Turner A F 1957 Induced transmission in absorbing films applied to band pass filter design *J. Opt. Soc. Am.* **47** 230–9
- [3] Nevière M and Vincent P 1980 Brewster phenomena in a lossy waveguide used just under the cut-off thickness *J. Opt. (Paris)* **11** 153–9
- [4] Ruiz-Urbieto M, Sparrow E M and Parikh P D 1975 Two-film reflection polarizers: theory and application *Appl. Opt.* **14** 486–92
- [5] Cox J T, Hass G and Hunter W R 1975 Infrared reflectance of silicon oxide and magnesium fluoride protected aluminium mirrors at various angles of incidence from 8 μm to 12 μm *Appl. Opt.* **14** 1247–50
- [6] Clapham P B, Downs M J and King R J 1969 Some applications of thin films to polarization devices *Appl. Opt.* **8** 1965–74
- [7] Lostis M P 1957 Etude et réalisation d'une lame demi-onde en utilisant les propriétés des couches minces *J. Phys. Rad.* **18** 518–28
- [8] Kretschmann E and Raether H 1968 Radiative decay of non-radiative surface plasmons excited by light *Z. Naturf.* **23** 2135–6
- [9] Otto A 1968 Excitation of non-radiative surface plasma waves in silver by the method of frustrated total reflection *Z. Phys.* **216** 398–410
- [10] Raether H 1977 Surface plasma oscillations and their applications *Physics of Thin Films* vol 9 (New York: Academic) pp 145–261
- [11] Abelès F 1976 Optical properties of very thin films *Thin Solid Films* **34** 291–302
- [12] Greenland K M and Billington C 1950 The construction of interference filters for the transmission of specified wavelengths *J. Phys. Radium* **11** 418–21
- [13] Harrick N J and Turner A F 1970 A thin film optical cavity to induce absorption of thermal emission *Appl. Opt.* **9** 2111–14
- [14] Banning M 1947 Practical methods of making and using multilayer filters *J. Opt. Soc. Am.* **37** 792–7
- [15] MacNeill S M 1946 *Beam Splitter* US Patent Specification 2 403 731
- [16] Clapham P B 1969 The preparation of thin film polarizers *Rep. OP. MET.* 7 National Physics Laboratory, Teddington
- [17] Songer L 1978 The design and fabrication of a thin film polarizer *Opt. Spectra* **12** 45–50
- [18] Blanc D, Lissberger P H and Roy A 1979 The design, preparation and optical measurement of thin film polarizers *Thin Solid Films* **57** 191–8
- [19] Netterfield R P 1977 Practical thin-film polarizing beam splitters *Opt. Acta* **24** 69–79
- [20] Thelen A 1981 Nonpolarizing edge filters *J. Opt. Soc. Am.* **71** 309–14
- [21] Thelen A 1976 Nonpolarizing interference films inside a glass cube *Appl. Opt.* **15** 2983–5
- [22] Knittl Z and Houserkova H 1982 Equivalent layers in oblique incidence: the problem of unsplit admittances and depolarization of partial reflectors *Appl. Opt.* **11** 2055–68
- [23] Born M and Wolf E 1975 *Principles of Optics* 5th edn (Oxford: Pergamon)
- [24] King R J 1966 Quarter wave retardation systems based on the Fresnel rhomb *J. Sci. Instrum.* **43** 617–22
- [25] Southwell W H 1979 Multilayer coatings producing 90° phase change *Appl. Opt.* **18** 1875

- [26] Southwell W H 1980 Multilayer coating design achieving a broadband 90° phase shift *Appl. Opt.* **19** 2688–92
- [27] Apfel J H 1981 Graphical method to design multilayer phase retarders *Appl. Opt.* **20** 1024–9
- [28] Baumeister P W 1967 Optical tunnelling and its applications to optical filters *Appl. Opt.* **6** 897–905 (This paper lists 49 references.)

Chapter 9

Production methods and thin-film materials

In this chapter, we shall deal briefly with the fundamental process, the machines that are used for the thin-film deposition and discuss some aspects of the properties of thin-film materials. Subsequent chapters will include a more detailed examination of some of the problems met in production.

Much of this chapter is concerned with the properties of materials, ways of measuring them, and some examples of the results of the measurements of the important parameters. Probably the most important properties from the thin-film point of view are given in the following list, although the order is not that of relative importance, which will vary from one application to another.

1. Optical properties such as refractive index and region of transparency.
2. The method which must be used for the production of the material in thin-film form.
3. Mechanical properties of thin films such as hardness or resistance to abrasion, and the magnitude of any built-in stresses.
4. Chemical properties such as solubility and resistance to attack by the atmosphere, and compatibility with other materials.
5. Toxicity.
6. Price and availability.
7. Other properties which may be important in particular applications, for example, electrical conductivity or dielectric constant.

Item 7 is not one on which we comment further here. On the question of price and availability, item 6, there is also little that can be said. The situation is changing all the time. Note, however, that price is of secondary importance to suitability. The cost of a failed batch of coatings is very great compared with the price of the source materials. Many companies are able to offer a wide range of materials completely ready for thin-film production, together with all the necessary information on the techniques that should be used.

9.1 The production of thin films

There is a considerable number of processes that can be and are used for the deposition of optical coatings. The commonest take place under vacuum and can be classified as physical vapour deposition (sometimes abbreviated to PVD). In these processes, the thin film condenses directly in the solid phase from the vapour. The word ‘physical’ as distinct from ‘chemical’ is intended to indicate the absence of any chemical reactions in the formation of the film. This is an oversimplification. Chemical reactions are, in fact, involved but the term chemical vapour deposition (sometimes abbreviated to CVD) is reserved for a family of techniques where the growing film differs substantially in composition and properties from the components of the vapour phase.

The physical vapour deposition processes can be classified in various ways but the most useful classifications for our purposes are based on the methods used for producing the vapour and on the energy that is involved in the deposition and growth of the films. Vacuum, or thermal, evaporation has for years been the principal physical vapour deposition process and because of its simplicity, its flexibility and its relatively low cost, and because of the enormous number of existing deposition systems, it is likely to continue so for some considerable time. It is, however, clear that it possesses major shortcomings, especially in respect of the microstructure of the films, and, particularly for high-performance specialised coatings, alternative processes, such as sputtering, are being adopted. In thermal evaporation, the material to be deposited, the evaporant, is simply heated to a temperature at which it vaporises. The vapour then condenses as a solid film on the substrates, which are maintained at temperatures below the melting point of the evaporant. Molecules travel virtually in straight lines between source and substrate and the laws governing the thickness of deposit are similar to the laws that govern illumination. In sputtering, the vapour is produced by bombarding a target with energetic particles, mostly ions, so that the atoms and molecules of the target are ejected from it. Such vapour particles have much more energy than the products of thermal evaporation and this energy has considerable influence on the condensation and film-growth processes. In particular the films are usually much more compact and solid. In other variants of physical vapour deposition, the condensation of thermally evaporated material is supplied with additional energy by direct bombardment by energetic particles. Such processes, together with sputtering, are known collectively as the energetic processes.

Although physical vapour deposition is the predominant class of deposition processes in optical coatings, the application of chemical vapour deposition is gradually increasing. The chemical reactions between the starting materials, the precursors, to form the material of the coating may be triggered in various ways but the most common is probably by means of an electrically induced plasma in the active vapour. Such processes are known collectively as plasma enhanced. Chemical vapour deposition is complementary to rather than a direct competitor of physical vapour deposition. It is especially useful in the deposition

of organic polymer films that are largely beyond the capabilities of physical vapour deposition. The boundary between the two classes of process is rather blurred.

In chapter 1 we saw how the subject could be said to begin with Fraunhofer's preparing of thin films by the chemical etching of glass and also by deposition from solution. These and similar methods have been used to some extent in optical thin-film work. Other techniques that, at different stages in the development of the subject, have been, and are still sometimes, employed, include anodic oxidation of aluminium to form a protective coating and the spraying of material onto a surface either in solution or in the form of a substance that can be chemically converted into the desired material later. Even the substance itself is sometimes sprayed on, possibly after vaporisation in a hot flame. Polymerisation of monomers deposited on surfaces by condensation or from solution is also used occasionally. Extrusion of self-supporting thin-film multilayers is yet another technique.

It is impossible to cover everything, or even anything, to the depth it deserves. There is a number of books that deal specifically with processes. Useful works include Vossen and Kern [1, 2] and Glocker and Shah [3]. We shall deal primarily with physical vapour deposition and especially with thermal evaporation since that is still the staple process.

9.1.1 Thermal evaporation

In thermal evaporation the vapour is produced simply by heating the material, known as the evaporant. Because of the reduced pressure in the chamber the vapour is given off in an even stream, the molecules appearing to travel in straight lines so that any variation in the thickness of the film that is formed is smooth, and depends principally on the position and orientation of the substrate with respect to the vapour source. The properties of the film are broadly similar to those of the bulk material, although, as we shall see, there are important differences in the detailed microstructure. Precautions that have to be taken to ensure good film quality include scrupulous cleanliness of the substrate surface, near normal incidence of the vapour stream and, sometimes, heating the substrate to temperatures of 200–300 °C (or even higher, depending on the material) before commencing deposition. The evaporation is carried out in a sealed chamber that is evacuated to a pressure usually of the order of 10^{-5} mb. The materials to be deposited are melted within the chamber, using one of a number of possible techniques that will be described. The complete plant consists of the chamber together with the necessary pumps, pressure gauges, power supplies for supplying the energy necessary to melt the evaporant, monitoring equipment for the measurement of the thin-film thickness during the process, substrate holding jigs, substrate heaters and the controls. Modern thin-film coating plants are shown in figures 9.1 and 9.2.

In order to evaporate the material, it must be contained in some kind of crucible and it must be heated until molten, unless it sublimes. There is a

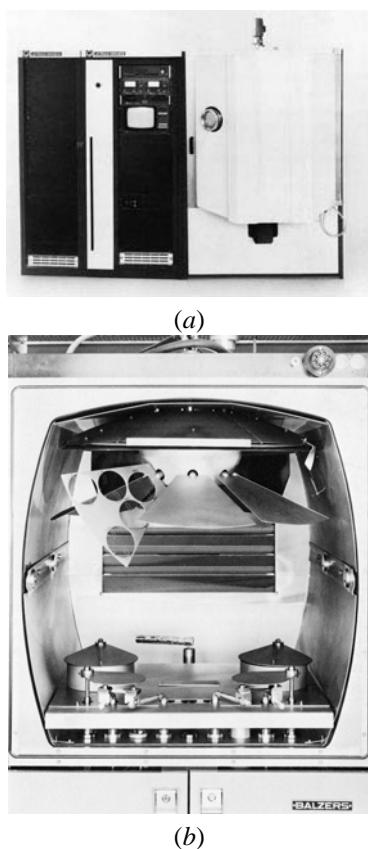
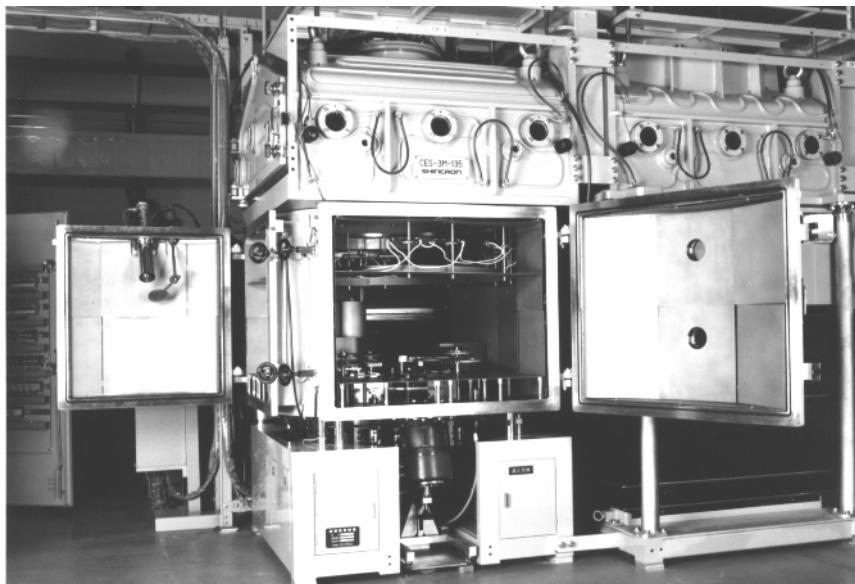


Figure 9.1. Thin-film coating machines. These are known as box coaters because the chamber is fabricated in the form of a box with a front door rather than as a bell jar on a base-plate. (a) Model A 1100 High Vacuum Deposition System. The LEYCOM process control computer is also shown. This displays on the screen the entire vacuum status of the system, the status of the evaporation process and the status of all pre- and post-deposition steps, all of these functions being computer controlled. Part of the photometer, which is used for real-time *in situ* optical thickness control, can be seen at the lower part of the front door. (Courtesy of Leybold Heraeus GmbH, Hanau, Germany.) (b) Internal view of the chamber of a BAK 760 High Vacuum Coating System. The upper part of the chamber is occupied by a reversible calotte so that substrates may be coated on both sides without breaking vacuum. The domed shape at the very top of the chamber, above the calotte, is a radiant heater. In the foreground at the base of the chamber, there are two thermal sources, each with a shutter and one charged with material. Towards the rear of the base, two electron beam sources are surrounded by circular shields and covered with shutters. The glow discharge electrode is a horizontal circular bar at the rear. (Courtesy of Balzers AG, Balzers, Liechtenstein.)



(a)

Figure 9.2. CES Series Continuous Vacuum Thin Film Coater. The operation is completely automatic. A continuous supply of jigs carrying preloaded substrates are heated under vacuum before passing into the coating chamber. Once coated they pass back out of the system and fresh jigs take their place. (a) The coating chamber. Some of the transport and heating chambers can be seen at the top of the photograph. (b) The interior of the coating chamber showing two electron-beam evaporation sources with automatic feed mechanisms for tablets on the right and granules on the left. (Courtesy of Shincron Co. Ltd, Tokyo, Japan.)

number of ways of achieving this. The simplest method is to make use of a crucible of refractory metal that acts also as a heater when an electric current is passed through it. The crucibles are elongated in shape with flat contact areas at either end and are commonly referred to as boats. Electrodes within the plant, which are insulated from the structure, act both as terminals and supports. The resistance of the boats is low and high currents, several hundred amps at low voltages, are required to heat them. Because of the high currents and especially to protect the sealing rings, the electrodes are normally water-cooled. Figure 9.4 shows a baseplate complete with a set of electrodes and figure 9.5 a molybdenum boat, mounted between electrodes, being charged with material. Tantalum, molybdenum and tungsten are all suitable for the manufacture of boats, tantalum and molybdenum being easily bent and formed, tungsten much less so. A wide range of materials can be evaporated from tantalum, and, of the three, it is the one most frequently used. However, some materials react with it (ceric oxide



(b)

Figure 9.2. (Continued)

for example) or with molybdenum, and require the less reactive but rather more difficult tungsten.

Considerable skill is required in the manufacture of tungsten boats. To avoid cracking, the tungsten strip should be heated to red heat before bending and only the simplest of shapes can be attempted. Fortunately, a wide range of preformed boats of high quality is available commercially. Certain evaporants react even with tungsten. In some cases a protective liner of alumina can be added, or an alumina crucible surrounded by a tungsten heater can even be used. In other cases, such as aluminium, the reaction is not very fast, and a tungsten wire helix is a satisfactory source. The aluminium, which wets the tungsten, forms droplets along the helix that has its axis horizontal. The area of tungsten in contact with the aluminium for a given evaporation rate is somewhat less, and the thickness of the wire somewhat greater, than for a boat, so that the tungsten is dissolved away more slowly and a greater proportion can be removed before failure. Different types of boat are shown in figure 9.6.

Materials like zinc sulphide or silicon monoxide, which sublime at not too high a temperature, can be heated in a crucible of alumina, or even fused silica, by radiation from above. A tungsten spiral just above the surface of the material can produce enough heat to vaporise it. This means that the hottest part of the material is the evaporating surface and so the material is much less prone to spitting. One example of such a source is shown in figure 9.6—the crucible is being held in the

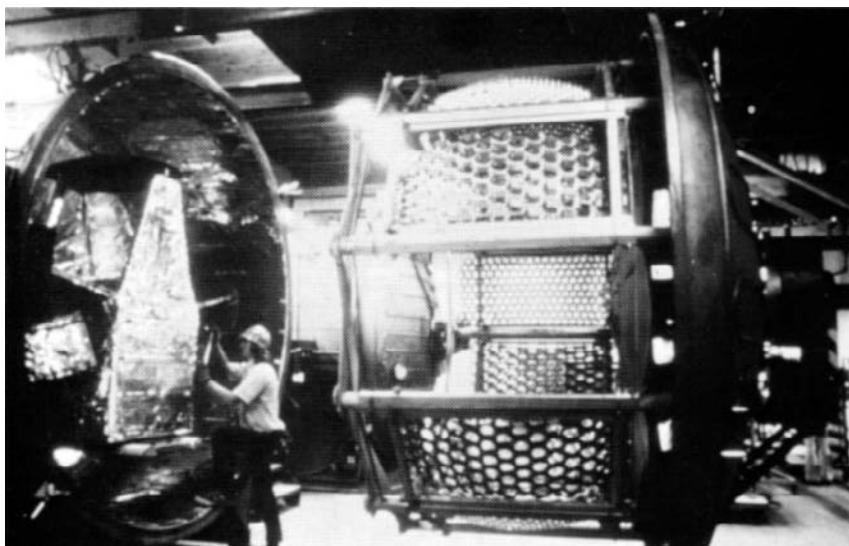


Figure 9.3. Preparing a very large plant for coating a batch of components. The substrate carrier is in the form of a horizontal drum that rotates around the sources and is carried by the chamber door that can be seen on the right. The chamber furniture, as is usual, is covered by aluminium foil for easy subsequent cleaning. (Courtesy of Optical Coating Laboratory Inc., Santa Rosa, California, USA.)

hand and the spiral is on the table. A development of this type of source is the 'howitzer' source that is shown in figure 9.7, which is particularly useful for zinc sulphide in the infrared as the capacity can be very great [4].

Germanium is an example of a material that reacts even with alumina. The reaction is not particularly fast, but the germanium films become contaminated and show higher longwave infrared absorption than is usual. Graphite has been found to be a useful boat material in this case. Supplied in rod form for use as furnace heating elements, it can be easily machined into almost any desired shape or form. Copper, graphite, or one of the refractory metals should be used to make the contacts to the graphite boats. At the high temperatures involved, steel and graphite interact so that the former tends to melt and pit badly and is, therefore, quite unsuitable.

A form of heating which avoids many of the difficulties associated with directly and indirectly heated boats is electron-beam heating, and this is now the preferred technique for most materials, especially the refractory oxides. In this method, the evaporant is contained in a suitable crucible, or hearth, of electrically conducting material, and is bombarded with a beam of electrons to heat and vaporise it. The portion of the evaporant that is heated is in the centre of the exposed surface, and there is a reasonably long thermal conduction path through



Figure 9.4. The base-plate of a thin-film coating plant showing the electrodes and the shutter used for terminating the layers.

the material to the hearth that can therefore be held at a rather lower temperature than the melting temperature of the evaporant, without prohibitive heat loss. This means that the reaction between the evaporant and the hearth can be inhibited, and the hearth is normally water-cooled to maintain its low temperature. Copper, because of its high thermal conductivity, is the preferred hearth material. The electrons are emitted by a hot filament, normally tungsten, and are attracted to the evaporant by a potential usually between 6 and 10 kV. Various types of electrodes and forms of focusing have been used at different times, but the arrangement that has now been almost universally adopted is what is known as the bent-beam type of gun. The hearth is at the ground potential and the filament is negative with respect to it. The filament and electrodes, usually a plate at filament potential situated close to the filament with a beam-defining slit through which the electrons pass, followed closely by the anode at the same potential as the hearth and incorporating a slightly larger slit so that the beam passes through it, are placed under the hearth, well out of reach of the emitted evaporant. The beam is bent around through rather greater than a semicircle by a magnetic field and focused on the material in the hearth. This avoids the problems of early electron beam systems that had filaments in line of sight of the hearth and hence considerably shortened life due to reactions with the evaporant. Supplementary magnetic fields derived from coils allow the position of the spot to be varied so that the mean can be placed in the centre of the hearth and a raster can be described which increases the area of heated material. This reduces the temperature necessary to maintain the same rate of deposition, improves the efficiency of use of the material in the



Figure 9.5. A molybdenum boat, mounted between electrodes in an Edwards E19E machine, being charged with material.

crucible and makes the electron beam source more stable. A typical electron beam source of this type is shown in figure 9.8.

The electron beam source is particularly useful for materials that react with boats or require very high evaporation temperatures, or both. Even in quite small sources, beam currents of up to 1 A at voltages of around 10 kV can be achieved and refractory oxides such as aluminium oxide, zirconium oxide and hafnium oxide, and reactive semiconductors such as germanium and silicon, can be evaporated readily. Furthermore, materials that can be evaporated quite satisfactorily by a directly heated boat can be evaporated still more easily by electron beam, and so the tendency is to use electron beam sources, once they are installed, for virtually all materials. To improve their flexibility, they can be constructed with multiple pockets in the hearth so that the same source can handle up to four different materials in a single coating cycle. Of course the capacity of each individual pocket in a multiple-pocket version is usually rather less than that of the single-pocket version of the same source. Also it is not currently possible to maintain the alternative crucibles at near evaporation temperatures implying a delay between layers as the source is brought up to temperature. For large-scale production, therefore, or for coatings for the infrared, it is normal to use two or

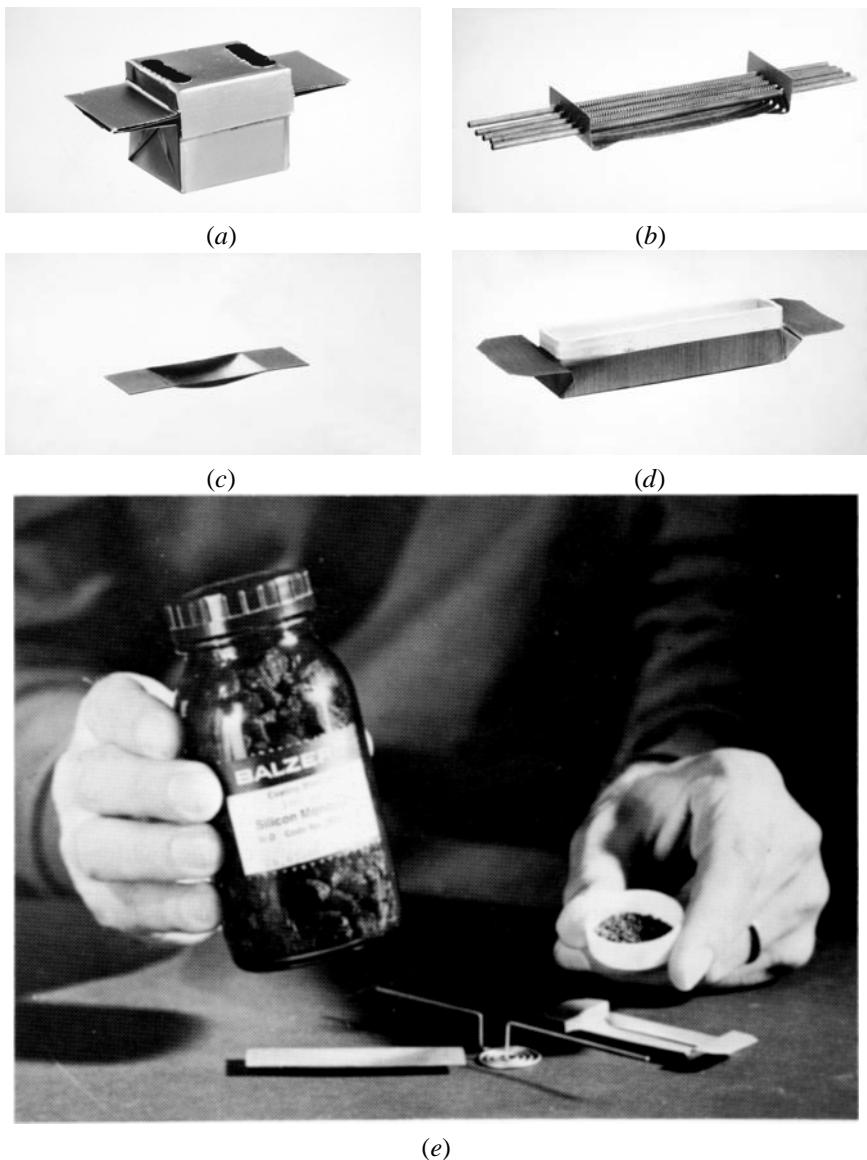


Figure 9.6. Various evaporation sources. (a) Tantalum box source (660 A, 1695 W for 1600 °C). (b) Tungsten source for large quantities of metals such as aluminium, silver and gold (475 A, 1400 W for 1800 °C). (c) Tungsten boat (325 A, 565 W for 1800 °C). (d) Aluminium oxide crucible with molybdenum heater. (e) Aluminium oxide crucible with tungsten filament. Two tungsten boats can also be seen. (Courtesy of Balzers AG.)

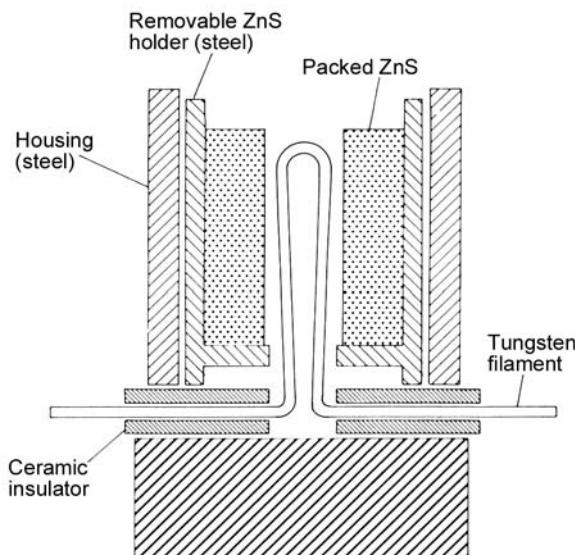


Figure 9.7. The howitzer—a source for evaporating large quantities of ZnS at high deposition rates. The removable ZnS holder shown as steel can also be made of fused silica or alumina and the hairpin filament can be replaced by a tungsten helix. (After Cox and Hass [4].)

more single-pocket sources.

The temperature of the substrate also plays a part in determining the properties of the condensed films. Usually it is the consistency of temperature from one coating run to the next which is of greater importance than the absolute level, although Ritchie [5], working in the far infrared beyond $12 \mu\text{m}$, found substrate temperature to be of critical importance and devised ways of controlling it to within 2°C of the experimentally determined optimum. Substrates are often of low thermal conductivity and are mounted on rotating jigs to ensure uniformity of film thickness so that the measurement of the absolute temperature of the substrates is difficult. The heating is usually by means of radiant elements placed a short distance behind the substrates or by tungsten halogen lamps placed so that they illuminate the front surfaces of the substrates, the latter method gaining in popularity. Measurement is most often carried out by placing a thermocouple just in front of the substrate carrier. This will not measure substrate temperature accurately but will give an indication of the constancy of process conditions; frequently this is all that is wanted, anyway. An improvement can be obtained by embedding the thermocouple in a block of material of the same type as the substrates. Thermocouples have been placed on the rotating jig and the signal led out through silver slip rings, but even in this case the temperature of the

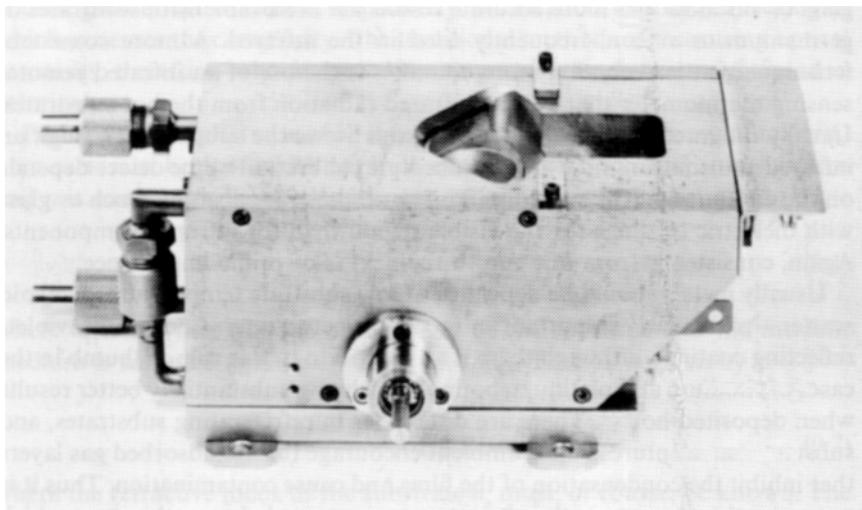


Figure 9.8. A four-pocket 'supersource'. This is an electron-beam source of the bent-beam type. The water-cooled crucible has four pockets that can be rotated into position at the focus of the electron beam that issues from the slot to the right of the opening in the top of the gun. The sides of the gun are the pole pieces of the focusing and deflecting magnet. (Reproduced by kind permission of Temescal, Berkeley, California, USA, a division of the BOC Group Inc.)

front surface of the substrates is still not necessarily known to any high degree of accuracy, especially if they are of material of low thermal conductivity such as glass or silica. Rather more accurate results are achievable with substrates of germanium or silicon, frequently used in the infrared. A more consistent technique that is becoming more common is the use of an infrared remote sensing thermometer that detects infrared radiation from the hot substrates. Usually mounted outside the chamber, this views the substrates through an infrared-transmitting window. The absolute calibration of the device depends on the emittance of the substrate. This varies less for substrates such as glass with dielectric coatings for the visible region than for infrared components. Again, consistency from one run to the next is of prime importance.

Usually metals should be deposited at low substrate temperatures to avoid scatter—particularly important in metal–dielectric filters and in ultraviolet-reflecting coatings, although there is an exception to this rule of thumb in the cases of rhodium and platinum, both of which give substantially better results when deposited hot [6, 7]. There are difficulties in refrigerating substrates, and substrate temperatures below ambient encourage thicker adsorbed gas layers that inhibit the condensation of the films and cause contamination. Thus it is not normal to operate with substrate temperatures below ambient, at which

adequate results are obtained. The softer dielectric materials such as zinc sulphide and cryolite can also be deposited at room temperature (except, as we shall see, if zinc sulphide is to be used in the infrared). The harder dielectric materials, however, usually require elevated substrate temperatures, often 200–300 °C. These materials include ceric oxide, magnesium fluoride and titanium dioxide. Some of the semiconductors for the infrared must be similarly treated. Frequently, optimum mechanical properties demand deposition at a temperature that is different from that for optimum optical properties and a compromise that depends on the particular application is necessary. Further details will be given when individual materials are discussed.

9.1.2 Energetic processes

The energetic processes, as the name suggests, are ones that involve energies rather greater than thermal. Thin films deposited by thermal evaporation have a pronounced columnar structure that is a major cause of coating instability and drift. This is discussed later in this chapter. The idea behind the energetic processes is to disrupt the columnar structure with its accompanying voids by supplying extra energy, and this does work well. Some of the energetic processes are old ones that have always involved extra energy and are now recognised as having certain advantages because of it. Although we describe the processes as energetic, it has been shown that momentum is the important quantity.

Sputtering is an old process that predates thermal evaporation. Momentum transfer from incident energetic ions is used to eject atoms and molecules from a target into the vapour phase. The kinetic energy and momentum of the ejected particles are high and so the growing film is subjected to a much greater impulse each time a fresh particle arrives, which disrupts the void and columnar structure. In the conventional form of sputtering, the target is metallic so that it conducts and the bombarding ions are derived from a DC discharge in the vicinity of the target. This discharge may be confined by crossed electric and magnetic fields when it is known as magnetron sputtering and this is the most common way in which the process is applied in optical coating. DC planar magnetron targets are most common; figure 9.9 shows a schematic form of such a target. The great advantage of magnetron sputtering is the much longer path length of the electrons so that the discharge can be maintained at a considerably lower pressure (0.3 Pa or 0.3×10^{-2} mb for example) than is required compared with conventional sputtering in the absence of the magnetic field.

There are, however, some disadvantages. The arrangement of magnets concentrates the discharge in the region between the pole pieces and the erosion of the target is greatest there, while other areas of the target show negligible erosion. With long rectangular targets, the appearance of the eroded region is not unlike the shape of a race track, a term often used to describe it. Target utilisation is therefore not good and so used targets are usually recovered rather than scrapped. Since the targets in DC magnetron sputtering are metallic, a process of reactive sputtering

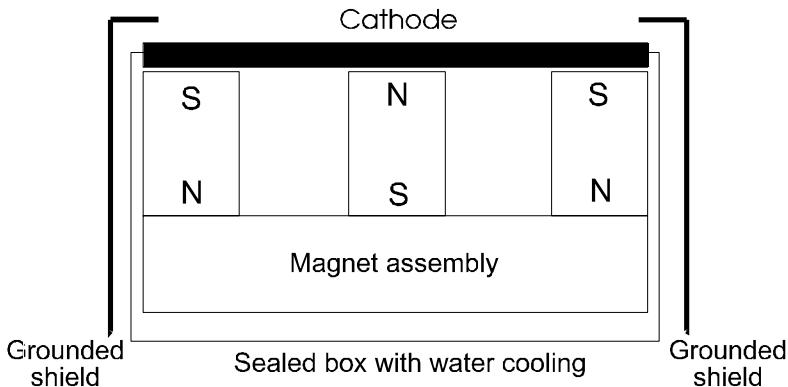


Figure 9.9. Schematic representation of a planar magnetron source. The target or cathode is connected to the negative supply. The structure of the coating machine including the grounded shield is the positive side of the supply. Electrons leaving the cathode surface move outward but are turned into a cycloidal path by the field of the magnets. The polarity of the magnets is unimportant as long as they are arranged with the outer poles opposite to the inner as shown.

must be used to produce oxides or nitrides and the sputtering gas, therefore, is usually a mixture of a noble gas such as argon and oxygen or nitrogen. This reactive gas reacts also with the target to produce a skin of oxide or nitride and the skin tends to build up in the less eroded regions. Electrons are very mobile and tend to collect on the surface of this skin charging it up like a capacitor that can discharge suddenly and violently. This arcing tends to produce molten droplets of material that are often embedded in the film. In the worst case the discharge can actually damage the target so badly as to render it unusable. The insulating skin also modifies the electrical properties of the sputtering system so that hysteresis appears making control difficult. These effects are particularly severe with silicon targets, and silicon oxide is the sole low-index material really suitable for sputtering. The problem is often called target poisoning.

There are several current solutions to the target poisoning problem. The target surface may be moved with respect to the magnets so that the region of high erosion moves over the surface and cleans it up. In the usual embodiment the target is made in cylindrical form and rotated about a longitudinal axis around the magnets and inside the grounded shield.

Another more recent form of solution involves twin magnetron targets that are connected to opposite poles of a mid-frequency power supply. The targets are now alternately the anode and cathode of the system. This discharges the effective capacitors before they can cause damage and also solves the problem of the disappearing anode. In normal single-target sputtering the chamber structure is the anode of the supply. The build up of insulating film over this structure

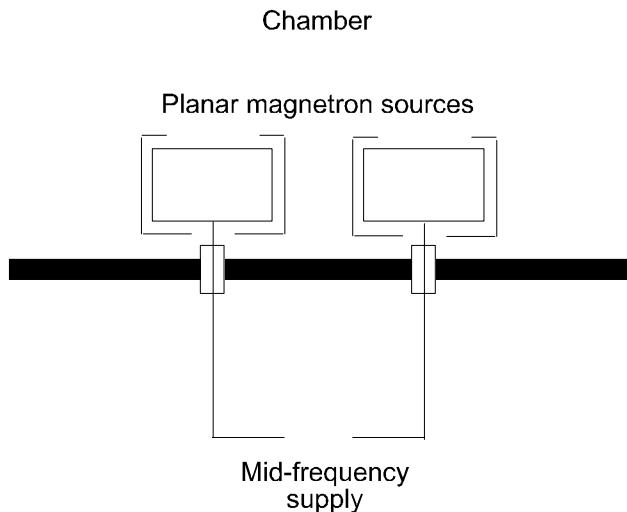


Figure 9.10. The twin magnetron arrangement in which two magnetron targets are connected to a mid-frequency power supply so that each is alternately anode and cathode. The arrangement avoids the charging problems of reactive DC sputtering without the complications of radio-frequency sputtering.

gradually makes the anode less and less effective with all kinds of implications for both control and deposition. The twin magnetron solution avoids this problem because the alternate source is the anode. The frequency is usually of the order of 40 kHz, high enough to avoid the charging problems but low enough so that the targets are effectively operating in the DC regime. Usually the twin magnetrons are planar but the process has also been used with rotating magnetrons.

Two other solutions are worthy of mention. The oxidation or nitriding may take place remote from the deposition. This requires that only a small amount of material be deposited then treated, then more deposited and then treated, and so on. The process is implemented by placing the substrates on a cylindrical drum that is then rotated rapidly and continuously past a linear magnetron sputtering source then past an ion source and round to the magnetron target again. This process is known as metemode, short for metal mode and is the subject of an issued patent [8]. An alternative process places the magnetron source inside a shroud where it can be operated in argon. The material escapes through a large aperture above the source in the centre of the shroud. Outside the shroud in the main chamber the material coats the substrates but the growing film is also bombarded with a beam of oxygen or nitrogen ions in the manner of ion-assisted deposition, described shortly. Enormous quantities of gas enter the deposition chamber and to remove the gas very fast, high capacity pumps are used. The films that grow are amorphous of very high packing density. This process is known as

microplasma and, at the time of writing, very little is known publicly about it except for an issued patent [9]. An advantage of the process appears to be that the geometry of the coating chamber can be similar to that for thermal evaporation. Presumably the increased positional stability of the magnetron sources is a further advantage.

Radio-frequency (RF) sputtering is a process that avoids the problems of an insulating target. It is much used in other areas of thin-film deposition but has not been popular in optical coatings mainly because of all the additional problems of radio-frequency systems such as screening and matching. At radio frequencies even a straight length of conductor can have an appreciable impedance so that grounding is a much greater problem than at low frequency. It is also a somewhat slower process. Nevertheless, in applications where speed is less important than quality it has been found remarkably reliable and stable, to the extent where even quite complex coatings can be controlled entirely by power, gas pressure and time with no ongoing layer thickness measurement whatsoever [10].

The most advanced form of sputtering uses a separate chamber to generate the ions that are then extracted and directed towards the target. This is known as ion-beam sputtering [11]. It is capable of a very high degree of film purity and the lowest published losses in optical coatings, 1 ppm or less, have been achieved with this process [12, 13]. Since the ion beam is usually neutralised by adding electrons, charging problems with insulating targets can be avoided and the process is as useful for insulating materials as for conductors. Ion-beam sputtering is slow compared with most other processes and it is not able to cope with deposition over large areas. It has not been generally adopted and its use is largely limited to special coatings where low loss is the important criterion.

Not all materials are suitable for sputtering. In particular the fluorides present considerable difficulties because of preferential sputtering of fluorine atoms. The film is then fluorine deficient and optically absorbing. The fluorine vacancies can be filled with oxygen—there is usually plenty of oxygen around—which removes the absorption, at least at longer wavelengths, but the film becomes an oxyfluoride with altered (usually raised) index of refraction and frequently degraded environmental resistance.

In reactive low-voltage ion plating [14, 15], a high-current beam of low-voltage electrons is directed into the region above the hearth in an electron beam source. This results in a very high degree of ionisation of evaporant material, usually a metal or suboxide so that the melt is conducting. Reactive gases, oxygen or nitrogen, fed separately into the chamber, are also highly ionised. There is a complete circuit from ion gun to electron beam source and back and it is completely isolated from the rest of the structure. The substrate carrier is also electrically isolated. There are many electrons and they are very mobile and so the isolated substrates acquire a charge that is negative with respect to the electron beam source. This attracts the positive ions from the source so that they arrive at the film surface with additional momentum that is transferred to the film and compacts it. Films are tough, hard and dense and usually amorphous. Because of



Figure 9.11. The Plasmacoat is a small machine intended principally for the coating of spectacle lenses but it can also be used for small batches of other types of coating. The process is one of reactive sputtering and the operation is entirely automatic. The coating chamber is permanently under vacuum. For loading, the substrate carrier drops down into the loading chamber leaving the coating chamber sealed off. The carrier can then be loaded through the access door. Once substrates are loaded the access door closes and the substrate carrier moves upwards back into the deposition chamber. (Courtesy of Applied Vision Ltd, Coalville, Leicestershire, England.)

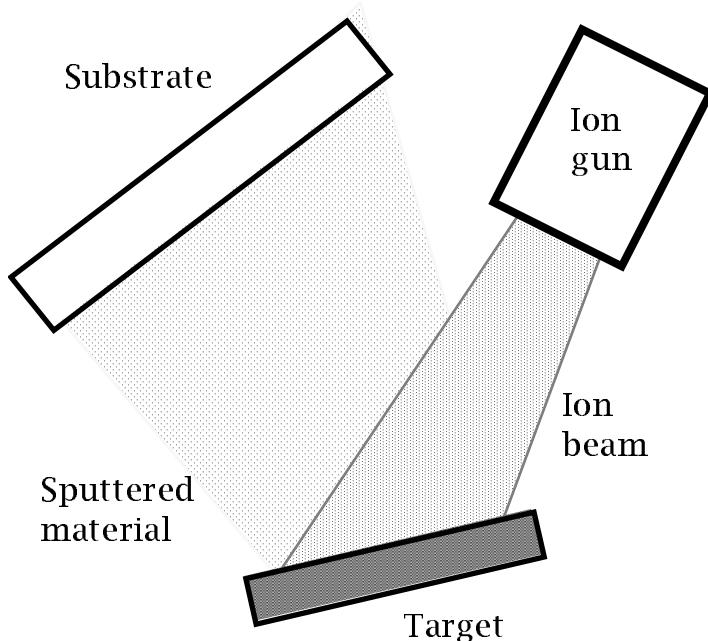


Figure 9.12. Ion-beam sputtering schematic. The ion-generating discharge is within the ion gun and therefore removed from the deposition chamber. This gives much higher quality films.

the very efficient reaction with the additional gas they are of high optical quality.

Ion-assisted deposition is an energetic process that has the great advantage that it is easy to implement in conventional equipment. It consists of thermal evaporation to which has been added bombardment of the growing film with a beam of energetic ions. All that is required to put it into operation in a conventional plant, therefore, is the addition of an ion gun. The most common types of ion sources for this purpose are broad-beam, often with extraction grids. Much of the published research and reported successes have been with the Kaufman or gridded type of ion gun. In that, the source of electrons is a hot filament and the extraction system consists of two closely aligned grids, the inner floating and acquiring the potential of the discharge so that it confines it within the gun, and the second applying a field to draw the positive ions out of the discharge chamber through the apertures in the inner grid. The beam of ions is neutralised outside the discharge chamber by adding electrons, usually from a hot filament, immersed in the beam to avoid space charge limitation, or from a separate hollow-cathode electron emitter. The grids are fragile and easily misaligned or damaged and so some effort has been put into the development of sources that do not require extraction grids and they are being used in increasing numbers in production. For



Figure 9.13. A Spector ion-beam sputtering system for the production of high-quality optical coatings especially narrowband filters for dense wavelength division multiplexing. (Courtesy of Ion Tech, Inc., Fort Collins, Colorado, USA.)

further information see Bovard [16] and Fulton [17].

The ionised plasma-assisted deposition process includes features of both ion-assisted deposition and low-voltage ion plating. It makes use of what is known as an advanced plasma source [18–20]. The source, which is insulated from the chamber and floats in potential, is of simple construction. A central indirectly heated cathode is made of lanthanum hexaboride. This lies along the axis of a vertical cylinder that is the anode. A noble gas, usually argon, is introduced into the source. The cylinder contains a solenoid that produces an axial magnetic field. The crossed electric and magnetic fields make the electrons move in cycloids with the usual increase in path length and degree of ionisation, so that an intense plasma is produced in the source. The fields do not confine the plasma axially and so it escapes from the source into the chamber. There the electrons, that are very mobile, escape preferentially to the chamber structure leaving the plasma charged positively without the need for isolated substrate holders. The deposition sources are thermal, usually electron beam, and they emit evaporant into the plasma where it gains energy and is partially ionised. The evaporant then condenses on the growing film with additional energy, as in ion plating, and is bombarded

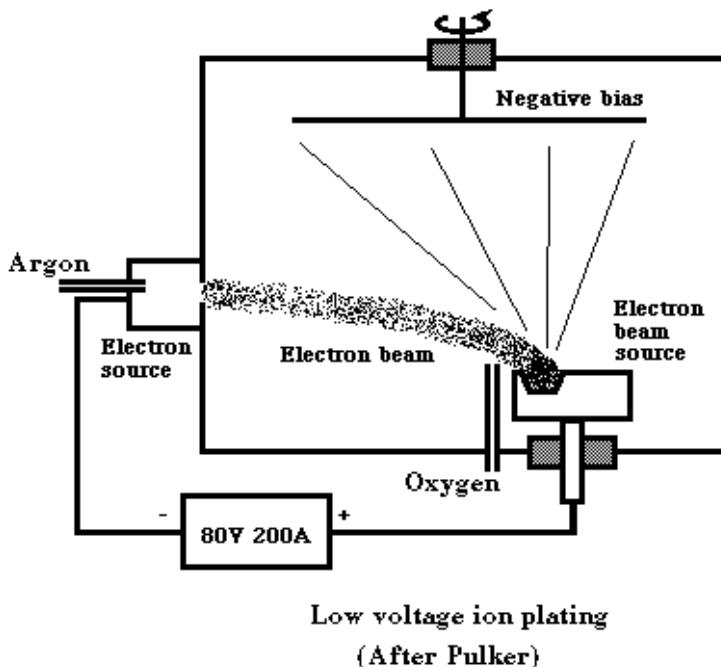


Figure 9.14. The low-voltage ion plating process. The negative bias on the electrically isolated substrates is acquired from the free electrons in the chamber. (After Pulker. See for example [15].)

simultaneously by ions from the plasma as in ion-assisted deposition. For reactive processes, the reacting gas is not fed into the source but into the plasma as it leaves the source. A ring-shower-shaped inlet tube is positioned just above the aperture of the source for this purpose. The process has been very successful in the production of narrowband filters for dense wavelength division multiplexing.

It seems clear that the major benefit of the energetic processes is an increase in film packing density. The improvements are achieved at comparatively low substrate temperatures which helps with the difficult coating of plastic substrates.

It has been theoretically demonstrated by advanced computer modelling [21,22] that the major effects are due to the additional momentum of the molecules, either supplied by collisions with the incoming energetic ions, or derived from the additional kinetic energy of the evaporant. Experimental evidence exists [23] that shows correlation of the effects with momentum rather than energy of the bombarding ions. Major benefits of these processes are the increased packing density of the films, making them more bulk-like and hence increasing their ruggedness, the improved adhesion resulting from a mixing of materials at the interfaces between layers, and a reduction of the sometimes

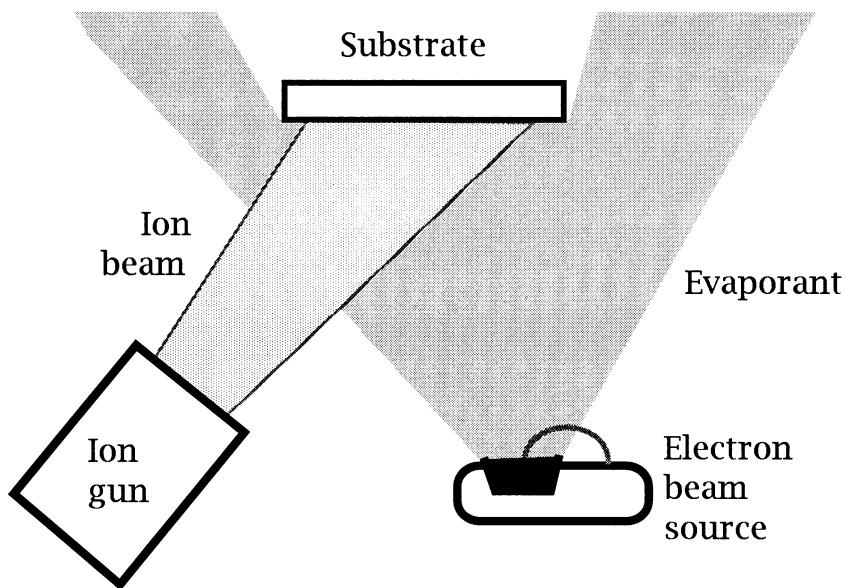


Figure 9.15. The addition of ion bombardment of the growing film transforms conventional thermal evaporation into ion-assisted deposition.

quite high tensile stress in the layers. The increase in packing density reduces also the moisture sensitivity and can actually eliminate it altogether [24]. The increased packing density also improves the stability of the films in other ways. Magnesium fluoride films resist high temperature oxidation better, for example [25]. The hardness and corrosion resistance of metal films, especially with dielectric overcoats [26], is improved by ion-assisted deposition but the optical properties tend to be slightly adversely affected, possibly by the implantation of a small fraction of the bombarding ions [27]. The increased reactivity of the bombarding ions permits the deposition of compounds, such as nitrides [28], that are difficult or impossible by normal vacuum evaporation.

9.1.3 Other processes

Physical vapour deposition processes are those most often used for the production of optical coatings. However, in the electronic device field, chemical vapour deposition is the principal method for thin-film deposition and there is increasing interest in it for optical purposes, usually with regard to very special requirements.

Chemical vapour deposition differs from physical vapour deposition in that the film material is produced by a reaction amongst components of the vapour that surrounds the substrates. The reaction may be induced by the temperature

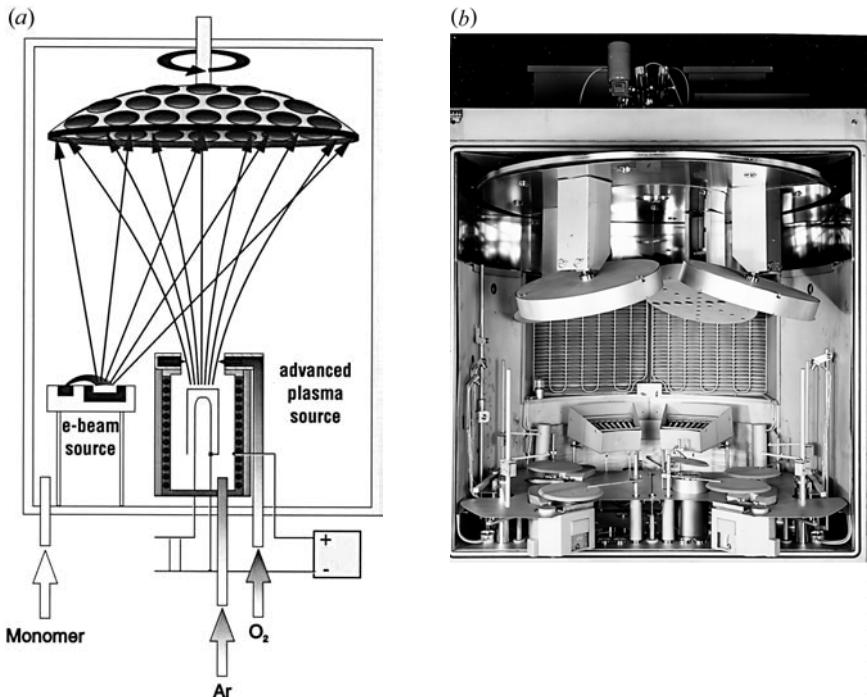


Figure 9.16. The advanced plasma source. (Courtesy of Leybold AG, Hanau, Germany.) (a) Diagram of the advanced plasma source (APS) and the arrangement of the machine for plasma ion-assisted deposition (PIAD). The monomer inlet shown is used in the construction of the final anti-smudge coat in the coating of spectacle lenses. (b) Photograph of the interior of the system showing the electron-beam sources and just slightly to the right of the centre the cylindrical advanced plasma source.

of the substrates themselves, when the process is the classical thermal chemical vapour deposition, or, and this is more usual in the optical field, it may be a plasma-induced process.

Usually the components, the reactants or precursors, will be introduced into a carrier gas that is permitted to flow through the system. This ensures a constant supply of the reactants to the growing interface and allows sufficient dilution so that the reaction is not so fast as to overwhelm the film growth.

In this classical form of chemical vapour deposition great problems are created by reactions that are too efficient. A reaction that proceeds rapidly tends to produce a film that is poorly packed and poorly adherent. The term *snow* is often used to describe it. The reactions must, therefore, be quite weak and this means that impurities that have strong reactions can play havoc with the process and severely limit the possible range of processes.

Because of all the difficulties, the classical thermal chemical vapour deposition process is not often used for optical coatings. Instead, pulsed processes have been largely adopted. Material added to a thin film is assimilated provided it is not immobilised by material deposited over it before it has had time to relax into favourable positions. The problem is not really the strength of the reaction but rather the large amount of material that arrives in a given time. Earlier material is buried under the weight of later material and cannot relax to a state of equilibrium, and snow is the result. If an efficient reaction can be made to deliver material at a correct rate then the film will be dense. It is the overall rate of deposition that determines the microstructure. Pulsing the reaction gives the control of rate that is required. The pulsing can most conveniently be achieved when a plasma-assisted process is involved [29].

A related process that is sometimes called plasma polymerisation, and sometimes plasma-enhanced (or induced) chemical vapour deposition or PECVD [30–32] is used to deposit dense organic layers with stable optical properties over curved and irregular surfaces with good uniformity. Plasma polymerisation is quite unlike normal polymerisation where monomers are linked into chains of repeat units. The plasma is characterised by energetic electrons that break the reactants into active fragments and these fragments link with each other to form the deposited film. Some of this combination may take place in the gaseous phase forming clusters that may deposit on the growing film or may be broken into fragments again by the plasma. Strong binding occurs so that the deposited film is tough and hard and dense. It is not strictly polymeric and contains free radicals that may combine with any oxygen that is also present. The mechanical properties can range from plastic to elastic and glass-like. Because the films are insulating, in fact they are used as capacitor dielectrics in some applications, RF discharges are usual for this process. Speed of deposition can be very high, up to $1 \mu\text{m min}^{-1}$ although rates of one-tenth to one-hundredth of this are more common.

The process has been used for some time in the semiconductor industry to deposit silicon dioxide. The normal precursor is tetraethoxysilane (TEOS) together with oxygen but the substrate temperature is usually quite high, at least 250°C , much higher than can be possible for plastic substrates. When the temperature is reduced to permit coating of plastic substrates, the film composition becomes much more complicated. Apart from the silicon oxide content they include, for example, silanol that results from reactions involving residual water vapour. There are, in fact, many silicone compounds that can be and have been used as precursors in the PECVD deposition of such silica-rich films. The feature that they tend to share is a backbone of alternate silicon and oxygen atoms. Apart from the tetraethoxysilane already mentioned other suitable compounds include hexamethyldisiloxane (HMDSO), tetramethoxysilane (TMOS), methyltrimethoxysilane (MTMOS) and trimethylmethoxysilane (TMMOS). As might be expected, they are toxic, although their toxicity varies. The make-up of the precursors determines to a large extent the character of the film. With organic silicone compounds or silanes present in the gas along with oxygen the

coatings are particularly tough and resistant to abrasion and form the basis for a number of different hard coats. The name hard coat is normally given to an initial layer over a plastic substrate that acts as a transition between the organic plastic and an overlying essentially inorganic optical coating. Fluorine compounds give films that have very low friction and are hydrophobic and are frequently used as the outermost anti-smudge layer in an antireflection coating. The precise details of the precursors are difficult to obtain. They are considered part of the know-how of the process.

There are many other techniques for the deposition of optical coatings. Probably the most important of these is the sol-gel process. The name sol-gel refers to those processes that involve a solution that undergoes a transition of the sol-gel type, that is, a solution is transformed into a gel. The common form of the sol-gel process starts with a metal alkoxide. This organometallic compound is hydrolysed when it is mixed with water in an appropriate mutual solvent. The solution is usually made slightly acidic to control the rates of reaction and to help the formation of a polymeric material with linear molecules. The result is a gradual transition to an oxide polymer with liquid-filled pores. This gel can be deposited over the surface of an optical component by dipping. The coating is then heat treated to remove the liquid in the pores and to densify it; the higher the temperature to which it is raised, the denser is the film. By treating the gel film at temperatures as high as 1000 °C complete densification is achieved. Lower temperatures give partial densification but already by 600 °C the film is largely impermeable. Typical materials are TEOS (tetraethylorthosilicate, $\text{Si}(\text{OC}_2\text{H}_5)_4$) for eventual films consisting of silica, and titanium tetraethoxide ($\text{Ti}(\text{OC}_2\text{H}_5)_4$) for films of titanium oxide. These materials are dissolved in ethanol and then hydrolysed by adding a little distilled water. In the case of the titanium compound, the rate of hydrolysis is much faster and so nitric acid is added to control the transformation and so the solution is made rather weaker.

There are quite considerable difficulties in producing multilayer coatings by the sol-gel process, and so, apart from some applications involving high durability antireflection coatings of a few layers, the process has never competed successfully with vacuum deposition.

Interest in the sol-gel process increased enormously when it was discovered that sol-gel deposited antireflection coatings had exceptionally high laser damage threshold [33]. The technique is much used therefore in producing antireflection coatings for components in the very large lasers for fusion experiments. These coatings are unbaked and quite porous, otherwise the refractive index would not be suitable for antireflection coatings for low-index materials. In uncontrolled environments such porous coatings take up moisture and other contamination and their index tends to vary over a period of time and their performance falls. Regular coatings must be baked at high temperature. However, the environment of the large lasers is tightly controlled and the fragility of the unbaked coatings can be viewed as an advantage if they have to be removed to permit recoating of the component.

9.1.4 Baking

A final stage of the manufacturing process for optical coatings that is seldom discussed is that of baking. This is probably the one aspect of coating production that might still be referred to as an art rather than a science. Baking consists of heating the coated component normally in air at temperatures of usually between 100 °C and 300 °C for a period of perhaps several hours.

A common reaction in most coating departments to a batch of coatings that exhibit less than acceptable properties is to bake the coatings in air for a time simply to see if their properties improve. They frequently do. There is no doubt that such treatment can improve the properties of the coatings in several respects.

Coated substrates that are to be used as laser mirrors cemented to laser tubes are almost invariably baked before mounting because it is believed that this increases their stability. There is no doubt that such treatment does reduce the drift that may occur at the early stages of laser operation but the reason for this is obscure.

Frequently the absorption in the layers falls. This may be simply a case of improved oxidation. We know that baking of titanium suboxides in air improves their transmittance and reduces their absorptance [34]. High-quality films are frequently amorphous and prolonged baking may induce a slow amorphous-to-crystalline transition in such films. This process may compete with the oxidation process so that an optimum period of baking may result. This may be one reason why details of baking are frequently considered proprietary.

Most of the work that has been reported on baking is with regard to narrowband filters frequently constructed from zinc sulphide and cryolite. Meaburn [35] was a particularly early worker in this area. He found that a process of baking at 90 °C for ten hours improved the stability of narrowband filters of zinc sulphide and cryolite enormously. This was especially so if they were protected afterwards by a cemented cover slip.

Title *et al* [36] reported a baking process called a *hard bake* with filters similar to those described by Meaburn. In the hard bake, filters were subjected to temperatures around 100 °C for a certain time. During the baking process the peak wavelength moved towards shorter wavelengths. After a critical time the rate of movement suddenly slowed and the filter became much more stable. Details of the shift and the time were considered proprietary and not included in the published account. This is consistent with a desorption process coupled with a diffusion process to be described shortly.

Richmond [37] and Lee [38] both conducted baking experiments on narrowband filters. They were interested in absorption and desorption processes in thin films. They found that the baking process did not appear to alter the amount of moisture absorbed and desorbed by the filters. The stability of the characteristic, in the sense of the total change for a given change in relative humidity, was essentially unaltered. The rate of change, however, was greatly increased so that the characteristic reached equilibrium very much faster. The

filters, therefore, appeared to be much more stable in the laboratory environment.

Müller [39] constructed computer models of the annealing process in thin films. The essential features of the models were thermally activated movements of atoms from a filled site to an available neighbouring and vacant site. He found that packing density did not change during this process but that there was a quite definite amalgamation of smaller voids into larger ones. This process appears to be a wandering of the voids through the material of the thin film but is really a process of surface diffusion around the interior of the voids. Once two voids meet there is an energetic advantage in combining but, once combined, no advantage in splitting. Thus the voids simply increase gradually in size as they reduce in number. The reason for the findings of Richmond and Lee, and probably also Title and Meaburn, now become clear. After deposition, the pore-shaped voids in the material are quite irregular in shape, especially at the interfaces between the layers. The annealing or baking process tends to remove the restrictions in the pores so that although their volume is unchanged their regular shape implies a much faster filling by capillary condensation when exposed to humidity. This means that equilibrium is reached much more rapidly and the filter appears much more stable when the environmental conditions are stable. In the case of already cemented filters the effective environment is quite stable although the filter stability may be disturbed by changes in temperature. However, when the temperature stabilises, equilibrium is rapidly established once again.

The improved stability of the integral laser mirror is probably also derived at least partly from this decrease in the time constant for it to reach equilibrium. Any drift of the mirrors after alignment in the laser would immediately cause fluctuations, almost invariably reductions, in laser output. If the mirror can reach equilibrium before the final alignment then, since the environment within the laser is reasonably stable from the point of view of moisture and consequent adsorption, the laser will be stable.

Müller [39] has also explained why it is that baking never seems to improve poor adhesion but invariably makes it worse. Here if the bonds that bind atoms together across an interface are weaker than those that bind similar atoms together in either material, then there is an energetic advantage for a void that reaches an interface to remain there. Voids therefore collect at such an interface and gradually weaken the adhesion further.

Much more work is required on the whole matter of baking and consequent filter stability before all becomes completely clear, but the oven is already an indispensable apparatus in virtually all coating shops.

We return to the matter of moisture adsorption in chapter 10.

9.2 Measurement of the optical properties

Once a suitable method of producing the particular thin film has been determined, the next step is the measurement of the optical properties. Many methods for this

exist and a useful earlier account is given by Heavens [40]. Measurement of the optical constants of thin films is also included in the book by Liddell [41]. A more recent survey is that of Borgogno [42]. Recently, the measurement of the optical properties of thin films has increased in importance to the extent that special purpose instruments are now available. These normally include the extraction software and are essentially push-button in operation. As always, however, even when automatic tools are available some understanding of the nature of the process and its limitations is still necessary. Here we shall be concerned with just a few methods that are frequently used.

In all of this it is important to understand that we never actually measure the optical constants n and k directly. Although thickness, d , is more susceptible to direct measurement, its value too is frequently the product of an indirect process. The extraction of these properties, and others, involves measurements of thin-film behaviour followed by a fitting process in which the parameters of a film model are adjusted so that the calculated behaviour of the model matches the measured data. The adjustable parameters of the model are then taken to be the corresponding parameters of the real film. The operation is dependent on a model that corresponds closely to the real film. The appropriateness of the model would be of less importance were we simply trying to recast the measurements in a more convenient form. Even an inadequate model with parameters appropriately adjusted can be expected to reconstitute the original measurements. But the parameters extracted are rarely used in that role. Rather they are used for predictions of film performance in different situations where film thickness may be quite different and where the film is part of a much more complex structure. This leads to the idea of *stability* of optical constants, a rather different concept from accuracy. Accurate fitting of measured data using an inappropriate model may reproduce the measurements with immense precision yet yield predictions for other film thicknesses that are seriously in error. Such parameters are lacking in stability. Stable optical constants might reproduce the measured results with only satisfactory precision but would have equal success in a predictive role. A good example might be where a film that is really inhomogeneous and free from absorption is modelled by a homogeneous and absorbing film. The extracted film parameters in this case can be completely misleading. It must always be remembered that the film model is of fundamental importance.

Almost as important as the model is the accuracy of the actual measurements. Calibration verification is an indispensable step in the measurement of the performance that will be used for the optical constant extraction. Remember that only two parameters are required to define a straight line but to verify linearity requires more. Small errors in measurement can have especially serious consequences in the extinction coefficient and/or assessment of inhomogeneity of the film. The samples themselves should be suitable for the quality of measurement. For example, a badly chosen substrate may deflect the beam partially out of the system so that the measurement is deficient or it may introduce scattering losses that are not characteristic of the film.

The calculation of performance given the design of an optical coating is a straightforward matter. Optical constant extraction is quite different. Each film is a separate puzzle. It may be necessary to try different techniques and different models. Repeat films of different thicknesses or on different substrates may be required. Some films may appear to defy rational explanation. A common film defect is a cyclic inhomogeneity that produces measurements that the usual simpler film models are incapable of fitting with sensible results. It is always worthwhile attempting to recalculate the measurements using the model and extracted parameters to see where deficiencies might lie. Because of all the caveats in this and the previous paragraphs, exact correspondence, however, does not necessarily indicate perfect extraction.

As we saw in chapter 2, given the optical constants and thicknesses of any series of thin films on a substrate, the calculation of the optical properties is straightforward. The inverse problem, that of calculating the optical constants and thicknesses of even a single thin film, given the measured optical properties, is much more difficult and there is no general analytical solution to the problem of inverting the equations. For an ideal thin film there are three parameters involved, n , k and d , the real and imaginary parts of refractive index and the geometrical thickness, respectively. Both n and k vary with wavelength, which increases the complexity. The traditional methods of measuring optical constants, therefore, rely on special limiting cases that have straightforward solutions.

Perhaps the simplest case of all is represented by a quarter-wave of material on a substrate, both of which are lossless and dispersionless, that is, k is zero and n is constant with wavelength. The reflectance is given by

$$R = \left(\frac{1 - n_f^2/n_m}{1 + n_f^2/n_m} \right)^2 \quad (9.1)$$

where n_f is the index of the film, n_m that of the substrate and the incident medium is assumed to have an index of unity. Then n_f is given by

$$n_f = n_m^{1/2} \left(\frac{1 - R^{1/2}}{1 + R^{1/2}} \right)^{1/2} \quad (9.2)$$

where the refractive index of the substrate n_m must, of course, be known. The measurement of reflectance must be reasonably accurate. If, for instance, the refractive index is around 2.3, with a substrate of glass, then the reflectance should be measured to around one-third of a per cent (absolute ΔR of 0.003) for a refractive index measurement accurate in the second decimal place. It is sometimes claimed that this method gives a more accurate value for refractive index than the original measure of reflectance since the square root of R is used in the calculation. This may be so, but the value obtained for refractive index will be used in the subsequent calculation of the reflectance of a coating, and therefore the computed figure can be only as good as the original measurement of reflectance.

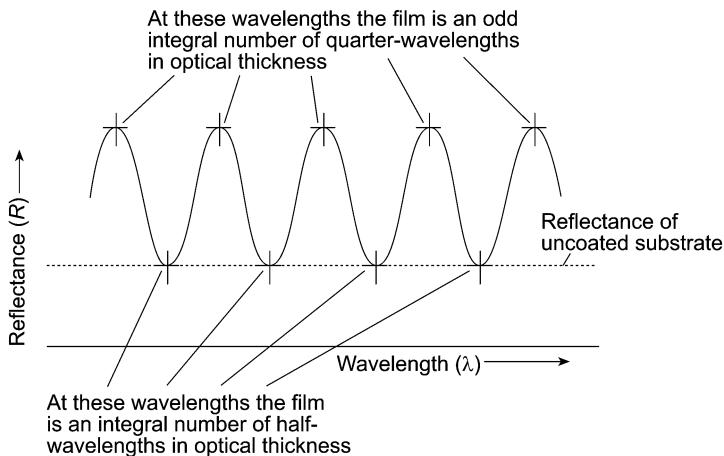


Figure 9.17. The reflectance of a simple thin film.

In the absence of dispersion, the curve of reflectance versus wavelength of the film will be similar to that in figure 9.17. The extrema correspond to integral numbers of quarter-waves, even numbers being half-wave absentees and giving reflectance equal to that of the uncoated substrate, and odd corresponding to the quarter-wave of equations (9.1) and (9.2). Thus it is easy to pick out those values of reflectance which correspond to the quarter-waves.

The technique can be adapted to give results in the presence of slight dispersion. The maxima in figure 9.17 will now no longer be at the same heights but, provided the index of the substrate is known throughout the range, the heights of the maxima can be used to calculate values for film index at the corresponding wavelengths. Interpolation can then be used to construct a graph of refractive index against wavelength. Results obtained by Hall and Ferguson [43] for MgF_2 are shown in figure 9.18.

This simple method yields results that are usually sufficiently accurate for design purposes. If, however, the dispersion is somewhat greater, or if rather more accurate results are required, then the slightly more involved formulae given by Hass *et al* [44] must be applied. It is still assumed that the absorption is negligible. If the curve of reflectance or transmittance of a film possessing dispersion is examined, it will easily be seen that the maxima corresponding to the odd quarter-wave thicknesses are displaced in wavelength from the true quarter-wave points, while the half-wave maxima are unchanged. This shift is due to the dispersion, and measurement of it can yield a more accurate value for the refractive index. In the absence of absorption the turning values of R , T , $1/R$ and $1/T$ must all coincide. Assuming that the refractive index of the incident medium is unity, that

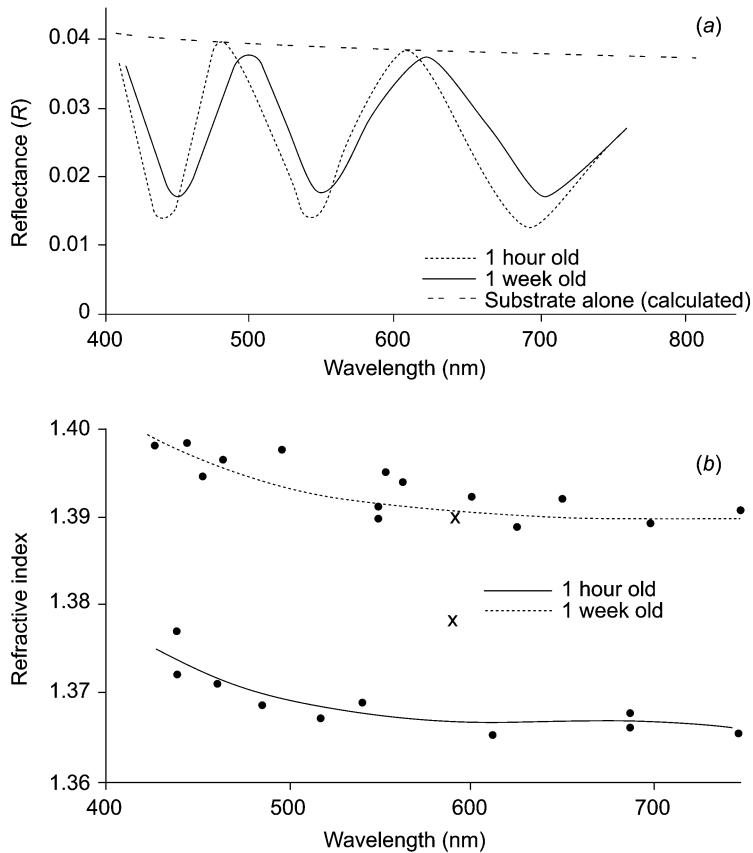


Figure 9.18. The refractive index of magnesium fluoride films. (a) The reflectance of a single film. (b) The reflectance result transforms into refractive index. The curves are formed by the results from many films. x denotes bulk indices of the crystalline solid. (After Hall and Ferguson [43].)

of the substrate n_m and of the film n_f then their expression for T becomes

$$T = \frac{4}{n_m + 2 + n_m^{-1} + 0.5n_m^{-1} \left(n_f - 1 - n_m^2 + n_m^2 n_f^{-2} \right) [1 - \cos(4\pi n_f d_f / \lambda)]}.$$

Since the turning values of T and $1/T$ coincide, the positions of the turning values can be found in terms of d/λ by differentiating the expression for $1/T$ and equating it to zero as follows:

$$\frac{1}{T} = \frac{n_m + 2 + n_m^{-1}}{4} + \frac{1}{8n_m} \left(n_f - 1 - n_m^2 + n_m^2 n_f^{-2} \right) \left(1 - \cos \frac{4\pi n_f d_f}{\lambda} \right)$$

i.e.

$$0 = \frac{d(1/T)}{d(t/\lambda)} = 0.25n'_f \left(n_m^{-1}n_f - n_s n_f^{-3} \right) \left(1 - \cos \frac{4\pi n_f d_f}{\lambda} \right) + 0.5\pi \left(n_m^{-1}n_f^2 - n_m^{-1} - n_m + n_m n_f^{-2} \right) \left(n_f n'_f \frac{t_f}{\lambda} \right) \sin \frac{4\pi n_f d_f}{\lambda}$$

where $n'_f = dn_f/d(d/\lambda)$. That the equation is satisfied exactly at all half-wave positions can easily be seen since both $\sin(4\pi n_f d_f/\lambda)$ and $(1 - \cos 4\pi n_f d_f/\lambda)$ are zero. At wavelengths corresponding to odd quarter-waves a shift does occur and this can be determined by manipulating the above equation into

$$\tan \frac{2\pi n_f d_f}{\lambda} = -2\pi \frac{n_f^5 - (1 + n_m^2) n_f^3 + n_m^2 n_f}{n_f^4 - n_m^2} \left(\frac{n_f}{n'_f} + \frac{d_f}{\lambda} \right). \quad (9.3)$$

Of course it is impossible to solve this equation immediately for n_f because there are too many unknowns. Generally the most useful approach is by successive approximations using the simpler quarter-wave formula (9.1) to obtain a first approximation for the index and the dispersion. It should be remembered that the reflection of the rear surface of the test glass should be taken into account in the derivation of the reflectance curve. It is also important that the test glass should be free from dispersion to a greater degree than the film, otherwise it must also be taken into account with consequent complication of the analysis.

If absorption is present, then formula (9.3) cannot be used. In the case of heavy absorption it can safely be assumed that there is no interference and the value of the extinction coefficient can be calculated from the expression

$$\frac{1 - R}{T} = \exp \left(\frac{4\pi k_f d_f}{\lambda} \right)$$

($4\pi k_f d_f/\lambda$ because we are dealing with energies not amplitudes) which gives [44] for k_f

$$k_f = \frac{\lambda}{4\pi d_f \log e} \log \left(\frac{1 - R}{T} \right) \quad (9.4)$$

where the two logarithms are to the same base, usually 10.

The thin-film designer is not too concerned with very accurate values of heavy absorption. Often it is sufficient merely to know that the absorption is high in a given region and the result given by (9.4) will be more than satisfactory. In regions where the absorption is significant but not great enough to weaken the single-film interference effects, a more accurate method can be used.

Equations (2.122) and (2.125) are valid for any assembly of thin films on a transparent substrate, n_m , and give

$$\frac{T}{1 - R} = \frac{\operatorname{Re}(n_m)}{\operatorname{Re}(BC^*)}. \quad (9.5)$$

For a single film on a transparent substrate, the values of B and C are given by

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} \cos \delta_f & (i \sin \delta_f) / N_f \\ i N_f \sin \delta_f & \cos \delta_f \end{bmatrix} \begin{bmatrix} 1 \\ n_m \end{bmatrix} = \begin{bmatrix} \cos \delta_f + i(n_m/N_f) \sin \delta_f \\ n_m \cos \delta_f + i N_f \sin \delta_f \end{bmatrix}.$$

Now

$$\delta_f = \varphi - i\psi = \frac{2\pi N_f d_f}{\lambda} = \frac{2\pi n_f d_f}{\lambda} - i \frac{2\pi k_f d_f}{\lambda}. \quad (9.6)$$

We shall assume the k is small compared with n and this implies that ψ will be small compared with φ . Now for φ sufficiently small

$$\cos \delta = \cos \varphi \cosh \psi + i \sin \varphi \sinh \psi \approx \cos \varphi + i\psi \sin \varphi$$

and

$$\sin \delta = \sin \varphi \cosh \psi - i \cos \varphi \sinh \psi \approx \sin \varphi - i\psi \cos \varphi$$

which yields the following expression for B and C

$$\begin{bmatrix} B \\ C \end{bmatrix} = \begin{bmatrix} [1 - (n_m/n_f)\psi] \cos \varphi - (n_m k_f / n_f^2) \sin \varphi + i[\psi + (n_m/n_f)] \sin \varphi \\ (n_m + n_f \psi) \cos \varphi + k_f \sin \varphi + i(n_f + n_m \psi) \sin \varphi \end{bmatrix}. \quad (9.7)$$

At wavelengths where the optical thickness is an integral number of quarter wavelengths, $\sin \varphi$ or $\cos \varphi$ is zero, and we can neglect terms in $\cos \varphi \sin \varphi$. The value of the real part of (BC^*) is then given by

$$\begin{aligned} \operatorname{Re}(BC^*) &= \cos^2 \varphi \left(1 + \frac{n_m}{n_f} \psi \right) (n_m + n_f \psi) + \sin^2 \varphi \left(\psi + \frac{n_m}{n_f} \right) (n_f + n_s \psi) \\ &= \left[n_m + \left(\frac{n_m^2}{n_f} + n_f \right) \psi \right] \end{aligned} \quad (9.8)$$

and when substituted in (9.5) yields

$$\frac{1-R}{T} = 1 + \left(\frac{n_m}{n_f} + \frac{n_f}{n_m} \right) \psi \quad (9.9)$$

giving for k_f (using the expression (9.6) in (9.9))

$$k_f = \left(\frac{\lambda}{2\pi d_f [(n_m/n_f) + (n_f/n_m)]} \right) \left(\frac{1-R-T}{T} \right). \quad (9.10)$$

This expression is accurate only close to the turning values of the reflectance or transmittance curves.

In the case of low absorption, the index should also be corrected. Hall and Ferguson [45] give the following expressions.

$$n_f = \left(\frac{n_m (1 + \sqrt{R})}{1 - \sqrt{R}} \right)^{1/2} + \frac{\pi k_f d_f}{\lambda} \left(\frac{1 + \sqrt{R}}{1 - \sqrt{R}} - n_m \right) \quad (9.11)$$

where R is the value of reflectance of the film at the reflectance maximum.

In the methods discussed so far, we have been assuming that the thickness of the film is unknown, except inasmuch as it can be deduced from the measurements of reflectance and transmittance, and the extrema have been the principal indicator of film thickness. However, it is possible to measure film thickness in other ways, such as multiple beam interferometry, or electron microscopy, or by using a stylus step-measuring instrument. Once there is an independent accurate measure of physical thickness, the problem of calculating the optical constants becomes much simpler. The most frequently used technique of this type was devised by Hadley (see Heavens [40] for a description). Since two optical constants, n_f and k_f , are involved at each wavelength, two parameters must be measured, and these can most conveniently be R and T . In the ideal form of the technique, if now a value of n_f is assumed, then by trial and error one value of k_f can be found, which, together with the known geometrical thickness and the assumed n_f , yields the correct measured value of R , and then a second value of k_f that similarly yields the correct value of T . A different value of n_f will give two further values of k_f , and so on. Proceeding thus, we can plot two curves of k_f against n_f , one corresponding to the T values and the other to the R values, and, where they intersect, we have the correct values of n_f and k_f for the film. The angle of intersection of the curves gives an indication of the precision of the result.

Hadley, at a time when such calculations were exceedingly cumbersome, produced a book of curves giving the reflectance and transmittance of films as a function of the ratio of geometrical thickness to wavelength, with n_f and k_f as parameters, which greatly speeded up the process. Nowadays, the method can be readily programmed and precision estimates incorporated. This method can be applied to any thickness of film, not just at the extrema, although maximum precision is achieved, as we might expect, near optical thicknesses of odd quarter-waves, while, at half-wave optical thicknesses, it is unable to yield any results. As with many other techniques, it suffers from multiple solutions, particularly when the films are thick, and in practice a range of wavelengths is employed, which adds an element of redundancy and helps to eliminate some of the less probable solutions.

Hadley's method involves simple iteration and does not require any very powerful computing facilities. Even in the absence of Hadley's precalculated curves, it can be accommodated on a programmable calculator of modest capacity. It does, however, involve the additional measurement of film thickness, which is of a different character from the measurements of R and T . This is the primary disadvantage. There is a problem with virtually all techniques that make independent measurements of thickness. Unless the thickness is very accurately determined and the model used for the thin film is well chosen, the values of optical constants that are derived may have quite serious errors. The source of the difficulty is that the extrema of the reflectance or transmittance curves are essentially fixed in position by the value of n and d . There is only a very small influence on the part of k . Should the value for d be incorrect then there is no

way in which a correct choice of n can satisfy both the value and the position of the extremum. What happens, then, is that the extremum position is assured by an apparent dispersion, usually enormous and quite false, and the values of n are seriously in error, sometimes showing abrupt gaps in the curve. The situation is often worse at the half-wave points than at the quarter-wave ones but, even in between the extrema, there are clear errors in level which tend to be alternately too high and then too low in between successive extremum pairs. A technique that has been used to avoid this difficulty is to permit some small variation of d around the measured value and to search for a value that removes to the greatest extent the incorrect features of the variation of n .

A different approach that has been developed by Pelletier and his colleagues in Marseille [46] and requires the use of powerful computing facilities, retains the measurement of R and T , but, instead of an independent measure of film thickness, adds the measurement of R' , the reflectance of the film from the substrate side. Now we have three parameters to calculate at each wavelength and three measurements, and it might appear possible that all three could be calculated by a process of iteration, rather like the Hadley method, but the Marseille group found the possible precision rather poor and it broke down completely when there was no absorption. To overcome this difficulty, the Marseille method uses the fact that the geometrical thickness of the film does not vary with wavelength, and therefore, if information over a spectral region is used, there will be sufficient redundancy to permit an accurate estimate of geometrical thickness. Then once the thickness has been determined, a computer method akin to refinement finds accurate values of the optical constants n_f and k_f over the whole wavelength region. For dielectric layers of use in optical coatings, k_f will usually be small, and often negligible, over at least part of the region and a preliminary calculation involving an approximate value of n_f is able to yield a value for geometrical thickness, which in most cases is sufficiently accurate for the subsequent determination of the optical constants. Given the thickness, R and T , as we have seen, should in fact be sufficient to determine n_f and k_f . But this would mean discarding the extra information in R' , and so the determination of the optical constants uses successive approximations to minimise a figure of merit consisting of a weighted sum of the squares of the differences between measured T , R and R' and the calculated values of the same quantities using the assumed values of n_f and k_f . Although seldom necessary, the new values of the optical constants can then be used in an improved estimate of the geometrical thickness, and the optical constants recalculated. For an estimate of precision, the changes in n_f and k_f to change the values of T , R and R' by a prescribed amount, usually 0.3%, are calculated. Invariably, there are regions around the wavelengths for which the film is an integral number of half-waves thick, where the errors are greater than can be accepted and results in these regions are rejected. In practice the films are deposited over half of a substrate, slightly wedged to eliminate the effects of multiple reflections, and measurements are made of R and R' and T and T' on both coated and uncoated portions of the substrate. This permits the optical

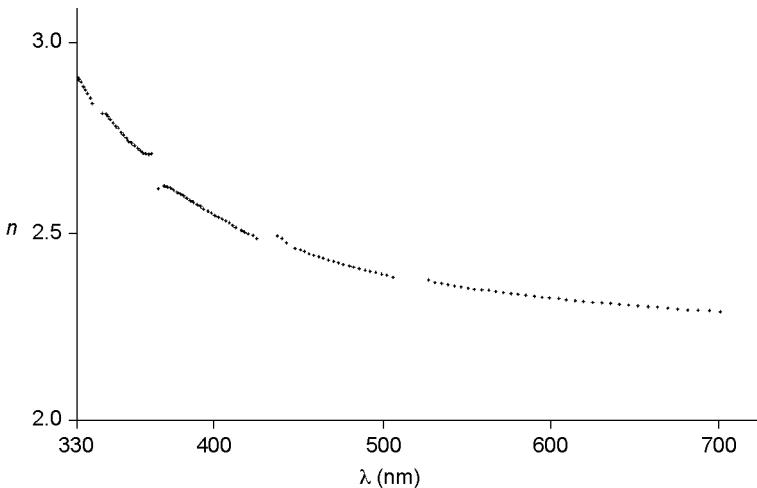


Figure 9.19. The refractive index of a film of zinc sulphide. The slight departure from a smooth curve is due to structural imperfections suggesting that even in this case of a very well-behaved optical material there is some very slight residual inhomogeneity. (After Pelletier *et al* [46].)

constants of the substrate to be estimated; the redundancy in the measurements of T and T' , the transmittance measured in the opposite direction, gives a check on the stability of the apparatus. A very large number of different dielectric thin-film materials have been measured in this way and a typical result is shown in figure 9.19.

A particularly useful and straightforward family of techniques is known as the envelope method. The results that they yield are particularly stable. The envelope method was first described in detail by Manifacier *et al* [47] and was later elaborated by Swanepoel [48]. Provided the absorption in a thin film is small then the transmittance at the quarter- and half-wave points is a fairly simple function of n_f , k_f and d_f . Unfortunately, the transmittances at these points for one single film can only be measured for different wavelengths. The optical constants of the film are functions of wavelength and an iterative process involving interpolation is necessary to extract their values. In their method, therefore, Manifacier *et al* begin by interpolating the actual values of transmittance by drawing two envelope curves around the transmittance characteristic for the film. These envelope curves are then supposed to mark the loci of quarter-wave and half-wave points assuming that the thickness of the film were to vary by a small amount. This gives at each wavelength point two values of transmittance corresponding to the two envelopes and therefore to the transmittances that a film of thickness an integral number of half-waves or of an odd number of quarter-waves would have at that particular wavelength. These transmittances are denoted

by T_{\max} and T_{\min} respectively for a film of high index on a substrate of lower index. For such a film we can write

$$\alpha = \frac{C_1 [1 - (T_{\max}/T_{\min})^{1/2}]}{C_1 [1 + (T_{\max}/T_{\min})^{1/2}]} \quad (9.12)$$

where

$$\begin{aligned} \alpha &= \exp(-4\pi k_f d_f/\lambda) \\ 4\pi k_f d_f/\lambda &= m\pi \text{ (quarter- or half-wave thickness)} \\ C_1 &= (n_f + n_0)(n_m + n_f) \\ C_1 &= (n_f - n_0)(n_m - n_f) \\ T_{\max} &= 16n_0 n_m n_f^2 \alpha / (C_1 + C_2 \alpha)^2 \\ T_{\max} &= 16n_0 n_m n_f^2 \alpha / (C_1 - C_2 \alpha)^2. \end{aligned} \quad (9.13)$$

Then from (9.12) and (9.13), if we define N as

$$N = \frac{n_0^2 + n_m^2}{2} + 2n_0 n_m \frac{T_{\max} - T_{\min}}{T_{\max} T_{\min}} \quad (9.14)$$

n_f is given by

$$n_f = \left[N + \left(N^2 - n_0^2 n_m^2 \right)^{1/2} \right]^{1/2}. \quad (9.15)$$

Once n_f has been determined, equation (9.12) can be used to find a value for α . The thickness d_f can then be found from the wavelengths corresponding to the various extrema and the extinction coefficient k_f from the values of d_f and α . The method has the advantage of explicit expressions for the various quantities, which makes it easily implemented on machines as small as programmable calculators. Unfortunately, as with many of the other techniques, the results can suffer from appreciable errors in the presence of inhomogeneity.

Computers bring the advantage that we no longer need to devise methods of optical constant measurement with the principal objective of ease of calculation. Instead, methods can be chosen simply on the basis of precision of results, regardless of the complexity of the analytical techniques that are required. This is the approach advocated by Hansen [49], who has developed a reflectance attachment making it possible to measure the reflectance of a thin film for virtually any angle of incidence and plane of polarisation, the particular measurements carried out being chosen to suit each individual film.

For rapid, straightforward measurement of refractive index, a method due to Abelès [50] is especially useful. It depends on the fact that the reflectance for p-polarisation is the same for substrate and film at an angle of incidence that depends only on the indices of film and incident medium, and not at all on either substrate index or film thickness, except, of course, that layers that are a half-wave

thick at the appropriate angle of incidence and wavelength will give a reflectance equal to the uncoated substrate regardless of index. It is fairly easy to use Snell's law and the expressions for equal p-admittances to give

$$n_f \sin \vartheta_f = n_0 \sin \vartheta_0$$

and

$$n_f / \cos \vartheta_f = n_0 / \cos \vartheta_0$$

so that

$$\tan \theta_0 = n_f / n_0. \quad (9.16)$$

The measurement of index reduces to the measurement of the angle θ_0 at which the reflectances are equal. Heavens [40] shows that the greatest accuracy of measurement is, once again, obtained when the layer is an odd number of quarter-waves thick at the appropriate angle of incidence. This is because there is then the greatest difference in the reflectances of the coated and uncoated substrate for a given angular misalignment from the ideal. It is possible to achieve an accuracy of around 0.002 in refractive index provided the film and substrate indices are within 0.3 of each other, but not equal. Hacskaylo [51] has developed an improved method based on the Abelès technique. It involves incident light that is plane polarised with the plane of polarisation almost but not quite parallel to the plane of incidence. The reflected light is passed through an analyser and the analyser angle, for which the reflected light from the uncoated substrate and from the film-coated substrate are equal, is plotted against the angle of incidence. A very sharp zero at the angle satisfying the Abelès condition is obtained, which permits accuracies of 0.0002–0.0006 in the measurement of indices in the range 1.2–2.3. It is not necessary for the film index to be close to the substrate index.

Values of R and φ for an opaque surface, for example, define completely and unambiguously the optical constants of the surface. Absolute reflectance is a difficult measurement and it is more usual to measure the way in which the unknown surface compares with a known reference—which introduces further difficulties. Phase is even more involved, requiring an interferometric operation as well as a known standard. Phase measurements are, therefore, quite rare and routine measurements of reflectance are almost always comparative. A major problem is the calibration and maintenance of suitable standards. There is, however, a way of avoiding such difficulties. At normal incidence there is only one value of reflectance and one of phase but at oblique incidence there are two, one pair for s-polarisation and the other for p-polarisation. In principle, therefore, it should be possible to use one as a reference for the other and this leads to the method known as ellipsometry. Two quantities are involved, the ratio of p- and s-reflectances $|\rho_p/\rho_s|$ and $(\varphi_p - \varphi_s)$ the relative phase shift. It is convenient to convert $|\rho_p/\rho_s|$ into an angle so that the parameters become the angles ψ and Δ , where

$$\tan \psi = \frac{|\rho_p|}{|\rho_s|} \quad \text{and} \quad \Delta = (\varphi_p - \varphi_s). \quad (9.17)$$

The implication of ψ and Δ is a change in the state of polarisation of the reflected light with respect to the polarisation of the incident, and so they are directly and simply related to the ellipticity and orientation of the polarisation of the beam. ψ and Δ are therefore known as the ellipsometric parameters and their study is known as ellipsometry.

Ellipsometry [52, 53] possesses several advantages and disadvantages over other measurement techniques. Advantages are the ability to use a single illuminated spot for both measurements and the absence of any reference samples that must be maintained. Although high accuracy is required, the measurement is simple involving straightforward manipulations of polarised light. Disadvantages are that the measurement is at oblique incidence, quite far from more normal measurements of performance, making it difficult to exercise instinct in judging the results. Although the measurement is a ratio, nevertheless the instrument must be carefully calibrated with regard to angle of incidence and alignment of polarisers and analysers. A limitation is that there are two parameters only, rather less than the number that must often be established for a complete description of the system.

A full description of ellipsometry and its techniques is beyond the scope of this book but some observations are appropriate. First of all the ellipsometric convention for phase shift is different from that normally used in optical coatings. The p-polarisation reference direction in the reflected beam is reversed, implying a difference of 180° in the values for p-polarised phase shift on reflection. The reason for this difference is the desirability in ellipsometry of arranging that the reference directions for s- and p-polarisations should coincide with the reference directions used in defining the elliptical polarisation state. It would be very difficult if these were changed in the reflected beam.

Two parameters, refractive index and extinction coefficient, are sufficient to define a simple surface. Since there are two ellipsometric parameters ψ and Δ , then it should be possible to make a determination of the surface parameters from a single ellipsometric measurement. This is indeed the case and there is a direct analytical connection between the two descriptions. Unfortunately, this is not the case with a thin film on a substrate. Even with the simplest film on a substrate that is already characterised, there are at least three parameters, n , k and d , necessary to define the film. The properties of films that are absorbing may depart only slightly from a surface of bulk material. In such cases it is often assumed that the extraction techniques used for a simple surface are applicable. The parameters, n and k , that are extracted in this way are usually referred to as the pseudo-optical constants. They exhibit, usually, the gross features of the real optical constants, although they may not be suitable for thin-film calculations and predictions.

In spectroscopic ellipsometry, the wavelength is varied. Since the film physical thickness is not sensitive to wavelength, this introduces an element of redundancy. It is then sufficient to introduce a small amount of additional information. This frequently takes the form of a prescribed spectral variation of optical constants. Other film parameters may then also be included. If there

is enough known information about the structure and makeup of the films the redundancy in the measurement can become so great that even simple multilayers may be evaluated. Spectroscopic ellipsometry does suffer from problems of insensitivity in certain regions that can be likened to the half-wave problem at normal incidence. The angle of incidence is another adjustable parameter that helps in such situations and it can also add to the redundancy. The combination is known as variable angle spectroscopic ellipsometry, frequently abbreviated to VASE.

We illustrate the extraction process by considering the simple case of a single wavelength, single angle measurement of a surface characterised by refractive index n and extinction coefficient k .

Let the incident medium be of index unity and let $\varepsilon = \tan \psi \exp i\Delta$. Then

$$\varepsilon = \frac{\rho_p}{\rho_s} = \frac{(\eta_{0p} - \eta_p)}{(\eta_{0p} + \eta_p)} \cdot \frac{(\eta_{0s} + \eta_s)}{(\eta_{0s} - \eta_s)} \quad (9.18)$$

where the symbols may be taken as the modified admittances and the sign convention of Δ may be considered corrected to the usual thin-film convention by adding or subtracting 180° . Then,

$$\varepsilon = \frac{(1 - y^2) - (\eta_p - \eta_s)}{(1 - y^2) + (\eta_p - \eta_s)} \quad (9.19)$$

where we have replaced the incident medium admittance by unity. Now

$$\eta_s = \frac{\sqrt{y^2 - \sin^2 \vartheta_0}}{\cos \vartheta_0} \quad (9.20)$$

and

$$\eta_p = \frac{y^2 \cos \vartheta_0}{\sqrt{y^2 - \sin^2 \vartheta_0}} \quad (9.21)$$

so that after some manipulation we can write

$$\gamma = \frac{1 - \varepsilon}{1 + \varepsilon} = \frac{(\eta_p - \eta_s)}{(1 - y^2)} = \frac{\eta_p \left(1 - \frac{\eta_s}{\eta_p}\right)}{(1 - y^2)} = \frac{\eta_p \left(1 - \frac{(y^2 - \sin^2 \vartheta_0)}{y^2 \cos^2 \vartheta_0}\right)}{(1 - y^2)} \quad (9.22)$$

i.e.

$$\gamma = \frac{\eta_p \sin^2 \vartheta_0}{y^2 \cos^2 \vartheta_0} = \frac{\sin^2 \vartheta_0}{\cos \vartheta_0 \sqrt{y^2 - \sin^2 \vartheta_0}}. \quad (9.23)$$

This gives

$$y^2 = \frac{\sin^4 \vartheta_0}{\gamma^2 \cos^2 \vartheta_0} + \sin^2 \vartheta_0. \quad (9.24)$$

There will be two solutions and the fourth quadrant solution will be the correct one. For more complicated systems the extraction process used is essentially a process of refinement of the parameters of a model so that its calculated behaviour matches the measured behaviour. The model will usually include the dispersion of the optical constants. The more information that is available for use in setting up a suitable model of the system the better.

Ellipsometry is especially useful for the derivation of the optical constants of opaque metal films. Provided they have a suitable thickness, high-performance metal films can be characterised by a measurement of a surface plasma resonance, discussed already in chapter 8. This tool involves a rather simpler optical arrangement than the ellipsometer but it is more limited in its application. The film in question is deposited over the base of a prism and the resonance is measured in the normal way. Usually a quite undemanding optical arrangement involving a simple goniometer with laser and collimator and receiver will suffice. The p-polarised resonance has three attributes, the angular position, the resonance width and the resonance depth. There are three attributes of the metal coating, n , k and d . n is primarily associated with the resonance width, k with the position, and d with the depth, so that the extraction process is a simple process of model fitting. There is one small problem associated with two possible solutions. The two solutions involve quite distinct values of d , except when the minimum reflectance is zero when the two solutions coincide. A simple technique for distinguishing the correct set of values is to ensure that the two thickness values are sufficiently far apart for the correct one to be recognised. This, of course, means that the sample should be prepared so that the minimum reflectance is sufficiently far from zero, yet the resonance is sufficiently well developed and is a limitation on the range of thicknesses that can be used. Alternatively, measurements at more than one wavelength may be performed. The correct solutions will be those with similar values of d . The technique has been used, for example, in studies of the influence of small changes in process parameters on the optical constants of metals [27].

Unfortunately, the behaviour of real thin films is often more complicated than we have been assuming. They are frequently inhomogeneous, that is, their refractive index varies throughout their thickness. They tend also to be anisotropic, although little work has been done on this aspect of their behaviour, but the possibility should be borne in mind when considering which methods to use for index determination.

Provided that the variation of index throughout the film is either a smooth increase or a smooth decrease, so that there are no extrema within the thickness of the film, the highest and lowest values being at the film boundaries, then we can use a very simple technique to determine the difference in behaviour at the quarter-wave and half-wave points, which would be obtained with an inhomogeneous film. We assume that the film is absorption-free and that its properties can be calculated by a multiple-beam approach, which considers the amplitude reflection and transmission coefficients at the boundary only. We assume that the index of that part of the film next to the substrate is n_b and that

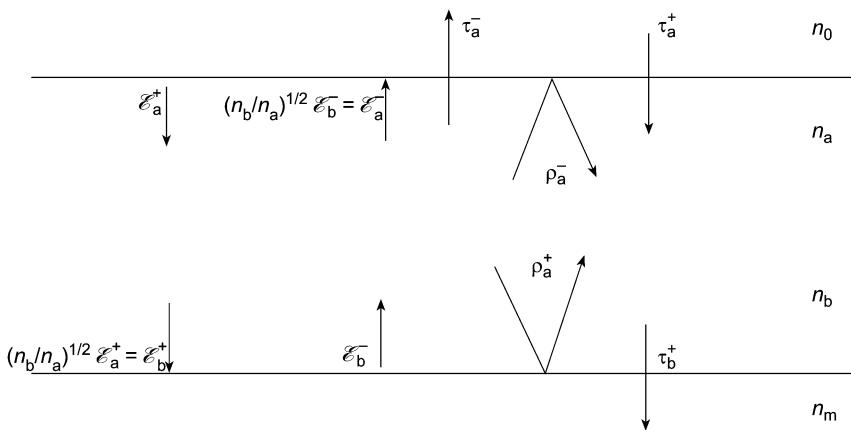


Figure 9.20. Inhomogeneous film quantities used in the development of the matrix expression for an inhomogeneous layer.

next to the surrounding medium is n_a . The corresponding admittances are y_b and y_a . The only reflections that take place are assumed to be at either of the two interfaces. There is one further complication, also indicated in figure 9.20, before we can sum the multiple beams to arrive at transmittance and reflectance. A beam propagating from the outer surface of the film to the inner is assumed to suffer no loss by reflection and, therefore, the irradiance is unaltered. Since irradiance is proportional to the square of the electric amplitude times admittance, a beam that is of amplitude E_a , just inside interface a, will have amplitude $(y_a/y_b)E_b$ at interface b. The correction will be reversed in travelling from b back to a. This is in addition to any phase changes. The inverse correction applies to magnetic amplitudes. Since the correction cancels out for each double pass it does not affect the result for resultant reflectance but it must be taken into account when the multiple beams are being summed for the calculation of transmittance. The derivation of the necessary expressions proceeds as in chapter 2. Here, for simplicity, we restrict ourselves to normal incidence. Oblique incidence is a very simple extension.

$$E_b = E_{1b}^+ + E_{1b}^- \\ H_b = y_b E_{1b}^+ - y_b E_{1b}^-$$

giving

$$E_{1b}^+ = 0.5 [(H_b/y_b) + E_b] \quad H_{1b}^+ = 0.5 [H_b + y_b E_b] \\ E_{1b}^- = 0.5 [-(H_b/y_b) + E_b] \quad H_{1b}^- = 0.5 [H_b - y_b E_b].$$

Then the various rays are transferred to interface a

$$\begin{aligned} E_{1a}^+ &= 0.5 [(H_b/y_b) + E_b] (y_b/y_a)^{1/2} e^{i\delta} \\ E_{1a}^- &= 0.5 [-(H_b/y_b) + E_b] (y_b/y_a)^{1/2} e^{-i\delta} \\ H_{1a}^+ &= 0.5 [H_b + y_b E_b] (y_a/y_b)^{1/2} e^{i\delta} \\ H_{1a}^- &= 0.5 [H_b - y_b E_b] (y_a/y_b)^{1/2} e^{-i\delta} \end{aligned}$$

giving

$$\begin{aligned} E_b &= E_{1b}^+ + E_{1b}^- \\ &= (y_b/y_a)^{1/2} (\cos \delta) E_b + \frac{i \sin \delta}{(y_a y_b)^{1/2}} H_b \\ H_b &= y_b E_{1b}^+ - y_b E_{1b}^- \\ &= i (y_a y_b)^{1/2} (\sin \delta) E_b + (y_a/y_b)^{1/2} (\cos \delta) H_b. \end{aligned}$$

The characteristic matrix for the layer is then given by

$$\begin{bmatrix} (y_b/y_a)^{1/2} \cos \delta & \frac{i \sin \delta}{(y_a y_b)^{1/2}} \\ i (y_a y_b)^{1/2} \sin \delta & (y_a/y_b)^{1/2} \cos \delta \end{bmatrix} \quad (9.25)$$

an expression originally due to Abelès [54]. The calculation of inhomogeneous layer properties has been considered in detail by Jacobsson [55].

Now we consider cases where the layer is either an odd number of quarter-waves or an integral number of half-waves. We apply the expression (9.25) in the normal way and find the well-known relations

$$R = \left(\frac{y_0 - y_a y_b / y_m}{y_0 + y_a y_b / y_m} \right)^2 \quad \text{for a quarter-wave} \quad (9.26)$$

and

$$R = \left(\frac{y_0 - y_a y_{\text{sub}} / y_b}{y_0 + y_a y_{\text{sub}} / y_b} \right)^2 \quad \text{for a half-wave.} \quad (9.27)$$

The expression for a quarter-wave layer is indistinguishable from that of a homogeneous layer of admittance $(y_a y_b)^{1/2}$, and so it is impossible to detect the presence of inhomogeneity from the quarter-wave result. The half-wave expression is quite different. Here the layer is no longer an absentee layer and cannot therefore be represented by an equivalent homogeneous layer. The shifting of the reflectance of the half-wave points from the level of the uncoated substrate in absorption-free layers is a sure sign of inhomogeneity and can be used to measure it.

The Hadley method of deriving the optical constants takes no account of inhomogeneity. Any inhomogeneity, therefore, introduces errors. The Marseille method, however, includes half-wave points and therefore has sufficient

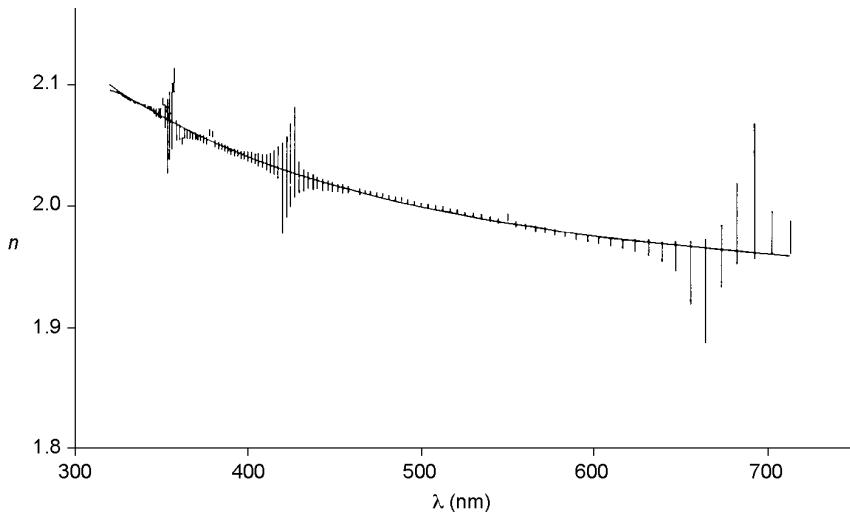


Figure 9.21. Values of mean index and the uncertainty n calculated for hafnium oxide using an inhomogeneous film model. The Cauchy coefficients for n are: $A = 1.9165$, $B = 2.198 \times 10^4 \text{ nm}^2$, $C = -3.276 \times 10^8 \text{ nm}^4$ and for $\Delta n/n$ are: $A' = -5.39 \times 10^{-2}$, $B' = -1.77 \times 10^3 \text{ nm}^2$. (After Borgogno *et al* [54].)

information to accommodate inhomogeneity. The matrix expression is a good approximation when the inhomogeneity is not too large and when the admittances y_a and y_b are significantly different from those of substrate and incident medium. To avoid any difficulties due to the model, the Marseille group actually uses a model for the layer consisting of at least ten homogeneous sublayers with linearly varying values of n but identical values of k and thickness d . The half-wave points still give the principal information on the degree of inhomogeneity. They are also affected by the extinction coefficient k and this has also to be taken into account. One half-wave point within the region of measurement can be used to give a measure of inhomogeneity that is assumed constant over the rest of the region. Several half-wave points can yield values of inhomogeneity that can be fitted to a Cauchy expression, that is an expression of the form

$$\frac{\Delta n}{n} = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4}. \quad (9.28)$$

Details of the technique are given by Borgogno *et al* [56]. Some of their results are shown in figure 9.21.

The envelope method has also been extended to deal with inhomogeneous films using the inhomogeneous matrix expression for the calculations [57]. The extinction coefficient k , as in the Marseille method, is assumed constant through the film.

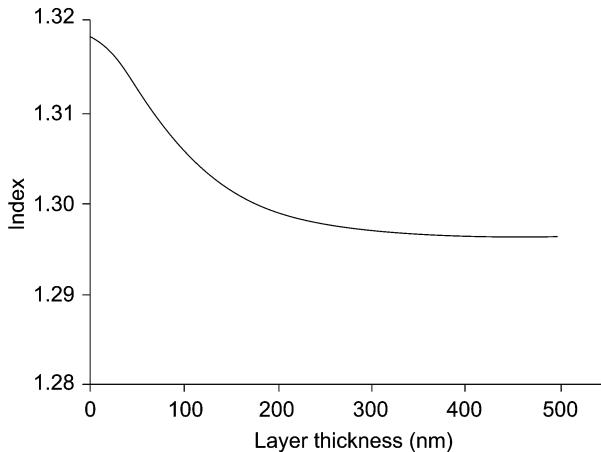


Figure 9.22. Graph of the index profile of cryolite layers at $\lambda = 633$ nm, derived from fitting a formula, $n^2 = A + [B/(t^2 + C)]$, where t is the thickness coordinate, to curves of the variation of reflectance *in vacuo* of a cryolite film deposited over a zinc sulphide film of varying thickness. $A = 1.6773$, $B = 5.0431 \times 10^2$ nm 2 , $C = 8.2986 \times 10^3$ nm 2 . (After Netterfield [58].)

Netterfield [58] measured the variation in reflectance of a film at a single wavelength as it was deposited. If the assumption is made that the part of the film which is already deposited is unaffected by subsequent material, then the values of reflectance associated with extrema can be used to calculate a profile of the refractive index throughout the thickness of the layer. Some results obtained for cryolite, in this way, are shown in figure 9.22.

9.3 Measurement of the mechanical properties

From the point of view of optical coatings, the importance of the mechanical properties of thin films is primarily in their relation to coating stability, that is, the extent to which coatings will continue to behave as they did when removed from the coating chamber, even when subjected to disturbances of an environmental and/or mechanical nature. There are many factors involved in stability, many of which are neither easy to define nor to measure and there are still great difficulties to be overcome. The approach used in quality assurance in manufacture, discussed further in chapter 10, is entirely empirical. Tests are devised which reproduce, in as controlled a fashion as possible, the disturbances to which the coating will be subjected in practice, and samples are simply subjected to these tests and inspected for signs of damage. Sometimes the tests are deliberately made more severe than those expected in use. Coating performance specifications are normally written in terms of such test levels.

Stress is measured by depositing the material on a thin flexible substrate that becomes deformed under the stress applied to it by a deposited film. The deformation is measured and the value of stress necessary to cause it calculated. The substrate may be of any suitable material; glass, mica, silica, metal, for example, have all been used. The form of the substrate is often a thin strip, supported so that part of it can deflect, and either the deflection is measured in some way or a restoring force is applied to restore the strip to its original position. Usually the deflection, or the restoring force, is measured continuously during deposition. Optical microscopes, capacitance gauges, piezoelectric devices and interferometric techniques are some of the successful methods.

A useful survey of the field of stress measurement in thin films in general is given by Hoffman [59]. A particularly useful paper which deals solely with dielectric films for optical coatings is that by Ennos [60]. Ennos used a thin strip of fused silica as substrate, simply supported at each end on ball bearings so that the centre of the strip was free to move. An interferometric technique with a helium-neon laser as the light source was used to measure the movement of the strip. The strip was made of one mirror of a Michelson interferometer of novel design, shown in figure 9.23. Since the laser light was plane polarised, the upper surface of the prism was set at the Brewster angle to eliminate losses by reflection of the emergent beam. Apart from the more obvious advantages of large coherence length and high collimation, the laser beam made it possible to line up the interferometer with the bell-jar of the plant in the raised position (see figure 9.23(b)). No high quality window in the plant was necessary, the glass jar of quite poor optical quality proving adequate. To complete the arrangement, the laser light was also directed on a test flat for the optical monitoring of film thickness. A typical record obtained with the apparatus is also shown in figure 9.23(c). The calibration of the fused-silica strip was determined both by calculation and by measurement of deflection under a known applied load.

Curves plotted for a wide range of materials showing the variation of stress in the films during the actual growth as a function both of film thickness and evaporation conditions are included in the paper, some examples being shown in figure 9.24. It is of particular interest to note the frequent drop in stress when the films are exposed to the atmosphere. This is principally due to adsorption of water vapour, an effect to be considered further towards the end of this chapter.

The interferometric technique has been further improved more recently by Roll and Hoffman [61]. Then Ledger and Bastien [62] have taken the Michelson interferometer of Ennos and replaced it by a cat's-eye interferometer, using circular disks as sensitive elements that are very much less temperature sensitive, and this has enabled the measurement of stress levels in optical films over a wide range of substrate temperatures. Examination of the differences in thermally induced stress for identical films on different substrate materials, when substrate temperature is varied after deposition, has permitted the measurement of the elastic moduli and thermal expansion coefficients of the thin-film materials. Although the measured value of expansion coefficient for bulk thorium fluoride

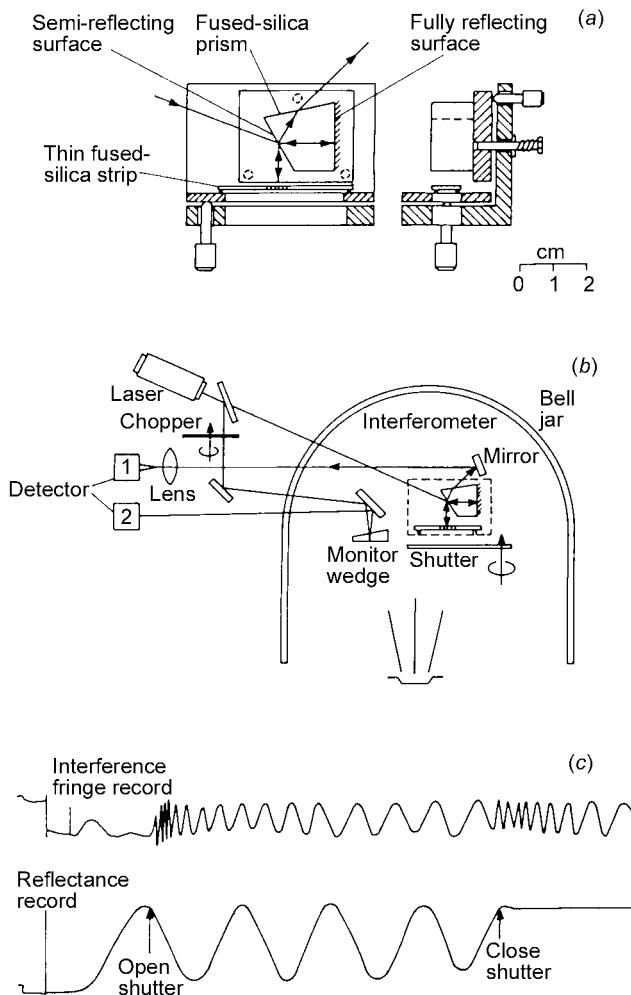


Figure 9.23. (a) Film-stress interferometer. (b) Experimental arrangement for continuous measurement of film stress during evaporation. (c) Recorder trace of fringe displacement and film reflectance. (After Ennos [60].)

crystals is small and negative, the values for thorium fluoride thin films were consistently large and positive, varying from 11.1×10^{-6} to $18.1 \times 10^{-6} \text{ }^{\circ}\text{C}$. Young's modulus for the same samples varies from 3.9×10^5 to $6.8 \times 10^5 \text{ kg cm}^{-2}$ (that is 3.9×10^{10} to $6.8 \times 10^{10} \text{ Pa}$).

Ledger and Bastien arranged the interferometer so that fringes were counted as they were generated at the centre of the interferometer during the deposition of the film and changes in the stress. An asymmetric shape to the fringes permitted

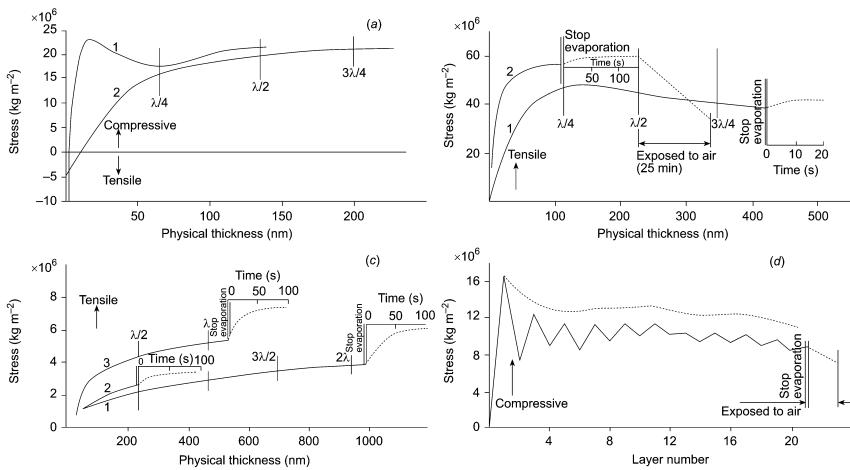


Figure 9.24. (a) Film stress in evaporated zinc sulphide on fused silica at ambient temperature. Evaporation rate $1:0.25 \text{ nm s}^{-1}$, $2:2.2 \text{ nm s}^{-1}$. (b) Film stress in magnesium fluoride. 1: Direct evaporation from molybdenum, evaporation rate 4.2 nm s^{-1} . 2: Indirect radiative heating, evaporation rate 1.2 nm s^{-1} . (c) Cryolite and chiolite evaporated by indirect radiative heating. 1: Cryolite, evaporation rate 3.5 nm s^{-1} . 2: Chiolite, evaporation rate 4 nm s^{-1} . (d) Zinc sulphide–cryolite multilayer. Twenty-one layers $(HL)^{10}H$. Resultant average stress after each evaporation plotted. Dashed curve shows upper limit of film stress reached during the warm-up period before the evaporation of a layer commenced. (After Ennos [60].)

the distinction between a fringe appearing and a fringe disappearing. This meant that the stress level would be lost if the fringe count failed at any stage. A group at the Optical Sciences Center [63] modified the interferometer to view a sufficiently large field that included a number of fringes. The fringe pattern was then interpreted in the manner of an interferogram to give the form of the surface of the deformable substrate. This effectively decoupled each measurement from all the others and permitted the stress to be determined unambiguously at any stage even if some intervening measurements were missed or skipped. The interferometer was used in a detailed study of titanium dioxide films deposited by thermal evaporation with or without ion assist.

Thermally evaporated films usually exhibit a tensile stress that is a consequence of the disorder which is frozen into the film, as freshly arriving material covers what is already existing. An increase in the rate of deposition gives less time for the material on the surface to reorganise itself and therefore should lead to an increase in tensile stress. This is clearly seen in figure 9.25

Under bombardment, the tighter packing of the films leads to an increase in compressive stress because of the transfer of momentum to the growing film; figure 9.26. In fact it is possible by careful control of the bombardment to achieve

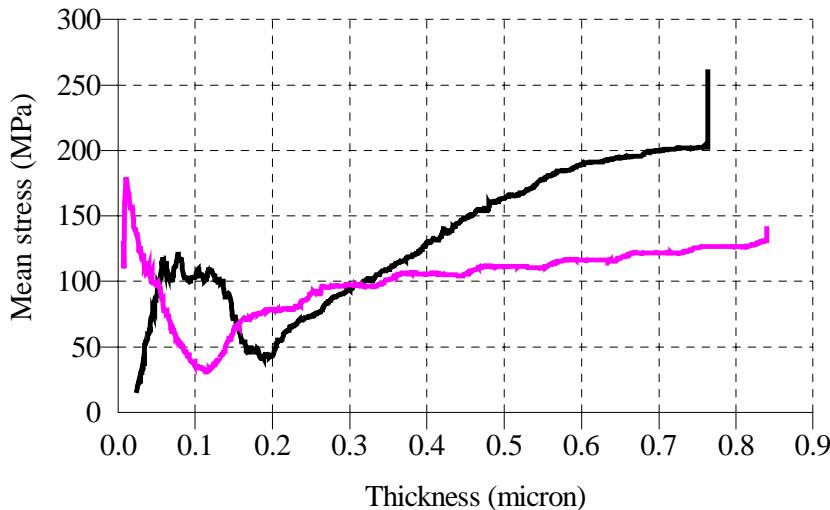


Figure 9.25. The mean (tensile) stress as a function of film thickness in titania films deposited at 0.7 nm s^{-1} (grey) and 0.97 nm s^{-1} (black). The higher rate of deposition leads to greater tensile stress. The vertical line at the end of each curve is a relaxation thought to be due to the disappearance of the thermal gradient present during deposition [63].

extremely low values. Unfortunately, not all materials exhibit such a simple relationship.

Pulker [64] has studied the relationship between stress levels and the microstructure of optical thin films, developing further some ideas of Hoffman. The work is surveyed in reference [25]. Good agreement between measured levels of stress and those calculated from the model has been achieved, but perhaps the most spectacular feature has been the demonstration, in accord with the theory, that small amounts of impurity can have a major effect on stress. The impurities congregate at the boundaries of the columnar grains of the films and reduce the forces of attraction between neighbouring grains, thus reducing stress. Small amounts of calcium fluoride in magnesium fluoride, around 4 mol%, reduce tensile stress by some 50%. Pellicori [65] has shown the beneficial effect of mixtures of fluorides in reducing cracking in low-index films for the infrared.

Windischmann [66] has discussed and modelled the stresses in ion-beam sputtered thin films. He identifies momentum transfer as the important parameter. This is in line with conclusions regarding ion-assisted deposition. The results of figure 9.26 agree with the Windischmann model.

Abrasion resistance is another mechanical property that is of considerable importance and yet extremely difficult to define in any terms other than empirical. This is probably principally because abrasion resistance is not a

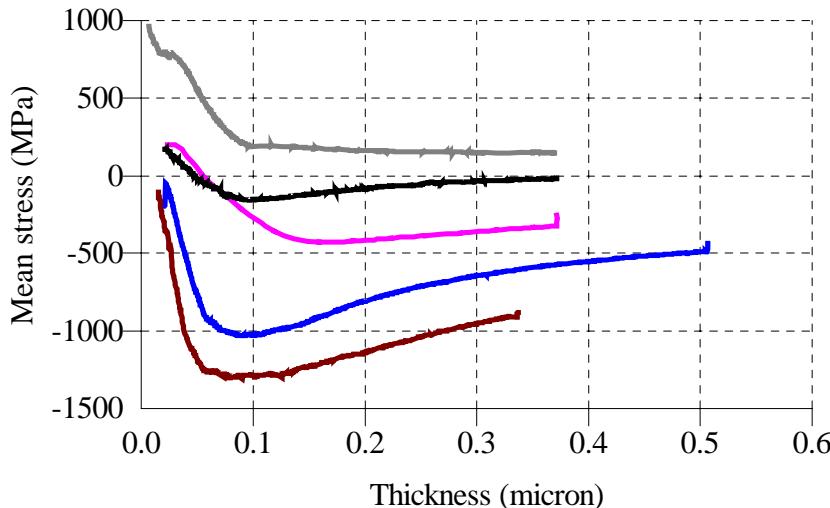


Figure 9.26. The mean stress as a function of thickness of a series of titania films deposited by ion-assisted deposition. The background gas was oxygen and the films were bombarded with 500 eV argon ions at levels from top to bottom of 0.16, 0.32, 0.48, 0.80 and 1.02 mA cm^{-2} [63].

single fundamental property but rather a combination of factors such as adhesion, hardness, friction, packing density and so on. Various ways of specifying abrasion resistance exist but all depend on arbitrary empirical standards. The standard sometimes involves a pad, made from rubber, which may be loaded with a particular grade of emery. The pad is drawn over the surface of the film under a controlled load for a given number of strokes. Signs of visible damage show that the coating has failed the test. Because the pad in early versions of the test was a simple eraser the test is sometimes known as the eraser test. Similar standard tests may be based on the use of cheesecloth or even of steel wool. Wiper blades and sand slurries have also been used to attempt to reproduce the kind of abrasion that results from wiping in the presence of mud. Most of the tests suffer from the fact that they do not give a measure of the degree of abrasion resistance but are merely of a go/no-go nature. There is a modification of the test, described in chapter 10, which does permit a measure of abrasion resistance to be derived from the extent of the damage caused by a controlled amount of abrasion. This is probably the best arrangement yet devised, but even here the results vary considerably with film thickness and coating design so that it is far from an absolute measure of a fundamental thin-film property. The scratch test, described shortly, is sometimes used to derive an alternative measure of abrasion resistance. Abrasion resistance is, therefore, primarily a quality-control tool. It will be considered further in chapter 10.

Adhesion is another important mechanical property that presents difficulties in measurement. What we usually think of as adhesion is the magnitude of the force necessary to detach unit area of the film from the substrate or from a neighbouring film in a multilayer. However, accurate measures of this type are impossible. Quality-control testing is, as for many of the other mechanical properties, of a go/no-go nature. A strip of adhesive tape is stuck to the film and removed. The film fails if it delaminates along with the tape. Jacobsson and Kruse [67] have studied the application of a direct-pull technique to optical thin films. In principle, the adhesive forces between film and substrate can be measured simply by applying a pull to a portion of the film until it breaks away, and, indeed, this is a technique which is used for other types of coatings, such as paint films. The test technique is straightforward and consists of cementing the flat end of a small cylinder to the film, and then pulling the cylinder, together with the portion of film under it, off the substrate, in as near normal a direction as possible. The force required to accomplish this is the measure of the force of adhesion. Great attention to detail is required. The end of the cylinder must be true, must be cemented to the film so that the thickness of cement is constant and so that the axis of the cylinder is vertical. The pull applied to the cylinder must have its line of action along the cylinder axis, normal to the film surface. The precautions to be taken, and the tolerances that must be held, are considered by Jacobsson and Kruse. Their cylindrical blocks were optically polished at the ends, and, in order more nearly to ensure a pull normal to the surface, the film and substrate were cemented between two cylinders, the axes of which were collinear. The mean value of the force of adhesion between 250 nm thick ZnS films and a glass substrate was found to be 2.3×10^7 Pa, which rose to 4.3×10^7 Pa when the glass substrate was subjected to 20 minutes of ion bombardment before coating. Zinc sulphide films evaporated on to a layer of SiO, some 150 nm thick, gave still higher adhesion figures of 5.4×10^7 Pa. The increases in adhesion due to the ion bombardment and the SiO were consistent, and the scatter in successive measures of adhesion was small, some 30% in the worst case.

An alternative method of measuring the force of adhesion is the scratch test, devised by Heavens [68], and improved and studied in detail by Benjamin and Weaver [69, 70], who applied it to a range of metal films. Again, in principle, it is a straightforward test that nevertheless is very complex in interpretation. A round-ended stylus is drawn across the film-coated substrate under a series of increasing loads, and the point at which the film under the stylus is removed from the surface is a measure of the adhesion of the film. Benjamin and Weaver were able to show that plastic deformation of the substrate under the stylus subjected the interface between film and substrate to a shear force, directly related to the load on the stylus by the expression [69]:

$$F = \left[a / \left(r^2 - a^2 \right)^{1/2} \right] - P \quad (9.29)$$

where

$$a = [W/(\pi P)]^{1/2}$$

P is the indentation hardness of the substrate

r is the radius of the stylus point

a is the radius of the circle of contact

W is the load on the stylus

F is the shear force.

The shear force is roughly proportional to the root of the load on the stylus. For the film just to be removed by drawing the stylus across it, the shear force had just to be great enough to break the adhesive bonds. Using this apparatus, Benjamin and Weaver were able to confirm, quantitatively, what had been qualitatively observed before, that the adhesion of aluminium deposited at pressures around 10^{-5} torr (1.3×10^{-5} mb) on glass was initially poor, of values similar to van der Waals forces, but that after some 200 hours it improved to reach values consistent with chemical bonding. Aluminium deposited at higher pressures, around 10^{-3} torr (1.3×10^{-3} mb), gave consistently high bonding immediately after deposition. This is attributed to the formation of an oxide-bonding layer between aluminium and glass, and a series of experiments demonstrated the importance of such oxide layers in other metal films on glass. On alkali halide crystals, the initial bonding at van der Waals levels showed no subsequent improvement with time. More recently the scratch test has been studied by Laugier [71, 72] who has included the effects of friction during the scratching action in the analysis. Zinc sulphide has been shown to exhibit an unusual ageing behaviour in that it occurs in two well-defined stages. After a period of some 18–24 hours after deposition the adhesion increases by as much as a factor of four from an initially low figure. After a period of three days the adhesion then begins to increase further, and after a further seven days reaches a final maximum that can be some 20 times the initial figure. This is attributed to the formation of zinc oxide at the interface between layer and substrate, first free zinc at the interface combining with oxygen that has diffused through the layer from the outer surface and then later zinc that has diffused to the boundary from within the layer.

Commercial instruments that apply these tests are now available and help to standardise the tests as far as is possible.

Unfortunately, none of these adhesion tests is entirely satisfactory. Some of the difficulties are related to consistency of measurement, but the greatest problem is the nature of the adhesion itself. The forces which attach a film to a substrate, or one film to another, are all very large (usually greater than 100 ton in⁻² or some 10^9 Pa) but also of very short range. In fact, they are principally between one atom and the next. The short range of the forces has two major consequences. First, the forces can be blocked by a single atom or molecule of contaminant, and so adhesion is susceptible to even the slightest contamination. A single monomolecular layer of contaminant is sufficient to destroy completely the adhesion between film and substrate. A small fraction of a monomolecular layer is enough to affect it adversely. Second, although the force of adhesion is large,

the work required to detach the coating, the product of the force and its range, can be quite small. Coatings usually fail in adhesion in a progressive manner rather than suddenly and simultaneously over a significant area, and in such peel failures, it is the work, rather than the force, required to detach the coating—the work of adhesion, as it is usually called—that is the important parameter. This work can be considered as the supply of the necessary surface energy associated with the fresh surfaces exposed in the adhesion failure together with any work lost in the plastic deformation of film and/or substrate.

With some metal films, particularly deposited on plastic, there is evidence that an electrostatic double layer gradually forms, which contributes positively to the adhesion. In the tape test, the adhesive forces are comparatively very weak, but their long range allows them to be applied simultaneously over a relatively large area. Thus the film is unlikely to be detached from the substrate unless it is very weakly bonded, and even then it may not be removed unless there is a stress concentrator that can start the delamination process. Sometimes this is provided by scribing a series of small squares into the coating and the tape will tend to lift out complete squares.

In the case of the direct-pull technique, it is exceedingly difficult to avoid a progressive failure rather than a simultaneous rupturing of the bonds over the entire area of the pin. Unevenness in the thickness of the adhesive, or a pull that is not completely central, can cause a progressive failure with consequent reduction in the force measured. Even when the greatest care is taken it is unlikely that the true force of adhesion will be obtained and the test is useful principally as a quality control vehicle. Poor adhesion will tend to give a very much reduced force.

The scratch test suffers from additional problems. Many of the films used in optical coatings shatter when a sufficiently high load is applied before any delamination from the substrate takes place. Such shattering dissipates additional energy and thus film hardness and brittleness enter into the test results. Rarely with dielectric materials does a clean scratch occur. Again the test becomes useful as a comparison between nominally similar coatings rather than an absolute one. Goldstein and DeLong [73] had some success in the assessment of dielectric films using microhardness testers to scratch the films. Most commercial scratch testers include a microscope, and visual examination of the nature of the failures is an important component of the test. Some also include sensitive acoustical detectors to detect the onset of damage. A stylus skidding over a surface is much quieter than one that is ploughing its way through and shattering the material as it goes.

The chemical resistance of the film is also of some significance, particularly in connection with the effects of atmospheric moisture, to be considered later. In this latter respect, the solubility of the bulk material is a useful guide, although it should always be remembered that, in thin-film form, the ratio of surface area to volume can be extremely large and any tendency towards solubility present in the bulk material greatly magnified. As in so many other thin-film phenomena, the magnitude of the effect depends very much on the particular thickness of material,

on the other materials present in the multilayer, on the particular evaporation conditions, as well as the type of test used. However, a broad classification into moisture resistant (materials such as titanium oxide, silicon oxide and zirconium oxide), slightly affected (materials such as zinc sulphide) and badly affected (materials such as sodium fluoride) can be made.

9.4 Toxicity

In thin-film work, as indeed in any other field where much use is made of a variety of chemicals, the possibility that a material may be toxic should always be borne in mind. Fortunately, most of the materials in common use in thin-film work are reasonably innocuous, but there are occasions where distinctly hazardous materials must be used. The thin-film worker would be wise to check this point before using a new material. The technical literature on thin films, being primarily concerned with physical and chemical properties, seldom mentions the toxic nature of the materials. For example, thorium fluoride, oxyfluoride and oxide are materials that are extensively covered in the literature, but for a long time there was little or no mention of the radioactivity of these materials. Recently there has been a growing realisation of the dangers associated with them and they are gradually being phased out. Some of the thallium salts are useful infrared materials, but these are particularly toxic.

Fortunately, manufacturers' literature is becoming a useful source of information on toxicity, and in any cases of doubt, the manufacturer should always be consulted. As long as toxic material is confined to a bottle there is little danger, but as soon as the bottle is opened, material can escape. A major objective, in the use of toxic materials, is to confine them in a well-defined space, in which suitable precautions may be taken. If material is allowed to escape from this space, so that dangerous concentrations can exist outside, then it may be impossible to prevent an accident. It may be necessary to include the whole laboratory in the danger zone and to take special precautions in cleaning up on leaving. Special clothing, extending to respirators, may even be required while in the laboratory. On the other hand, machines may be isolated from the remainder of the production area by special dust-containing cabinets complete with air circulation and filtration units.

Most of the material evaporated in a process ends up as a coating on the inside of the plant and on the jigs and fixtures, where it usually forms a powdery deposit. The greatest danger is in the subsequent cleaning. Some of the solvents and cleaning fluids that can be used in the process give off harmful vapours. A good rule when dealing with potentially hazardous chemicals is to limit the total quantity on the premises to a minimum and especially the amount that is out of safe storage at any time. This puts an upper bound on the magnitude of any major disaster but also, even if no other precautions are taken, minimises any leakage. It is also good from the psychological point of view. It should also be remembered

that many poisons are cumulative in action, and while a slight dose received in the course of a short experiment may not be particularly harmful, the same dose, repeated many times in the course of several years, may do irreparable damage. Thus, the research worker may get away with a particular process that is operated only enough times to prove it, but the production worker will be expected to operate this process day in and day out, possibly for years. The safety standards in the production shop must therefore be of the highest standard and workers should be aware of them without being dismayed by them. It should be remembered, too, that in an emergency the laboratory may be vacated rapidly. It is then important, particularly for any emergency workers, that the hazardous materials should be well contained and their situation known. Good housekeeping is indispensable.

The thin-film worker in industry should make certain that the medical officer of the works is fully aware of the materials currently in use, so that any necessary precautions can be taken before any trouble occurs.

There are, of course, legal requirements. But legal requirements may not represent sufficient prudent precautions. In general, unless positively dangerous materials are involved, the same precautions should be taken as in any chemical laboratory.

9.5 Summary of some properties of common materials

So far, little has been said about the actual properties of the more useful materials employed in thin-film work. The list which follows is far from being exhaustive, but gives the more important properties of some commonly used materials. Often the properties of a particular material appear to vary from plant to plant and sometimes even from operator to operator. This is a symptom of the lack of tight control, which is unfortunately a frequent feature of optical thin-film work, and generally the worker should measure the particular parameters in his own plant and process. Published figures tend to be more of a guide than anything else. This lack of control, of course, is usually understood by the thin-film practitioner and could be altered, but only with the expenditure of much time and money, which always poses the question whether the market for a thin-film product is sufficiently large to justify the outlay.

The material probably used more than any other in thin-film work is magnesium fluoride. This has an index of approximately 1.39 in the visible (see figure 9.18) and is used extensively in lens blooming. In the simplest case this is generally a single layer. Early workers used fluorite but this was found to be rather soft and vulnerable and was subsequently replaced by magnesium fluoride. Magnesium fluoride can be evaporated from a tantalum or molybdenum boat, and the best results are obtained when the substrate is hot at a temperature of some 200–300 °C. When magnesium fluoride is evaporated, trouble can sometimes be experienced through spitting and flying out of material from the boat. This is thought to be caused by thin coatings of magnesium oxide round the grains

of magnesium fluoride in the evaporant. Magnesium oxide has a rather higher melting point than magnesium fluoride and the grains tend to explode once they have reached a certain temperature. It is important, therefore, to use a reasonably pure grade of material, preferably one specifically intended for thin-film deposition.

Magnesium fluoride tends to suffer, as do many of the fluorides, from rather high tensile stress. In single films the total shear force transmitted across the magnesium fluoride interface with either the substrate or underlying materials is not usually dangerously high but in multilayers containing many magnesium fluoride layers, such as high reflectors, the total strain energy and consequent shear loading can become high enough for spontaneous destruction of the coating to occur. Thus magnesium fluoride is not recommended for use in structures containing many layers.

Probably the easiest materials of all to handle are zinc sulphide and cryolite. They have a good refractive index contrast in the visible, the index of zinc sulphide being around 2.35 and that of cryolite around 1.35. Both materials sublime rather than melt, and can be deposited from a tantalum or molybdenum boat or else from a howitzer (described on p 399). Although these materials are not particularly robust, they are so easy to handle that they are very much used, especially in the construction of multilayer filters for the visible and near infrared which can subsequently be protected by a cemented cover slip. The substrates need not be heated for the deposition of the materials when intended for the visible region. Zinc sulphide is also a particularly useful material in the infrared out to about $25\ \mu\text{m}$. In the infrared, however, the substrates must be heated for the best performance. The conditions are given by Cox and Hass [4], who state the best conditions to be on substrates which have been heated to around $150\ ^\circ\text{C}$ and cleaned with an effective glow discharge just prior to the evaporation and certainly not more than five minutes beforehand. Films produced under these conditions will withstand several hours' boiling in 5% salt water, exposure to humid atmospheres and cleaning with detergent and cotton wool.

A trick, which has sometimes been used with zinc sulphide to improve its durability, is bombardment of the growing film with electrons. This can be achieved by positioning a negatively biased hot filament, somewhere near the substrate carrier, in such a way that the filament is shielded from the arriving evaporant, but is in line of sight of the substrates. This process is still not entirely understood, but it has been suggested [74] that an important factor is the modification of the crystal structure of the zinc sulphide layers by electron bombardment. Resistively heated boats produce a mixture of the cubic zinc blende and the hexagonal wurtzite structure, while electron-beam sources produce purely the zinc blende modification. The hexagonal form is a high temperature modification which, it is suspected, will tend to transform into the lower temperature cubic modification, particularly when water vapour is present, a transformation accompanied by a weakening of adhesion, and even delamination. Deliberate electron bombardment of growing zinc sulphide films

from boat sources results in films with entirely cubic structure and with the improved stability expected from that structure.

For more durable films in the visible region, use can be made of a range of refractory oxide layers. More of these are available for the role of high-index layer than low-index.

Cerium dioxide is a high-index material that is not now as commonly used as it once was. It can be evaporated from a tungsten boat (it reacts strongly with molybdenum, producing dense white powdery coatings that completely cover the inside of the system). The procedure to be followed is given by Hass *et al* [75]. Unless the material is one of the types prepared especially for vacuum evaporation, it should first be fired in air at a temperature of around 700–800 °C. If this procedure is not followed the films will have a lower refractive index. Even with these precautions cerium dioxide is an awkward material to handle. It tends to form inhomogeneous layers and the index varies throughout the evaporation cycle as the material in the tungsten boat is used up. It is therefore difficult to achieve a very high performance from cerium dioxide layers, in terms of maximum transmission from a filter or from an antireflection coating, and its chief use tended to be in the production of high-reflectance coatings, for high-power lasers for example, where high reflectance coupled with low loss is the primary requirement and transmission in the pass region is not as important.

Titanium dioxide is nowadays preferred over cerium oxide and is probably one of the most common high-index materials for the visible and near infrared. It has the advantage of the highest index of any of the transparent high-index materials. It is extremely robust but has a rather high melting point of 1925 °C, which makes it very difficult to evaporate directly from a boat source. Tungsten boats are most useful. One of the most successful early methods [34] was the initial evaporation of pure titanium metal which is then subsequently oxidised in air by heating it to temperatures of 400–500 °C. To obtain the highest possible index it is important to evaporate the titanium metal as quickly as possible at as low a pressure as possible so that little oxygen is dissolved in the film. On oxidation in air, indices of around 2.65 can be attained. If the deposit is partially oxidised beforehand, the index is usually rather lower, of the order of 2.25. Other early methods involved the reaction between atmospheric moisture and titanium tetrachloride. Titanium dioxide is formed when atmospheric moisture mixes with the vapour of hot titanium tetrachloride and can be made to condense on the surface of a component that is introduced into the vapour. Best results on glass are obtained when the temperature of the glass is maintained at around 200 °C.

Both of these methods are useful for single layers but are almost impossibly complicated where multilayers are required. More modern alternative methods involve what is known as reactive deposition using either evaporation from electron-beam sources or sputtering.

Reactive evaporation was developed as a useful process in the early 1950s, Auwärter and colleagues in Europe and Brinsmaid in the United States being major contributors [76–78]. The problem with the direct evaporation of titanium

dioxide is that the very high temperatures that are required cause the titanium dioxide to be reduced so that absorption appears in the film. It was found that the reduced titanium oxide can be reoxidised to titanium dioxide during the deposition by ensuring that there is sufficient oxygen present in the atmosphere within the chamber. It appears that the oxidation takes place actually on the surface of the substrate rather than in the vapour stream, and the pressure of the residual atmosphere of oxygen must be arranged to be high enough for the necessary number of oxygen molecules to collide with the substrate surface. If the pressure is too high, then the film becomes porous and soft. There is therefore a range of pressures over which the process works best, usually 5×10^{-5} to 3×10^{-4} mbar. However, it is not possible to give hard and fast figures because they vary from plant to plant and depend on the particular evaporation conditions such as substrate temperature and speed of evaporation. The conditions must therefore be established by trial and error in each process. A suboxide is normally used as starting material. There are two reasons for this. The suboxide usually melts at a lower temperature than the dioxide or the metal and so is useful when a tungsten boat must be used. However, the reduction of the oxide in melting and vaporising has been mentioned. This causes the composition of the vapour to vary unless the evaporation is what is known as congruent, that is the composition of the vapour is the same as the composition of the material in the source. Experimental evidence shows that congruent evaporation is obtained when the composition is near either Ti_2O_3 or Ti_3O_5 [79]. It is usual to use a starting material that has one or other of these compositions. The evaporation should proceed slowly enough to ensure that complete oxidisation takes place. This means that several minutes should be allowed for a thickness corresponding to a quarter-wave in the visible region. Provided the rate of evaporation is kept substantially constant then the refractive index of the film can be as high as 2.45 in the visible region. The titanium dioxide remains transparent throughout the visible, the absorption in the ultraviolet becoming intense at around 350 nm.

Titanium oxide is also used with success in sputtering processes. Sputtering is the process of bombardment of the material to be deposited with high-energy positive ions so that molecules are ejected and deposited on the substrate. Reactive sputtering is the same process except that the gas in the chamber is one which can and does react with the material as it is sputtered. Usually this gas is oxygen and in this case it reacts with the titanium to produce titanium dioxide without requiring any subsequent oxidation. The problems of poisoning of the sputtering cathodes and the various solutions have already been mentioned in connection with reactive sputtering. The rotating cylindrical magnetron and the mid-frequency twin magnetron are two current solutions.

The most complete account of the properties of titanium dioxide, and the way in which they depend on deposition conditions, is that of Pulker *et al* [80]. The behaviour is exceedingly complicated and the results depend on starting material, oxygen pressure, rate of deposition and substrate temperature. The evaporation of Ti_3O_5 as the starting material gave more consistent results than were obtained

with other possible starting materials. With other forms of titanium oxide, the composition varied as the material was used up, tending in each case towards Ti_3O_5 .

Apfel [81] has pointed out the slight conflict between high optical properties and durability. Optical absorption falls as the substrate temperature is reduced and the residual gas pressure is raised. At the same time, the durability of the layers is adversely affected, and a compromise, which depends on the actual application, is usually necessary. Substrate temperatures between 200–300 °C are usually satisfactory, with gas pressures around 10^{-4} torr (1.3×10^{-4} mbar).

The low-index material that is normally used in conjunction with titanium dioxide is silicon dioxide (silica). Indeed there is virtually no choice amongst the oxides. The usual current method for the evaporation of silicon oxide uses an electron-beam source. Chunks of silica or machined plates are used as source material and a slight background pressure of oxygen may sometimes be used. The silicon oxide forms amorphous layers that are dense and resistant. As with most materials, a high substrate temperature during deposition is an advantage.

The high melting temperature of silica makes it difficult to evaporate it directly from heated boats. However, it is possible to use a reactive method [76, 77] that avoids this problem. Silicon monoxide is a convenient starting material, which, in its own right, is a useful material for the infrared. The silicon monoxide can be evaporated readily from a tantalum boat or, as the material sublimes rather than melts, a howitzer source. Provided there is sufficient oxygen present, the silicon monoxide will oxidise to the form Si_2O_3 that has a refractive index of 1.52–1.55 and exhibits excellent transmission from just on the longwave side of 300 nm out to 8 μm [82].

An interesting effect involving the ultraviolet irradiation of films of Si_2O_3 has been reported [83]. With ultraviolet intensity corresponding to a 435 W quartz-envelope Hanovia lamp at a distance of 20 cm, the refractive index of the film, after around five hours' exposure, drops to 1.48 (at 540 nm). This change in refractive index appears to be due to an alteration in the structure of the film, rather than in the composition, that remains Si_2O_3 . At the same time as the reduction in refractive index, an improvement in the ultraviolet transmission is observed, the films becoming transparent to beyond 200 nm. Longer exposure to ultraviolet, around 150 hours, does eventually alter the composition of the films to SiO_2 . These changes appear to be permanent. Si_2O_3 is a particularly useful material for protecting aluminium mirrors, and this method of improvement by ultraviolet irradiation opens the way to greatly improved mirrors for the quartz ultraviolet. The effect has been studied in some detail by Mickelsen [84] who proposes an explanation involving electron traps.

Heitmann [85] made considerable improvements to the reactive process by ionising the oxygen in a small discharge tube through which the gas is admitted to the coating chamber. The degree of ionisation is not high, but the reactivity of the oxygen is improved enormously, and the titanium oxide and silicon oxide films produced in this way have appreciably less absorption than those deposited by the

conventional reactive process. The silicon oxide films show infrared absorption bands characteristic of the SiO form rather than the more usual Si_2O_3 . The technique has been further improved by Ebert [86] and his colleagues who have developed a more efficient hollow-cathode ion source, and extended the method to materials such as beryllium oxide, with useful transmittance in the ultraviolet.

Other materials found useful in thin films are the oxides and fluorides of a number of the lanthanides or rare earths. Ceric oxide [75], although possibly strictly not a rare earth, has already been mentioned. Cerium fluoride forms very stable films of index 1.63 at 550 nm when evaporated from a tungsten boat.

Similarly, the oxides of lanthanum, praseodymium and yttrium, and their fluorides, form excellent layers when evaporated from tungsten boats. Their properties are summarised in chapter 15. A full account of their properties is given by Hass *et al* [44]. The properties of the rare earth oxides have been shown [87] to have improved transparency, especially in the ultraviolet, when electron-beam evaporation is used.

A detailed study of the fluorides of the lanthanides and their usefulness in the extreme ultraviolet, in fact there is little else that can be used in that region, has been performed by Lingg [88, 89].

Then there is a number of other hard oxide materials which were extremely difficult to evaporate until the advent of the high-power electron-beam gun, and so were used only relatively infrequently, if at all. Zirconium dioxide [87, 90] is a very tough, hard material which has good transparency from around 350 nm to some 10 μm . It tends to give inhomogeneous layers, the degree of inhomogeneity depending principally on the substrate temperature. Hafnium oxide [87, 91] has good transparency to around 235 nm, and an index around 2.0 at 300 nm, so that it is a good high-index material for that region. Both yttrium and hafnium oxide have been found to be good protecting layers for aluminium in the 8–12 μm region [92, 93], which avoid the drop in reflectance at high angles of incidence associated with SiO_2 and with Al_2O_3 .

In the infrared many more possibilities are available. Semiconductors all exhibit a sudden transition from opacity to transparency at a certain wavelength known as the intrinsic edge. This wavelength corresponds to the energy gap between the filled valence band of electrons and the empty conduction band. At wavelengths shorter than this gap, photons are absorbed in the material because they are able to transfer their energy to the electrons in the filled valence band by lifting them into the empty conduction band. At wavelengths longer than this value, the photon energy is not sufficient, and apart from a little free carrier absorption, there is no mechanism for absorbing the energy and the material appears transparent until the lattice vibration bands at rather long wavelengths are encountered. For the more common semiconductors, silicon and germanium, the intrinsic edge wavelengths are 1.1 μm and 1.65 μm respectively. Thus both of these materials are potentially useful in the infrared. A great advantage that they possess is their high refractive index, 3.5 for silicon and 4.0 for germanium.

Silicon, however, is not at all easy to evaporate because it reacts strongly

with any crucible material, and almost the only way of dealing with it in thermal evaporation is to use an electron gun with a water-cooled crucible so that the cold silicon in contact with the crucible walls acts as its own container. The high thermal conductivity of silicon makes it necessary to use high power. Sputtering is a viable process and, in fact, most large-area silicon dioxide coatings are produced by the reactive sputtering of silicon from magnetron targets. The poisoning problem in reactive sputtering and its solutions have already been mentioned. Germanium, on the other hand, is a most useful material and straightforward techniques have been devised to handle it. Tungsten boats can be used provided that the total thickness of material to be deposited is not too great, 2 or 3 μm say, because germanium does react with tungsten. Molybdenum boats have been used with greater success [91]. A quite satisfactory method is to use a crucible made from graphite and heated directly or indirectly when the germanium films obtained are extremely pure and free from absorption. Again, the method of choice nowadays is the electron-beam source when the hearth material can be graphite or water-cooled copper.

There are other semiconductors of use as follows. Tellurium [95, 96] has an index of 5.1 at 5 μm , good transmission from 3.5 μm to at least 12 μm , and can be evaporated easily from a tantalum boat. Lead telluride [5, 97–104] has an even higher index of around 5.5 with good transmission from 3.4 μm out to beyond 20 μm . A tantalum boat is the most suitable source. Care must be taken not to overheat the material; the temperature should be just enough to cause the evaporation to proceed, otherwise some alteration in the composition of the film will take place, causing an increase in free-carrier absorption and consequent fall-off in longwave transparency. The substrates should be heated, best results being obtained with temperatures around 250 °C, but as this will be too great for the low-index film which is usually zinc sulphide, a compromise temperature which is rather lower, usually around 150 °C, is often used for both materials. One difficulty with lead telluride is the ease with which it can be upset by impurities that cause free-carrier absorption. It is extremely important to use pure grades of material and this applies to the accompanying zinc sulphide as well as the lead telluride, especially if the material is to be used at the longwave end of its transparent region. Lead telluride also appears to be incompatible with a number of other materials, particularly some of the halides, presumably because material diffuses into the lead telluride generating free carriers. An annealing process which can in certain circumstances improve the transmission of otherwise absorbing films of lead telluride in the region beyond 12 μm is described by Evans and Seeley [99].

Lead telluride can in some circumstances behave in a curious way immediately after deposition [101, 102]. The optical thickness of the material is observed to grow during a period of around 15 minutes while the layer is still under vacuum. Typical gains in optical thickness of a half-wave layer are of the order of 0.007 full waves, although in any particular case it varies considerably and can often be zero. The reasons for this behaviour are not clear but the layers

do not exhibit any further instability, once they have ceased growing. It is simply a matter of allowing for this behaviour in the monitoring process.

A wide range of low-index materials is used in the infrared. Zinc sulphide [4, 45] in comparison with the high-index semiconductors has a relatively low index. If an electron-beam source is not available, then zinc sulphide should be deposited from a tantalum boat, or, better still, a howitzer, on substrates freshly cleaned by a glow discharge and held at temperatures of around 150 °C, if the maximum durability is to be obtained. Zinc sulphide films so treated will withstand boiling for several hours in 5% salt solution, cleaning with cotton wool, and exposure to moist air, without damage [4]. Silicon monoxide is another possibility [4, 105]. It can also be deposited from a tantalum boat or a howitzer. The deposition rate should be fast and the pressure low, of the order of 10^{-5} torr (1.3×10^{-5} mb) or less if possible. The refractive index is around 1.85 at 1 μm and falls to 1.6 at 7 μm . A strong absorption band prevents use of the material beyond 8 μm . Thorium fluoride, unfortunately radioactive, has been much used in the past, although it is less in favour nowadays because of its radioactivity, and there are many other materials, such as fluorides of lead, lanthanum, barium, cerium, for example, and oxides such as titanium, yttrium, hafnium and cerium. Some details of these and other materials are given in chapter 15.

The nitrides of silicon and aluminium are tough, hard materials with excellent transparency from the ultraviolet through to around 10 μm in the infrared. They have not been much used in optical coatings because of the difficulty of thermal evaporation. The process of reactive evaporation of the metal in nitrogen does not work because the nitrogen, unless it is in atomic form, does not readily combine with the metal. Evaporation of aluminium, for example, in a residual atmosphere of nitrogen results in bright aluminium films whereas evaporation in oxygen gives aluminium oxide. The situation has changed completely with the introduction of the energetic processes, and especially ion-assisted deposition, into batch optical coatings. The nitrogen beam from the ion source used in these processes reacts strongly with the metal to form dense, hard and tough nitride films of good transparency. There is another enormous advantage in these materials. The oxynitrides represent a continuous range of compositions between the pure oxide and the pure nitride. The oxide is of rather lower refractive index and the refractive index of the oxynitride ranges smoothly with composition from that of the oxide to that of the nitride. The composition of the film is a function of the reacting gas composition and this can readily be varied to alter the film index in a well-controlled manner. Hwangbo and colleagues [28] investigated the ion-assisted deposition of aluminium oxynitride. They used aluminium metal as source material. A particularly straightforward way of controlling the index of aluminium oxynitride films from 1.65 to 1.83 at 550 nm was to bombard the growing film with a constant flux of nitrogen from the ion gun and to supply a variable quantity of oxygen to the process simply as a background gas. The reactivity of the oxygen is so great that any small quantity is taken up preferentially by the film. In fact in the oxynitride process it is virtually

impossible to eliminate oxygen entirely and so the achievable high index does not quite reach the value that would be associated with the pure nitride. Hwangbo was able to construct simple rugate filters with the sole variable during the process being the background pressure of oxygen, all other quantities, bombardment, evaporation rate, and so on, being held constant. Placido [106] has constructed rugate structures of very many accurately controlled cycles from aluminium oxynitride using reactive RF sputtering of aluminium metal in a mixture of oxygen and nitrogen.

Bovard and colleagues [107] produced silicon nitride films using low-voltage ion plating. Here there was no oxygen in the chamber and the films were pure nitride giving a refractive index of 2.05 at 550 nm. The range of variation in index from silicon oxynitride films is potentially very great.

Mixtures of materials are receiving attention both in deliberately inhomogeneous films and in homogeneous films where an intermediate index between the two components of the mixture is required to improve the evaporation properties of an otherwise difficult material.

Jacobsson and Martensson [108] used mixtures of cerium oxide and magnesium fluoride, of zinc sulphide and cryolite, and of germanium and magnesium fluoride, with the relative concentration of the two components varying smoothly throughout the films to produce inhomogeneous films with a refractive index variation of a prescribed law. Some of the results they obtained for antireflection coatings were mentioned in chapter 3. To produce the mixture, two separate sources, one for each material, were used; they were evaporated simultaneously but with independent rate controls. Apparently no difficulty in obtaining reasonable films was experienced, the mixing taking place without causing absorption to appear.

Fujiwara [109, 110] was interested in the production of homogeneous films for antireflection coatings [111]. The three-layer quarter–half–quarter coating for glass requires a film of intermediate index which is rather difficult to obtain with a simple material, and the solution adopted by Fujiwara was to use a mixture of two materials, one having a refractive index lower than the required value and the other higher. The two combinations that were tried successfully were cerium oxide and cerium fluoride, and zinc sulphide and cerium fluoride. These were simply mixed together in powder form in a certain known proportion by weight and then evaporated from a single source. The mixture evaporated giving an index that was sufficiently reproducible for antireflection coating purposes. The range of indices obtainable with the cerium oxide–cerium fluoride mixture was 1.60–2.13, and with the cerium fluoride–zinc sulphide mixture 1.58–2.40. One interesting feature of the second mixture was that, although zinc sulphide on its own is not particularly robust, in the form of a mixture with more than 20% by weight of cerium fluoride the robustness was greatly increased, the films withstanding boiling in distilled water for 15 minutes without any deterioration. Curves are given for refractive index against mixing ratio in the papers.

Mixtures of zinc sulphide and magnesium fluoride have also been studied

by Yadava *et al* [112]. The refractive index of the mixture varies between the indices of magnesium fluoride and zinc sulphide, depending on the mixing ratio, and the absorption edge varies from that of zinc sulphide to that of magnesium fluoride in a nonlinear fashion. The same authors [112, 113] have studied the use of assemblies of large numbers of alternate very thin discrete layers of the components instead of mixtures. For a wide range of material combinations, ZnS–MgF₂, ZnS–MgF₂–SiO, Ge–ZnS, ZnS–Na₃AlFs for example, the results were similar to those expected from the evaporation of mixtures of the same materials.

Silica is a particularly difficult material to evaporate because of its high melting point and also because of its transparency to infrared, which makes it difficult to heat. It was found by workers at the Libbey-Owens-Ford Glass Company [114] that silica could be thermally evaporated readily if some pretreatment was carried out. This consisted of combining the silica with a metallic oxide, a vast number of different oxides being suitable. The oxide can be mixed intimately with the silica, coated on the outer surface of silica chunks or, in some cases where the oxide has a rather lower melting temperature than the silica, mixed very crudely. Only a small quantity of the oxide is required and the evaporation is carried out in the conventional manner from a tungsten source. The oxides mentioned include aluminium, titanium, iron, manganese, cobalt, copper, cerium and zinc. Working along similar lines it has been discovered by workers at Balzers AG [115, 116] that cerium oxide mixed with other oxides improves the oxidation and increases the transparency and ease of evaporation. Materials such as titanium dioxide are difficult to evaporate without absorption, and the most successful method is reactive evaporation in oxygen, which produces absorption-free films, although the process is rather time consuming because the evaporation must proceed slowly. With the addition of a small amount of cerium oxide—the mixture can vary from 1:1 to 8:1 titanium oxide (the monoxide, the dioxide or even the pure metal) to cerium oxide—hard films free from absorption, even when evaporated quickly at pressures of 10⁻⁵ torr, are readily obtained. Apparently this effect is not limited to titanium oxide, and a vast range of different materials which have been successfully tried is given. Other rare earth oxides and mixtures of rare earth oxides can also take the place of the cerium dioxide.

Stetter and his colleagues [90] have pointed out the advantage of oxygen-depleted materials as source material for electron-beam evaporation, in that composition changes little if at all during evaporation, which leads to more consistent film properties. The extra oxygen is supplied, in the usual way, from the residual atmosphere in the plant. The depleted materials also have higher thermal and electrical conductivity. A mixture of ZrO₂ and ZrTiO₄, sintered at high temperature under high vacuum and oxygen-depleted, was developed. This material, designated 'Substance no 1', when evaporated from an electron-beam system in a residual oxygen pressure of 1–2 × 10⁻⁴ torr (1.3–2.5 × 10⁻⁴ mb) with substrate temperature 270 °C, and condensation rate of the order of 10 nm min⁻¹, gives homogeneous layers of refractive index 2.15 (at 500 nm). Such a value of index is ideal for the quarter–half–quarter antireflection coating for the visible

region. This has prompted further work on mixtures [117] and there are now several similar materials available. H1 is from the zirconia/titania system with index 2.1 at 500 nm and good transparency from 360 nm to 7 μm but with some difficulties in evaporation because of incomplete melting. H2 from the praseodymium/titanium oxide system has a similar index and the advantage of ease of evaporation but suffers from a more restricted range of good transmittance, 400 nm to 7 μm , and localised slight absorption in the transparent region. H4 is a lanthanum/titanium oxide combination with again refractive index 2.1 at 500 nm and transmission region from 360 nm to 7 μm that melts completely and so is normally preferred over the other two materials. M1 is a mixture of praseodymium/aluminium oxide with index on heated substrates of 1.71 at 500 nm and good transparency from 300 nm to longer wavelengths.

Butterfield [118] has produced films of a mixture of germanium and selenium. For composition varying from 30 to 50 atomic % of germanium, glassy films with refractive index in the range 2.4–3.1, with good transparency from 1.5–15 μm , could be produced. The starting material was an alloy of germanium and selenium in the correct proportions, produced by melting the pure substances in an evacuated quartz tube. The evaporation source was a graphite boat.

It is likely that much more work will be carried out on mixtures, because of the apparent ease with which the deposition can be performed to give a wide range of refractive indices, many of which are not available by other means. The theory of the optical properties of mixtures is covered in a useful review by Jacobsson [53], who also gives further information on mixtures, and on inhomogeneous layers.

References

- [1] Vossen J L and Kern W 1978 *Thin Film Processes* (New York: Academic)
- [2] Vossen J L and Kern W 1991 *Thin Film Processes II* (San Diego: Academic)
- [3] Glocker D A and Shah S I 1995 *Handbook of Thin Film Process Technology* (Bristol: Institute of Physics)
- [4] Cox J T and Hass G 1958 Antireflection coatings for germanium and silicon in the infrared *J. Opt. Soc. Am.* **48** 677–80
- [5] Ritchie F S 1970 Multilayer filters for the infrared region 10–100 microns *PhD Thesis* (University of Reading)
- [6] Hass G and Ritter E 1967 Optical film materials and their applications *J. Vacuum Sci. Technol.* **4** 71–9
- [7] Coulter J K, Hass G and Ramsay J B 1973 Optical constants and reflectance and transmittance of evaporated rhodium films in the visible *J. Opt. Soc. Am.* **63** 1149–53
- [8] Scobey M A, Seddon R I, Seeser J W, Austin R R, LeFebvre P M and Manley B W Optical Coating Laboratory, Inc. 1989 *Magnetron Sputtering Apparatus and Process* USA Patent 4 851 095
- [9] Scobey M A Optical Corporation of America 1996 *Low Pressure Reactive Magnetron Sputtering Apparatus and Method* USA Patent 5 525 199

- [10] Placido F 1998 *Radio Frequency Sputtering of Optical Coatings Including Rugate Filters* Private communication (Department of Physics, University of Paisley)
- [11] Wei D T, Kaufman H R and Lee C-C 1995 Ion beam sputtering *Thin Films for Optical Systems* ed F R Flory (New York: Marcel Dekker) pp 133–201
- [12] Lalezari R, Rempe G, Thompson R J and Kimble H J 1992 Measurement of ultralow losses in dielectric mirrors *Topical Meeting on Optical Interference Coatings (Tucson, AZ)* (Washington: Optical Society of America) pp 331–3
- [13] Mackowski J M, Pinard L, Dognin L, Ganau P, Lagrange B, Michel C and Morgue M 1998 Different approaches to improve the wavefront of low-loss mirrors used in the VIRGO gravitational wave antenna *Optical Interference Coatings* (Washington: Optical Society of America) pp 18–20
- [14] Pulker H K, Bühler M and Hora R 1986 Optical films deposited by a reactive ion plating process *Proc. Soc. Photo-Opt. Instrumentation Eng.* **678** 110–14
- [15] Pulker H K and Guenther K H 1995 Reactive physical vapor deposition processes *Thin Films for Optical Systems* ed F R Flory (New York: Marcel Dekker) pp 91–115
- [16] Bovard B G 1995 Ion-assisted deposition *Thin Films for Optical Systems* ed F R Flory (New York: Marcel Dekker) pp 117–32
- [17] Fulton M L 1994 Applications of ion-assisted deposition using a gridless end-Hall ion source for volume manufacturing of thin-film optical filters. *Proc. Soc. Photo-Opt. Instrumentation Eng.* **2253** 374–93
- [18] Matl K, Klug W and Zöller A 1991 Ion-assisted deposition with a new plasma source *Mater. Sci. Eng.* **A140** 523–7
- [19] Pongratz S and Zöller A 1992 Plasma ion-assisted evaporative deposition of surface layers *Annual Rev. Mater. Sci.* **22** 279–95
- [20] Zöller A, Beißwenger S, Götzelmann R and Matl K 1994 Plasma ion assisted deposition: a novel technique for the production of optical coatings *Proc. Soc. Photo-Opt. Instrumentation Eng.* **2253** 394–402
- [21] Müller K-H 1986 Monte Carlo calculation for structural modifications in ion-assisted thin film deposition due to thermal spikes *J. Vacuum Sci. Technol.* **4** 184–8
- [22] Müller K-H 1988 Models for microstructure evolution during optical thin film growth *Proc. Soc. Photo-Opt. Instrumentation Eng.* **821** 36–44
- [23] Targove J D, Lingg L J and Macleod H A 1988 Verification of momentum transfer as the dominant densifying mechanism in ion-assisted deposition *Optical Interference Coatings (Tucson, AZ)* (Washington: Optical Society of America) pp 268–71
- [24] Martin P J, Macleod H A, Netterfield R P, Pacey C G and Sainty W G 1983 Ion-beam-assisted deposition of thin films *Appl. Opt.* **22** 178–84
- [25] Messerly M J 1987 Ion-beam analysis of optical coatings *PhD Dissertation* (University of Arizona)
- [26] Sainty W G, Netterfield R P and Martin P J 1984 Protective dielectric coatings produced by ion-assisted deposition *Appl. Opt.* **23** 1116–19
- [27] Hwangbo C K, Lingg L J, Lehan J P, Macleod H A, Makous J L and Kim S Y 1989 Ion-assisted deposition of thermally evaporated Ag and Al films *Appl. Opt.* **28** 2769–78
- [28] Hwangbo C K, Lingg L J, Lehan J P, Macleod H A and Suits F 1989 Reactive ion-assisted deposition of aluminum oxynitride thin films *Appl. Opt.* **28** 2779–84

- [29] Segner J 1995 Plasma impulse chemical vapor deposition *Thin Films for Optical Systems* ed F R Flory (New York: Marcel Dekker) pp 203–29
- [30] Möhl W, Lange U and Pacquet V 1994 Optical coatings on plastic lenses by PICVD-technique *Proc. Soc. Photo-Opt. Instrumentation Eng.* **2253** 486–91
- [31] Hora R and Wohlrab C 1993 Plasma polymerization: a new technology for functional coatings on plastics *36th Annual Technical Conference* (Albuquerque, NM: Society of Vacuum Coaters) pp 51–5
- [32] Wohlrab C and Hofer M 1995 Plasma polymerization of optical coatings on organic substrates: equipment and processes *38th Annual Technical Conference* (Albuquerque, NM: Society of Vacuum Coaters) pp 222–30
- [33] Thomas I M 1993 Sol-gel coatings for high power laser optics: past present and future *Proc. Soc. Photo-Opt. Instrumentation Eng.* **2114** 232–43
- [34] Hass G 1952 Preparation, properties and optical applications of thin films of titanium dioxide *Vacuum* **2** 331–45
- [35] Meaburn J 1967 A search for nebulosity in the high galactic latitude radion spurs *Z. Astrophys.* **65** 93–104
- [36] Title A M, Pope T P and Andelin J P 1974 Drift in interference filters. Part 1 *Appl. Opt.* **13** 2675–9
- [37] Richmond D 1976 Thin film narrow band optical filters *PhD Thesis* (Newcastle upon Tyne Polytechnic)
- [38] Lee C C 1983 Moisture adsorption and optical instability in thin film coatings *PhD Dissertation* (University of Arizona)
- [39] Müller K-H 1985 A computer model for postdeposition annealing of porous thin films *J. Vacuum Sci. Technol.* **3** 2089–92
- [40] Heavens O S 1964 Measurement of optical constants of thin films *Physics of Thin Films* ed G Hass and R E Thun (New York: Academic) pp 193–238
- [41] Liddell H M 1981 *Computer-Aided Techniques for the Design of Multilayer Filters* (Bristol: Adam Hilger)
- [42] Borgogno J-P 1995 Spectrophotometric methods for refractive index determination *Thin Films for Optical Systems* ed F R Flory (New York: Marcel Dekker) pp 269–328
- [43] Hall J F Jr and Ferguson W F C 1955 Dispersion of zinc sulfide and magnesium fluoride films in the visible spectrum *J. Opt. Soc. Am.* **45** 74–5
- [44] Hass G, Ramsay J B and Thun R 1959 Optical properties of various evaporated rare earth oxides and fluorides *J. Opt. Soc. Am.* **49** 116–20
- [45] Hall J F and Ferguson W F C 1955 Optical properties of cadmium sulphide and zinc sulphide from 0.6 micron to 14 micron *J. Opt. Soc. Am.* **45** 714–18
- [46] Pelletier E, Roche P and Vidal B 1976 Détermination automatique des constantes optiques et de l'épaisseur de couches minces: application aux couches diélectriques *Nouv. Rev. Opt.* **7** 353–62
- [47] Manifacier J C, Gasiot J and Fillard J P 1976 A simple method for the determination of the optical constants n , k and the thickness of a weakly absorbing thin film *J. Phys. E* **9** 1002–4
- [48] Swanepoel R 1983 Determination of the thickness and optical constants of amorphous silicon *J. Phys. E* **16** 1214–22
- [49] Hansen W 1973 Optical characterization of thin films: theory *J. Opt. Soc. Am.* **63** 793–802
- [50] Abelès F 1950 La détermination de l'indice et de l'épaisseur des couches minces

- transparentes *J. Phys. Rad.* **11** 310–14
- [51] Hacskaylo M 1964 Determination of the refractive index of thin dielectric films *J. Opt. Soc. Am.* **54** 198–203
- [52] Rivory J 1995 Ellipsometric measurements *Thin Films for Optical Systems* ed F R Flory (New York: Marcel Dekker) pp 299–328
- [53] Azzam R M A 1995 Ellipsometry *Handbook of Optics* ed M Bass (New York: McGraw Hill) pp 27.1–27
- [54] Abelès F 1950 Recherches sur la propagation des ondes électromagnétiques sinusoïdales dans les milieux stratifiés. Applications aux couches minces *Ann. Phys.* **5** 596–640
- [55] Jacobsson R 1975 Inhomogeneous and coevaporated homogeneous films for optical applications *Phys. Thin Films* **8** 51–98
- [56] Borgogno J P, Lazarides B and Pelletier E 1982 Automatic determination of the optical constants of inhomogeneous thin films *Appl. Opt.* **21** 4020–9
- [57] Arndt D P, Azzam R M A, Bennett J M, Borgogno J P, Carniglia C K, Case W E, Dobrowolski J A, Arndt D P, Gibson U J, Hart T T et al 1984 Multiple determination of the optical constants of thin-film coating materials *Appl. Opt.* **23** 3571–96
- [58] Netterfield R P 1976 Refractive indices of zinc sulphide and cryolite in multilayer stacks *Appl. Opt.* **15** 1969–73
- [59] Hoffman R W 1976 Stresses in thin films: the relevance of grain boundaries and impurities *Thin Solid Films* **34** 185–90
- [60] Ennos A E 1966 Stresses developed in optical film coatings *Appl. Opt.* **5** 51–61
- [61] Roll K 1976 Analysis of stress and strain distribution in thin films and substrates *J. Appl. Phys.* **47** 3224–9
- [62] Ledger A M and Bastien R C 1977 *Intrinsic and Thermal Stress Modeling for Thin-Film Multilayers* (Norwalk, CT: The Perkin Elmer Corporation)
- [63] Bovard B G, Lega X C d, Hahn S-H and Macleod H A 1991 *Intrinsic Stress in Titanium Dioxide Thin Films Produced by Ion-Assisted Deposition* Private communication (Optical Sciences Center, University of Arizona)
- [64] Pulker H K 1982 Stress, adherence, hardness and density of optical thin films *Proc. Soc. Photo-Opt. Instrumentation Eng.* **325** 84–92
- [65] Pellicori S F 1984 Stress modification in cerous fluoride films through admixture with other fluoride compounds *Thin Solid Films* **113** 287–95
- [66] Windischmann H 1987 An intrinsic stress scaling law for polycrystalline thin films prepared by ion beam sputtering *J. Appl. Phys.* **62** 1800–7
- [67] Jacobsson R and Kruse B 1973 Measurement of adhesion of thin evaporated films on glass substrates by means of the direct pull method *Thin Solid Films* **15** 71–7
- [68] Heavens O S 1950 Some features influencing the adhesion of films produced by vacuum evaporation *J. Phys. Rad.* **11** 355–60
- [69] Benjamin P and Weaver C 1960 Measurement of adhesion of thin films *Proc. R. Soc. A* **254** 163–76
- [70] Benjamin P and Weaver C 1960 Adhesion of metal films to glass *Proc. R. Soc. A* **254** 177–83
- [71] Laugier M 1981 The development of the scratch test technique for the determination of the adhesion of coatings *Thin Solid Films* **76** 289–94
- [72] Laugier M 1981 Unusual adhesion-aging behaviour in ZnS thin films *Thin Solid Films* **75** L19–20

- [73] Goldstein I S and DeLong R 1982 Evaluation of microhardness and scratch testing for optical coatings *J. Vacuum Sci. Technol.* **20** 327–30
- [74] Bangert H and Pfefferkorn H 1980 Condensation and stability of ZnS thin films on glass substrates *Appl. Opt.* **19** 3878–9
- [75] Hass G, Ramsay J B and Thun R 1958 Optical properties and structure of cerium dioxide films *J. Opt. Soc. Am.* **48** 324–7
- [76] Auwärter M 1960 *Process for the Manufacture of Thin Films* USA Patent 2 920 002
- [77] Vogt A 1957 *Improvements in or Relating to the Manufacture of Thin Light-Transmitting Layers* UK Patent 775 002
- [78] Brinsmaid D S, Keenan W J, Koch G J and Parsons W F Eastman Kodak Co 1957 *Method of Producing Titanium Dioxide Coatings* USA Patent 2 784 115
- [79] Chiao S C, Borard B G and Macleod H A 1998 Repeatability of the composition of titanium oxide films produced by evaporation of Ti_2O_3 *Appl. Opt.* **37** 5284–90
- [80] Pulker H K, Paesold G and Ritter E 1976 Refractive indices of TiO_2 films produced by reactive evaporation of various titanium-oxide phases *Appl. Opt.* **15** 2986–91
- [81] Apfel J H 1980 The preparation of optical coatings for fusion lasers *Int. Conf. on Metallurgical Coatings (San Diego)*
- [82] Ritter E 1962 Zur Kenntnis des SiO und Si_2O_3 —Phase in dünnen Schichten *Opt. Acta* **9** 197–202
- [83] Bradford A P, Hass G, McFarland M and Ritter E 1965 Effect of ultraviolet irradiation on the optical properties of silicon oxide films *Appl. Opt.* **4** 971–6
- [84] Mickelsen R A 1968 Effects of ultraviolet irradiation on the properties of evaporated silicon oxide films *J. Appl. Phys.* **39** 4594–600
- [85] Heitmann W 1971 Reactive evaporation in ionized gases *Appl. Opt.* **10** 2414–18
- [86] Ebert J 1982 Activated reactive evaporation *Proc. Soc. Photo-Opt. Instrumentation Eng.* **325** 29–38
- [87] Smith D and Baumeister P W 1979 Refractive index of some oxide and fluoride coating materials *Appl. Opt.* **18** 111–15
- [88] Lingg L J, Targove J D, Lehan J P and Macleod H A 1987 Ion-assisted deposition of lanthanide trifluorides for VUV applications *Proc. Soc. Photo-Opt. Instrumentation Eng.* **818** 86–92
- [89] Lingg L J 1990 Lanthanide trifluoride thin films: structure, composition and optical properties *PhD Dissertation* (University of Arizona)
- [90] Stetter F, Esselborn R, Harder N, Friz M and Tolles P 1976 New materials for optical thin films *Appl. Opt.* **15** 2315–17
- [91] Baumeister P W and Arnon O 1977 Use of hafnium dioxide in multilayer dielectric reflectors for the near uv *Appl. Opt.* **16** 439–44
- [92] Lubezky I, Ceren E and Klein Z 1980 Silver mirrors protected with Yttria for the 0.5 to 14 μm region *Appl. Opt.* **19** 1895
- [93] Cox J T and Hass G 1978 Protected Al mirrors with high reflectance in the 8–12-mm region from normal to high angles of incidence *Appl. Opt.* **17** 2125–6
- [94] Datta U 1979 *Molybdenum Boats Good for Germanium Evaporation* Private communication (New Delhi, India)
- [95] Moss T S 1952 Optical properties of tellurium in the infra-red *Proc. Phys. Soc.* **65** 62–6
- [96] Greenler R G 1955 Interferometry in the infrared *J. Opt. Soc. Am.* **45** 788–91
- [97] Smith S D and Seeley J S 1968 *Multilayer Filters for the Region 0.8 to 100 Microns* (Air Force Cambridge Research Laboratories)

- [98] Seeley J S, Hunneman R and Whatley A 1981 Far infrared filters for the Galileo-Jupiter and other missions *Appl. Opt.* **20** 31–9
- [99] Evans C S and Seeley J S 1968 Properties of thick evaporated layers of PbTe *Paper presented at the Colloquium on IV–VI Compounds (Paris)*
- [100] Evans C S, Hunneman R, Seeley J S and Whatley A 1976 Filters for the V2 band of CO₂: monitoring and control of layer deposition *Appl. Opt.* **15** 2736–45
- [101] Evans C S, Hunneman R and Seeley J S 1976 Increments at the interface between layers during infra-red filter manufacture *Opt. Acta* **23** 297–303
- [102] Evans C S, Hunneman R and Seeley J S 1976 Optical thickness changes in freshly deposited layers of lead telluride *J. Phys. D* **9** 321–8
- [103] Yen Y-H, Zhu L-X, Zhang W-D, Zhang F-S and Wang S-Y 1984 Study of PbTe optical coatings *Appl. Opt.* **23** 3597–601
- [104] Zhang K G, Seeley J S, Hunneman R and Hawkins G J 1989 Optical and semiconductor properties of lead telluride coatings *Proc. Soc. Photo-Opt. Instrumentation Eng.* **1112** 393–402
- [105] Hass G and Salzberg C D 1954 Optical properties of silicon monoxide in the wavelength region from 0.24 to 14.0 microns *J. Opt. Soc. Am.* **44** 181–7
- [106] Placido F 1997 *RF Sputtering of Aluminium Oxynitride Rugates. Micrographs of Rugate Structures* Private communication (Department of Physics, University of Paisley)
- [107] Bovard B B, Ramm J, Hora R and Hanselmann F 1989 Silicon nitride thin films by low voltage reactive ion plating: optical properties and composition *Appl. Opt.* **28** 4436–41
- [108] Jacobsson R and Martensson J O 1966 Evaporated inhomogeneous thin films *Appl. Opt.* **5** 29–34
- [109] Fujiwara S 1963 Refractive indices of evaporated cerium dioxide–cerium fluoride films *J. Opt. Soc. Am.* **53** 880
- [110] Fujiwara S 1963 Refractive indices of evaporated cerium fluoride–zinc sulphide films *J. Opt. Soc. Am.* **53** 1317–18
- [111] Kogaku N Nippon Kogaku K K 1965 *Surface-Coated Optical Elements* UK Patent 1 010 038
- [112] Yadava V N, Sharma S K and Chopra K L 1974 Optical dispersion of homogeneously mixed ZnS–MgF₂ films *Thin Solid Films* **22** 57–66
- [113] Yadava V N, Sharma S K and Chopra K L 1973 Variable refractive index optical coatings *Thin Solid Films* **17** 243–52
- [114] Libbey-Owens-Ford Glass Company 1947 *Method of Coating with Quartz by Thermal Evaporation* UK Patent 632 442
- [115] Kraus T and Rheinberger P Balzers Patent und Lizenz Anstalt 1962 *Use of a Rare Earth Metal in Vaporizing Metals and Metal Oxides* US Patent 3 034 924
- [116] Balzers Patent und Lizenz Anstalt 1962 *Improvements in and Relating to the Oxidation and/or Transparency of Thin Partly Oxidic Layers* UK Patent 895 879
- [117] Fritz M, Koenig F, Merck E and Feiman S 1992 New materials for production of optical coatings *35th Annual Technical Conf. Proc. (Albuquerque, NM: Society of Vacuum Coaters)* pp 143–7
- [118] Butterfield A W 1974 The optical properties of Ge_xSe_{1-x} thin films *Thin Solid Films* **23** 191–4

Chapter 10

Factors affecting layer and coating properties

10.1 Microstructure and thin-film behaviour

One of the most significant features of optical thin films is the way in which their properties and behaviour differ from those of identical materials in bulk form. This is, of course, also true for thin films in areas other than optics. Almost always, the performance of the film is poorer than that of the corresponding bulk material. Refractive index is usually lower, although, very occasionally, for some semiconductor materials it can be slightly higher, losses greater, durability less and stability inferior. There is also a sensitivity to deposition conditions, especially substrate temperature.

Heitman [1] has studied the influence of parameters, such as the residual gas pressure within the plant and the rate of deposition, on the refractive indices of cryolite and thorium fluoride. Raising the residual gas (nitrogen) pressure from 4×10^{-6} torr (5.3×10^{-6} mb) in one case, and 2×10^{-6} torr (2.6×10^{-6} mb) in another, to 2×10^{-5} torr (2.6×10^{-5} mb) had no measurable effect, within the accuracy of the experiment ($\pm 0.1\%$ for thorium fluoride and $\pm 0.3\%$ for cryolite) while a further increase in residual pressure to 2×10^{-4} torr (2.6×10^{-4} mb) gave a drop in index of 1.5% for cryolite, and 1.4% for thorium fluoride. At this higher pressure, the mean free path of the nitrogen molecules was less than the distance between boat and substrate, and the decrease in refractive index was probably caused by increased porosity of the layers. This tends to confirm that the mean free path of the residual gas molecules should be kept longer than the source–substrate distance, but that any further increases in mean free path beyond this have little effect. Heitman concluded that the mean free path of the molecules is the important parameter, not the ratio of the numbers of evaporant molecules to residual gas molecules impinging on the substrate in unit time, which appeared to have no effect on refractive index. He also found that changes in the rate of deposition, from a quarter-wave in 0.5 min (measured at 632.8 nm) to a quarter-

wave in 1.5 min, caused a decrease in refractive index of 0.6% in both cases, but that a further decrease to a quarter-wave in 5 min produced only slight variations.

Heitman's results are probably best interpreted in terms of slight changes in film structure, induced by the variations in deposition conditions. Layer structure is, in fact, the most significant factor in determining the properties of optical thin films and the way in which they differ from the same material in bulk form. During the past two decades, there has been an increasing interest in the structure of, and structural effects in, optical thin films.

A useful technique for the study of thin-film structure, which immediately yielded important results, is electron microscopy. Its use in the examination of thin-film coatings has involved the development of techniques for fracturing multilayers and for replicating the exposed sections. Pearson, Lissberger, Pulker and Guenther [2–5] have all made substantial contributions in this area and their results show that the layers in optical coatings have, almost invariably, a pronounced columnar structure, with the columns running across the films normal to the interfaces. To their investigations, we can add those of Movchan and Demchishin [6] and then Thornton [7, 8], who investigated the effects of substrate temperature and, in Thornton's case, residual gas pressure, on the structure of evaporated and sputtered films. This showed that a critical parameter in vacuum deposition of thin films is the ratio of the temperature of the substrate T_s to the melting temperature T_m of the evaporant. For values of this ratio lower than around 0.5, the structure of the layers is intensely columnar, the columns running along the direction of growth. Increased gas pressure forces the growth into a more pronounced columnar mode even for slightly higher values of substrate temperature.

Because the most useful materials in optical thin films are all of high melting point, substrate temperatures can never be higher than a small fraction of the evaporant melting temperature, and so the structure of thin films is almost invariably a columnar one, with the columns running along the direction of growth, normal to the film interfaces. The columns are several tens of nanometres across and roughly cylindrical in shape. They are packed in an approximately hexagonal fashion with gaps in between the columns, which take the form of pores running completely across the film, and there are large areas of column surface which define the pores and are in this way exposed to the surrounding atmosphere. The columnar structure of a film of zinc sulphide is shown in figure 10.1 [9].

Packing density p defined as:

$$p = \frac{\text{Volume of solid part of film (i.e. columns)}}{\text{Total volume of film (i.e. pores plus columns)}}$$

is a very important parameter. It is usually in the range 0.75–1.0 for optical thin films, most often 0.8–0.95, and seldom as great as unity. A packing density that is less than unity reduces the refractive index below that of the solid material of the columns. A useful expression that is reasonably accurate for films of low index [10, 11] connects the index of the film n_f that of the solid part of the film n_s and

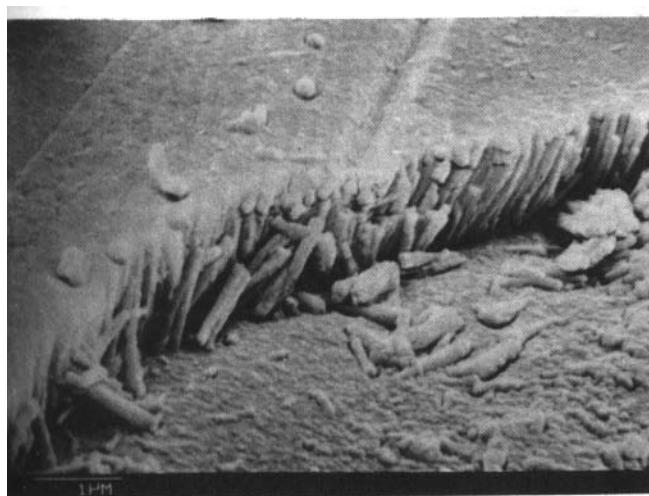


Figure 10.1. The columnar structure of a zinc sulphide film. Part of the film has been mechanically removed leaving the columnar structure visible in the cross section. (After Reid *et al* [9].)

of the voids n_v with the packing density p :

$$n_f = p n_s + (1 - p) n_v. \quad (10.1)$$

The behaviour of films of higher index, 2.0 and above, can be rather more complicated but in many cases a linear law as in equation (10.1) is sufficiently accurate and is, therefore, often employed. If the value of packing density has been derived from optical measurements by using equation (10.1), as is frequently the case, then, of course, the expression can, and should, be used. In any event, it gives an indication of the correct trend. For an alternative expression that is more complicated and gives a better fit in many of the more complicated cases, although still far from ideal, the paper by Harris and colleagues [11] should be consulted.

Packing density is a function of substrate temperature, usually, but not always, increasing with substrate temperature, and of residual gas pressure, decreasing with rising pressure. Film refractive index, therefore, is also affected by substrate temperature and residual gas pressure. The columns frequently vary in cross-sectional area as they grow outwards from the substrate surface, which is a major cause of film inhomogeneity. Substrate temperature is a difficult parameter to measure and to control so that consistency in technique, heating for the same period each batch, identical rates of deposition, pumping for the same period before commencing deposition, and so on, is of major importance in assuring a stable and reproducible process. Changing the substrate dimensions,

especially substrate thickness, from one run to the next can cause appreciable changes in film properties. Such changes are even more marked in the case of reactive processes where the residual gas pressure is raised, and where a reaction between evaporant and residual atmosphere takes place at the growing surface of the film. Thus it should not be surprising that a very high proportion of test runs are required in any manufacturing sequence.

Various modelling studies [12–15] have confirmed that the columnar growth results from the limited mobility of the material on the surface of the growing film. It diffuses over the surface under thermal excitation until it is buried by arriving material. Diffusion through the bulk of the material is not significant. Thus lower substrate temperature and higher rates of deposition lead to more pronounced columns and reduced packing density. The energetic processes involve an element of bombardment of the growing films. The transfer of momentum drives the material deeper into the film and, although the columnar structure may persist to some extent, squeezes out the voids. The packing density is normally close to or equal to unity. The results of the higher packing density are almost all favourable. The consequences described in this chapter of the columnar microstructure are all less serious in the energetically deposited films. (See figure 10.2 [16].)

A second level of microstructure in thin films is their crystalline state. This is less well understood but considerable progress has been made. Optical thin films are deposited from vapour that has been derived from sources at comparatively very high temperature. The substrates on which the films grow are at relatively very low temperature. There is therefore a great lack of equilibrium between growing film and arriving vapour. The film material is rapidly cooled or quenched, and this not only influences the formation of the columnar microstructure but it also affects the crystalline order. The material that is condensing will attempt to reach the equilibrium form appropriate to the temperature of the substrate, but the correct rearrangement of the molecules will take a certain time, and the film will tend to pass through the higher temperature forms during this rearrangement. If the rate of cooling is greater than the rate of crystallisation, then a higher temperature form will be frozen into the layer. The very rapid cooling rate normally existing in thin films implies the presence of quite high temperature forms and there are often mixtures of phases. This explains an, at first sight, curious behaviour of thin films. Frequently there is an inversion in the crystalline structure in that at low substrate temperatures a predominance of high-temperature crystalline forms are found, whereas at high substrate temperatures, more low-temperature material appears to form. The low substrate temperature leads to a higher quench rate and the rest follows [17]. Amorphous forms, corresponding to a quite high temperature, can often be frozen by very rapid cooling, and are enhanced by a higher temperature of the arriving species. For example sputtering, where additional kinetic energy is possessed by the arriving molecules, often gives amorphous films. The low voltage ion-plating technique, again with high incident energy, appears virtually invariably to give amorphous films. The high temperature forms are often only metastable and may change their

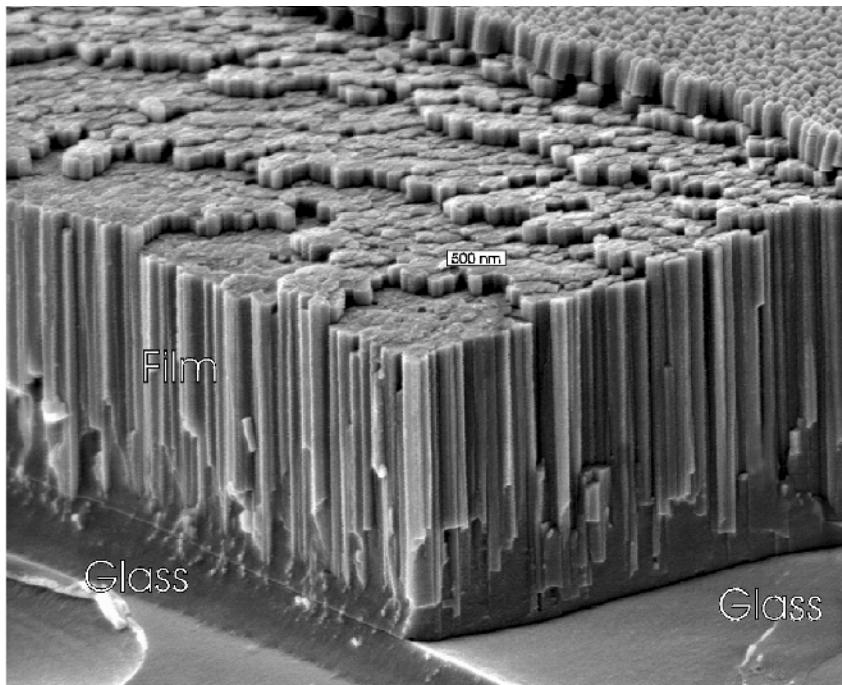


Figure 10.2. Compact microstructure of an aluminium oxynitride rugate structure deposited by radio frequency reactive sputtering of aluminium. The packing density is very high but some columnar features remain. The fractures at the outer surface tend to be in the nitrogen-rich parts of the rugate cycle leading to the stepped appearance. (Courtesy of Professor Frank Placido [16].)

structure at quite low temperatures leading to problems of various kinds. Some films deposited in amorphous form by sputtering may sometimes be induced to recrystallise, in a manner described as explosive, by a slight mechanical disturbance, such as a scratch, or by laser irradiation [18].

Samarium fluoride has two principal crystalline forms, a hexagonal high-temperature form and an orthorhombic low-temperature form. Table 10.1 shows the results of thermal evaporation and ion-assisted deposition which both lead to this apparently inverted structure [17]. Zirconia has three principal structures, monoclinic, tetragonal and cubic in ascending temperature. Klinger and Carniglia [19] found that very thin zirconia shows a cubic structure, but becomes monoclinic when thicker than a quarter-wave at 600 nm. This behaviour can be explained by a lower rate of quenching when the film is thicker and less thermally conducting. Alumina, normally amorphous in thin-film form, can recrystallise in the electron microscope when subjected to the electron bombardment necessary for viewing [20]. Amorphous zirconia, which can occur

Table 10.1. Samarium Fluoride (SmF_3) [17].

Normal high temperature form		Hexagonal
Normal low temperature form		Orthorhombic
Thermal evaporation	Substrate temperature of $100\text{ }^\circ\text{C}$	Hexagonal (111)
	Substrate temperature $\geq 200\text{ }^\circ\text{C}$	Orthorhombic (111) with some hexagonal
Ion-assisted deposition	Substrate temperature $100\text{ }^\circ\text{C}$	Hexagonal (110) with some (111)
	Higher bombardment at substrate temperature $100\text{ }^\circ\text{C}$	Hexagonal (110) with appearance of new peak $\text{SmF}_2(111)?$

when films are very thin, has been shown to exhibit similar behaviour [21].

Thin films, therefore, are complicated mixtures of different crystalline phases, some being high-temperature metastable states. Such behaviour is clearly very material- and process-dependent and each specific system requires individual study. What is a good structure for one application may not be so for another. The low scattering of the amorphous phases make them attractive for certain applications, but their high-temperature or high-flux behaviour may not be as satisfactory. Much more needs to be done in attempting to improve our understanding.

The columnar structure and the crystalline structure can be considered as essentially regular intrinsic features of film microstructure. Then, in addition, there are defects that can be thought of as local disturbances of the intrinsic features. A principal and very important class of defect is the nodule. Nodules are inverted conical growths that propagate through the film or multilayer. They can occur in all processes. They start at a seed that is usually a very small defect or irregularity and it appears that virtually any irregularity, even minute ones, may act as a seed. Scratches on the substrate, pits, dust, contamination, material particles ejected from the source, loose accumulations of material in the vapour phase, perhaps even local electric charges, can all cause nodules to start growing. Once the nodule starts, it continues to grow until it forms a domed protrusion at the outer surface of the multilayer. The nodule itself is very much larger than the defect that causes it. It is not, in itself, a contaminant. It is made up of exactly the material of the remainder of the coating. It is simply growing in a different way. The outer surface of the nodule is a quite sharp boundary between it and the remainder of the coating. This sharp boundary is a region of weakness and there is frequently a fissure around the nodule, either partially or completely,

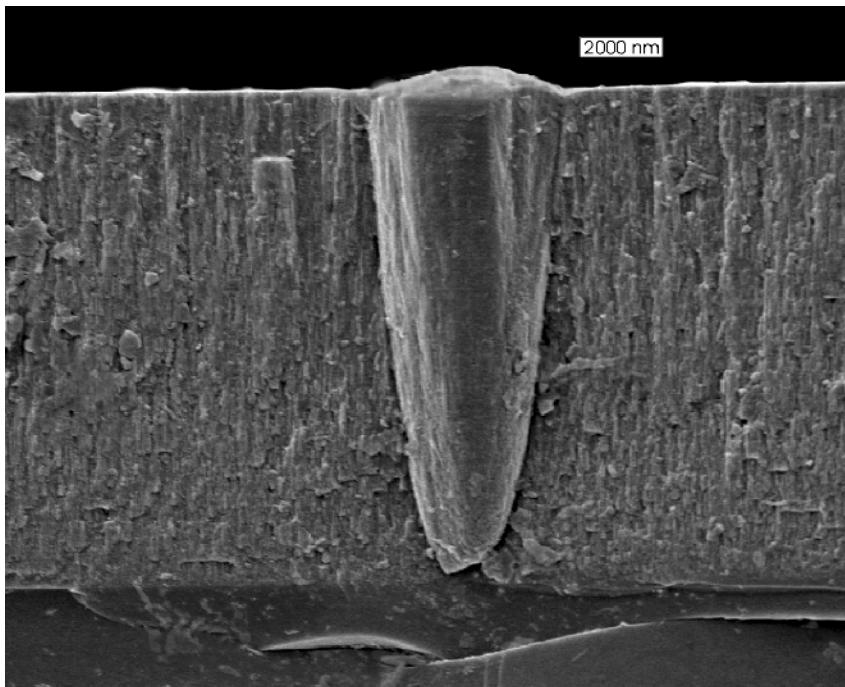


Figure 10.3. A nodule. The film is a rugate structure of aluminium oxynitride deposited by radio-frequency (RF) reactive sputtering of aluminium. The film has been broken across its width to show a cross-section that includes a complete nodule. The sharpness of the boundary is clear and the weakness is shown by the fact that the crack in the film circles around the nodule rather than passing through it. The shape and the domed protrusion at the outer surface (upper) of the film system are typical. (Courtesy of Professor Frank Placido [16].)

and the nodule may sometimes be detached from the coating completely, leaving a hole behind. Nodules are present in almost all coatings. The only way of suppressing them appears to be a move towards perfection in the substrate, its surface and its preparation, and in the coating deposition. The incidence of nodules over superpolished substrates, for example, is much reduced compared with conventional substrates. A typical nodule is shown in figure 10.3 and the hole left by a detached nodule in figure 10.4.

Variation in refractive index is not the only feature of film behaviour associated with the columnar structure. The pores between the columns permit the penetration of atmospheric moisture into the film, where, at low relative humidity, it forms an adsorbed layer over the surfaces of the columns and, at medium relative humidity, actually fills the pores with liquid water due to capillary

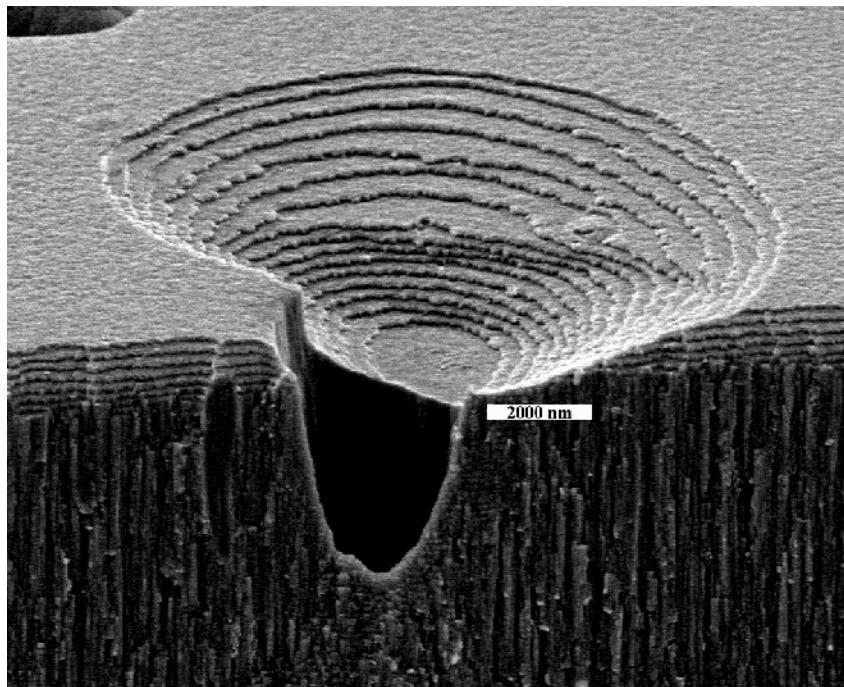


Figure 10.4. The hole left by the detachment of a nodule. Part of the outer part of the structure has been removed along with the nodule. The stepped appearance is once again caused by preferential cracking in the nitrogen-rich part of the aluminium oxynitride rugate structure. (Courtesy of Professor Frank Placido [16].)

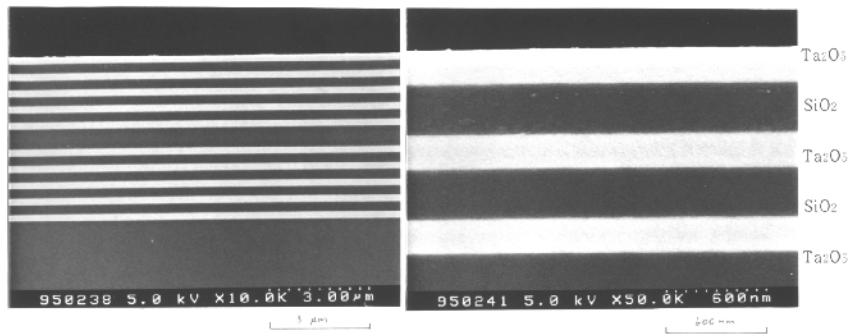


Figure 10.5. A micrograph showing the compact amorphous structure of a narrowband filter of silica and tantalum produced by ion-assisted deposition using an RF ion-gun. (Courtesy of Shincron Co. Ltd, Tokyo, Japan.)

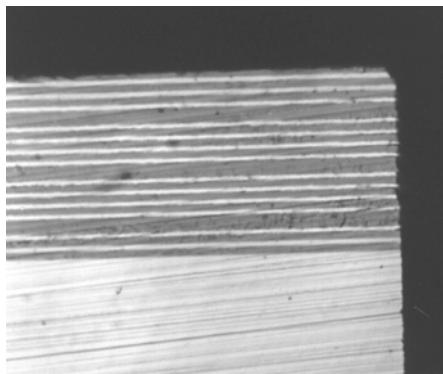


Figure 10.6. The structure of a multiple-cavity filter for the far infrared constructed from lead telluride and zinc sulphide. This particular filter was one of a set for the region 6–18 μm required to have a size of 1.2 mm \times 0.45 mm for use in the High Resolution Dynamic Limb Sounder (HIRLDS) and the high quality of the diamond sawn edge of the component is clear from the micrograph. The scale of the micrograph can be assessed from the 4 μm physical thickness of the cavity layers. (Courtesy of Roger Hunneman, University of Reading, England.)

condensation. Moisture adsorption has been the subject of considerable study by Ogura [22, 23], who used the variation in adsorption with relative humidity to derive information on the pore structure of the films. The moisture, since it has a different refractive index (around 1.33) from the 1.0 of the air, which it displaces from the voids, causes an increase in the refractive index of the films. Since the geometrical thickness of the film does not change, the increase of film index during adsorption is accompanied by a corresponding increase in optical thickness. Exposure of a film to the atmosphere, therefore, usually results in a shift of the film characteristic to a longer wavelength. Such shifts in narrowband filters have been the subject of considerable study. Schildt *et al* [24] found that for freshly prepared filters of zinc sulphide and magnesium fluoride, constructed for the region 400–500 nm, the variation in peak wavelength could be expressed as

$$\Delta\lambda = q \log_{10} P$$

where q is a constant varying from around 1.4 for filters which had aged to around 8.3 for freshly prepared filters, and P is the partial pressure of water vapour measured in torr (P should be replaced by $0.76 \times P$ if P is measured in mb) and $\Delta\lambda$ is measured in nm. $\Delta\lambda$ was arbitrarily chosen as zero when the pressure was 1 torr (1.3 mb). This relationship was found to hold good for the pressure range 1 to approximately 20 torr (1.3–26 mb). The filters settled down to the new values of peak wavelength some 10–20 minutes after exposure to a new

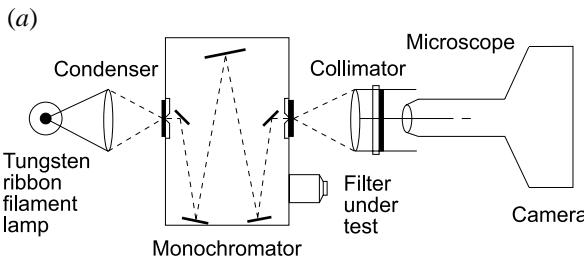


Figure 10.7. Moisture-penetration patterns in a multilayer of zinc sulphide and cryolite. (a) Sketch of the apparatus for observing the phenomenon. Short slits that are virtually pin holes are used in the monochromator. (After Macleod and Richmond [27].) (b) Photograph of moisture-penetration patterns in a zinc sulphide and cryolite filter some two weeks after coating. The relative humidity was approximately 50% during this time. The upper photograph was taken at a wavelength of 488.5 nm and the lower at 512.8 nm. The dark patches of the upper photograph correspond to the light patches of the lower showing that a wavelength shift rather than absorption is responsible for the patterns. (After Lee [29].)

level of humidity began. They found that the shifted values of peak wavelength could be stabilised by cementing cover slips over the layers using an epoxy resin. Koch [25, 26] showed that the characteristics of narrowband filters became quite unstable during adsorption until the filters reached an equilibrium state. Macleod and Richmond [27], Richmond [28] and Lee [29] have made detailed studies of the effects of adsorption on the characteristics of narrowband filters. The results are applicable to all types of multilayer coating. The shifts in the characteristics are due, as we have seen, to the filling of the pores of the film with liquid water. In multilayers, the pores of one film are not always directly connected with the pores of the next, and the penetration of atmospheric moisture is frequently a slow and complex process in which a limited number of penetration pores take part, from which the moisture spreads across the coating in increasing circular patches. The primary entry points for the moisture are thought to be nodules where capillary condensation can take place in the fissures that often surround them. The coating may take several weeks to reach equilibrium and, afterwards, will exhibit some instability should the environmental conditions change. The patches, which can sometimes be seen with the naked eye as a flecked or mottled appearance, can be made more visible if the coating is viewed in monochromatic light, at or near a wavelength for which there is a rapid variation of transmittance. The edge of an edge filter, or the pass band of a narrowband filter, are especially suitable. Wet patches show a shift in wavelength that changes them from high to low transmittance, or vice versa, and they can be readily photographed as was done in figures 10.7 and 10.8.

The drift of the filters towards longer wavelengths, which occurs on exposure to the atmosphere, varies considerably in magnitude with both the materials and the spectral region and there is frequently considerable hysteresis on desorption.

(b)

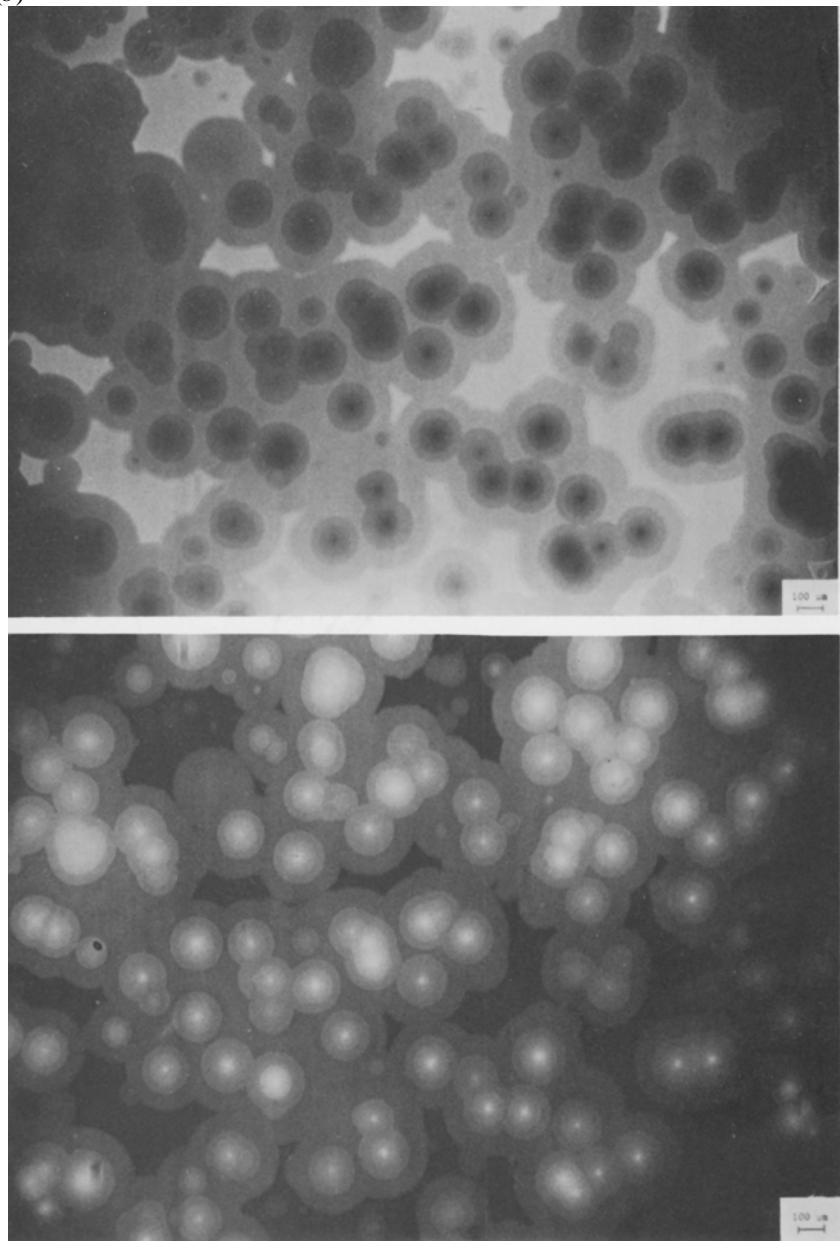


Figure 10.7. (Continued)

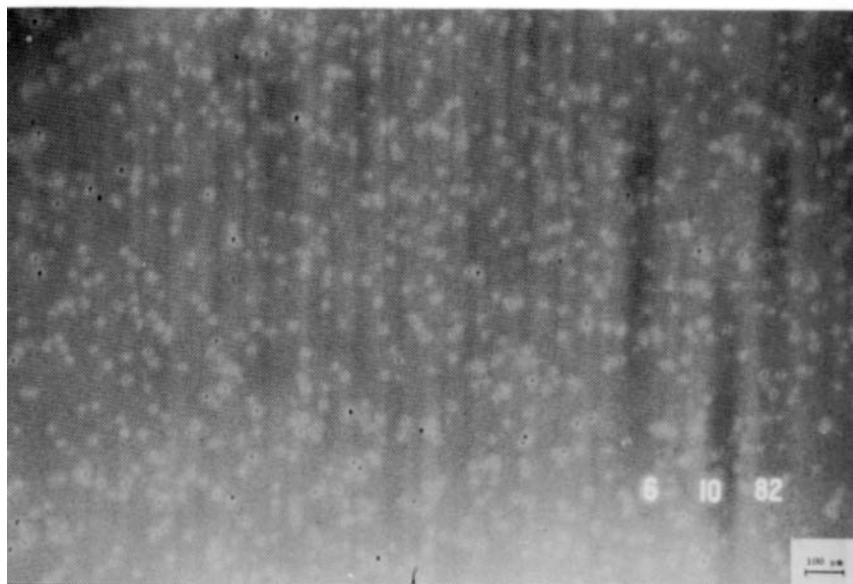
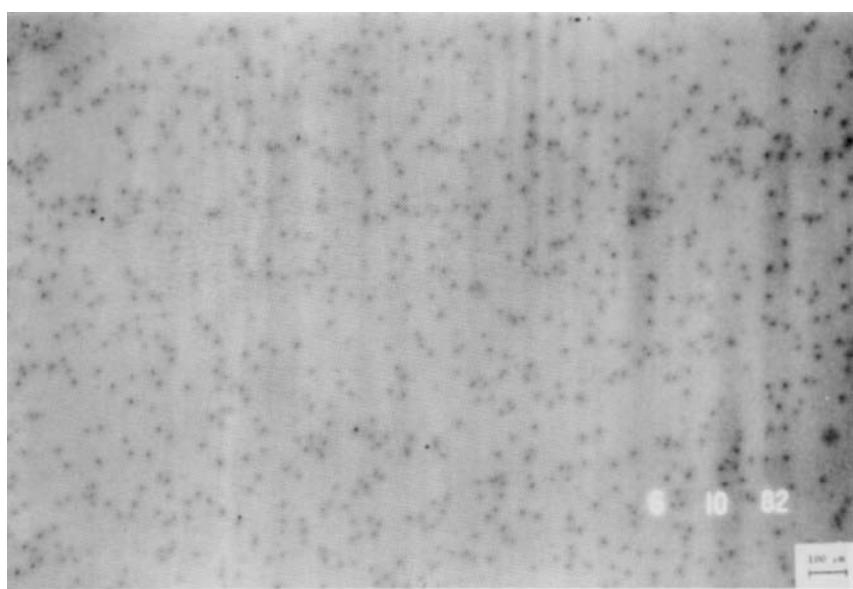


Figure 10.8. Moisture-penetration patterns in a multilayer of zirconium dioxide and silicon dioxide. The photographs were taken immediately after removal from the coating chamber. The wavelength for the upper photograph was 543 nm, and that for the lower 553 nm. (After Lee [29].)

In the infrared the layers are thick, and many of the semiconductor materials that are used as high-index layers have high packing density. This means that moisture-induced drift is less of a general problem than it is in the visible and ultraviolet regions of the spectrum, although it is important in some applications. In the visible region, drifts can be as high as 10 nm, and sometimes greater, towards longer wavelengths. The gradual stabilisation of the coating as it reaches equilibrium is frequently referred to as ageing or settling. The energetic processes can usually suppress completely the moisture-induced drifts and have been almost universally adopted for suitable coatings. It should be noted, however, that not all materials respond well to the brutal bombardment that is characteristic of the energetic processes. Metals suffer from the inevitable implantation of the bombarding species. Their optical properties are degraded by the scattering of conduction electrons that results. Fluorides lose fluorine and so the bombardment must be strictly limited otherwise the concentration of vacancy defects becomes too great. Oxygen tends to fill the vacancies and form oxyfluorides that are neither as rugged as the original fluorides nor as useful in the ultraviolet.

It is not simply in generating optical shifts that moisture is a problem for coatings. It has major mechanical and sometimes chemical effects as well. The stress in the coating is transmitted across the gaps between the columns, again by short-range forces. These forces can be very easily blocked by water molecules. An alternative explanation of the phenomenon is that the moisture, which coats the surfaces of the columns, reduces the surface energy to something approaching that of liquid water. Since the surface energy is an important factor in the stress/strain balance in the film, the result of the moisture adsorption is a change in the stress level. The stress is usually tensile and the moisture reduces it, usually significantly. We have already mentioned Pulker's work [30] on impurities in thin films and their reduction of stress levels in a similar way. Adhesion, too, is affected by moisture. The materials used for thin films have usually very high surface energies and then the work of adhesion is correspondingly high. The presence of liquid water in a film can cause a reduction in the surface energy of the exposed surfaces of at least an order of magnitude. If water is present at the site of an adhesion failure and can take part in a process of bond transfer, rather than bond rupture followed by adsorption, then it will reduce the work of adhesion, and it is more likely that the failure will propagate. There is frequently enough strain energy in a film to supply the required work. The penetration sites for the moisture patches are probably associated with defects which may act as stress concentrators where adhesion failures driven by the internal strain energy in the films may originate. All the ingredients for a moisture-assisted adhesion failure are present and it is frequently at such sites that delamination is first observed. Blistering is a similar form of adhesion failure frequently associated with moisture penetration sites and a compressively strained film.

We have already mentioned in chapter 7 that changes in temperature cause changes in the spectral characteristics of coatings, narrowband filters having characteristics that are probably most sensitive to such alterations. We must divide

the coatings into those that have been simply thermally evaporated and those that have been produced by an energetic process.

Most of the work that has been reported has been in respect of conventionally thermally evaporated coatings. For small temperature changes, the principal effect is a simple shift towards longer wavelengths with increasing temperature. For the materials commonly used in the visible region of the spectrum, the shift is of the order of $0.003\% \text{ }^{\circ}\text{C}^{-1}$, while for infrared filters it can be greater, and a useful figure is $0.005\% \text{ }^{\circ}\text{C}^{-1}$, although it can be as high as $0.0125\% \text{ }^{\circ}\text{C}^{-1}$. It must be emphasised that these figures depend strongly on the particular materials used. Filters of lead telluride and zinc sulphide can actually have negative coefficients greater than $0.01\% \text{ }^{\circ}\text{C}^{-1}$ and, using these materials, it is even possible to design a filter that has zero temperature coefficient [31]. With greater positive changes of, say, $60\text{ }^{\circ}\text{C}$ or more, it is usual for the moisture in the filter to desorb partially, causing an abrupt shift towards shorter wavelengths (see figure 10.9). This shift is not recovered immediately on cooling to room temperature, and so considerable hysteresis is apparent in the behaviour [32]. Subsequent temperature cycling, before readsorption of any moisture, will then exhibit no hysteresis. Eventually, if maintained at room temperature, the filter will readsorb moisture and drift gradually back to its initial wavelength. Exposure to higher temperatures still, over $100\text{ }^{\circ}\text{C}$, can cause permanent changes which appear to be related to minute alterations in the structure of the layers, altering the adsorption behaviour so that some materials become less ready to adsorb moisture while others show more rapid adsorption [27–29]. A frequently applied empirical treatment, already mentioned in chapter 9, involves baking of filters at elevated temperatures, usually several hundred degrees Celsius, for some hours. The baking process reduces residual absorption, particularly in reactively deposited oxide films, and improves the subsequent stability of the coatings. Part of the baking process appears to involve the opening up of the pores in the films, by smoothing out restrictions, so that moisture adsorption processes are more rapid and the films reach equilibrium in normal atmospheres much more quickly.

Films that have been deposited by the energetic processes usually exhibit lower temperature coefficients than thermally evaporated, even when the effects of moisture desorption and adsorption are removed. This is at first sight a quite surprising result. But the explanation appears to lie in the microstructure. The lateral thermal expansion of the loosely packed columns in the thermally evaporated films enhances the drifts due to temperature changes. In the energetically deposited films, the material is virtually bulk-like in that there are no voids in between any residual columns and so the material exhibits bulk-like properties. The change in characteristics with a change in temperature now corresponds to what would be expected from bulk materials. Indeed, Takahashi [33] has shown that for multiple-cavity narrowband filters, once the design and materials are chosen, the expansion coefficient of the substrate dominates the behaviour and can even change the sense of the induced spectral shift. The stress induced in the coating by the differential lateral expansion

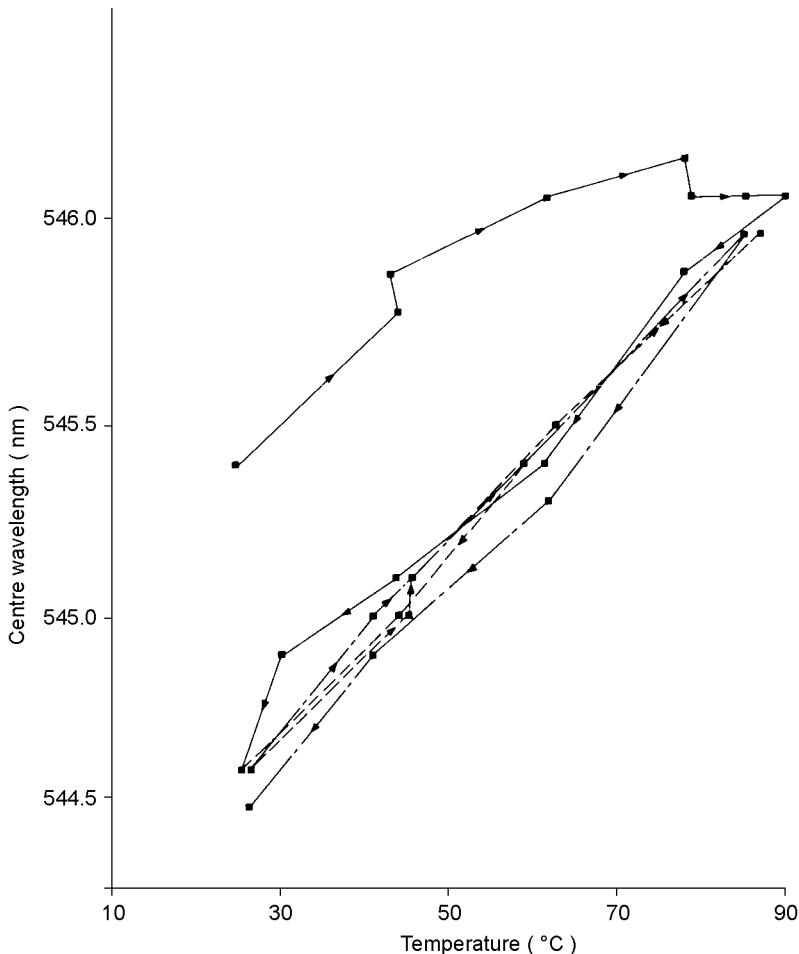


Figure 10.9. Record of the variation of peak wavelength with temperature for a filter with L = cryolite and H = Air|(HL)'6H(LH)'|Glass zinc sulphide. (After Roche *et al* [32].)

and contraction of substrate and coating is translated by Poisson's ratio into a swelling or reduction normal to the film surfaces. As a result of this modelling and improved understanding, temperature coefficients of peak wavelength shift at 1550 nm of $3 \text{ pm } ^\circ\text{C}^{-1}$ (pm is picometre, i.e. 0.001 nm so that $3 \text{ pm } ^\circ\text{C}^{-1}$ at 1550 nm represents $0.0002\% \text{ }^\circ\text{C}^{-1}$) have routinely been achieved in energetically deposited tantalum/silica filters for communication purposes and shifts as low as $1 \text{ pm } ^\circ\text{C}^{-1}$ are possible.

Coatings that are subjected to very low temperatures usually shift towards shorter wavelengths, consistent with their behaviour at elevated temperatures.

Filters are not usually affected mechanically except for laminated components that run the risk of breaking because of differential contraction and/or expansion.

There are losses associated with all layers, which can be divided into scattering and absorption. In absorption, the energy, which is lost from the primary beam, is dissipated within the coating and usually appears as heat. In scattering, the flux lost is deflected and re-emerges from the coating in a different direction. Absorption is a material property which may be intrinsic or due to impurities. A deficiency of oxygen, for example, can cause absorption in most of the refractory oxide materials. Scattering is usually due to defects in the coating that can be classified into volume or surface defects. Surface defects are simply a departure from the smooth flat surfaces of the ideal film. Such departures can be due to roughness of the substrate surface which tends to be reproduced at each interface in a multilayer, or to the columnar structure of the layers which results in a nodular appearance of the film boundaries. Volume defects are local variations of optical constants and are usually dust particles, pinholes or fissures in the coating.

Losses in thin films are of particular importance in the laser field where they determine the limiting performance of multilayers. A major problem in the production of high-quality laser coatings is dust that emanates from the sources and from the powdery deposit that forms on the cold walls of the chamber. If this dust can be eliminated, only possible if the strictest attention is paid to detail and the most involved precautions are taken, then the remaining source of scattering loss is the roughness of the interfaces between the layers and between multilayer and substrate. If great care is exercised, then, in the visible and near infrared regions, the total losses, that is, absorption and scattering, can be reduced below 0.001% (for some very special applications losses towards one-tenth of this figure have been achieved) and the power handling capability of the coatings can be of the order of 5 J cm^{-2} for pulses of 1 ns or less at $1.06 \mu\text{m}$. Recent useful surveys of scattering in thin-film systems have been written by Duparré [34–36] and by Amra [37, 38].

Laser damage is still a very active research topic. The best bulk crystals can exhibit intrinsic damage thresholds that are ultimately connected with multi-photon events causing the raising of electrons into the conduction band. Damage in thin-film systems, on the other hand, is dominated by the defects in the films so that the intrinsic level is not reached. In continuous wave applications, particularly in the infrared, thermal effects associated with absorption, either local or general, appear to be the principal source of damage, small defects appearing less important. In most other cases local defects are the problem. The particular nature of the defects may vary considerably, from inclusions to cracks or fissures, but considerable attention in recent years has been paid to the nodules that tend to grow through the films from any substrate imperfections. These nodules are poorly connected thermally to the film and this is suspected to be an important factor in the initiation of damage. In those spectral regions where water absorbs strongly, considerable importance is attached to the presence of liquid water

within the films. In other parts of the spectrum its role is less clear, but it may well play a part. Laser damage has been surveyed recently by Koslowski [39].

10.2 Sensitivity to contamination

Optical coatings are rarely used in an ideal environment. They are subjected to all kinds of environmental disturbances ranging from abrasion to high temperature and humidity. These cause performance degradation that mostly originates in an actual irreversible and usually visible destruction of the layers. However, performance may be degraded in a rather less spectacular way by the simple acquisition of a contaminant that may have no aggressive effect on the layers other than a reduction of the level of performance of the coating as a whole. The action of water vapour that is adsorbed by a process of capillary condensation and causes a spectral shift of the coating is well known. Here we are concerned with much smaller amounts of absorbing material, such as carbon, in the form of submolecular thicknesses either at some point during the construction of the coating or, more usually, over the surface after deposition.

Although there are many tests for the assessment of the resistance of a coating to most environmental disturbances there is no standard test for the measurement of susceptibility to contamination. Yet it can be shown that the response of coatings can vary enormously, depending on many factors including design, wavelength, and even on errors committed during deposition. The reason may be that, often, careful cleaning will restore the performance but this does not avoid the degradation in between cleanings, and more frequent cleanings are required for more susceptible coatings.

Fortunately it is possible to make some predictions of coating response to low levels of contamination and, especially, to make assessments of comparative sensitivity [40, 41]. Electric field distribution and potential absorption are the keys to understanding the phenomenon.

If the contamination layer is on the front surface then it receives the full irradiance that enters the multilayer, and the admittance at the contamination layer determines the reflectance as well as the potential absorptance. The key expressions involving absorptance, A , and potential absorptance, \mathcal{A} , have already been derived in chapter 2.

$$\mathcal{A} = \left(\frac{2\pi nkd}{\lambda} \right) \left(\frac{2}{\text{Re}(Y)} \right) \quad (10.2)$$

and

$$A = (1 - R)\mathcal{A}. \quad (10.3)$$

Then we can write

$$A = (1 - R)\mathcal{A}$$

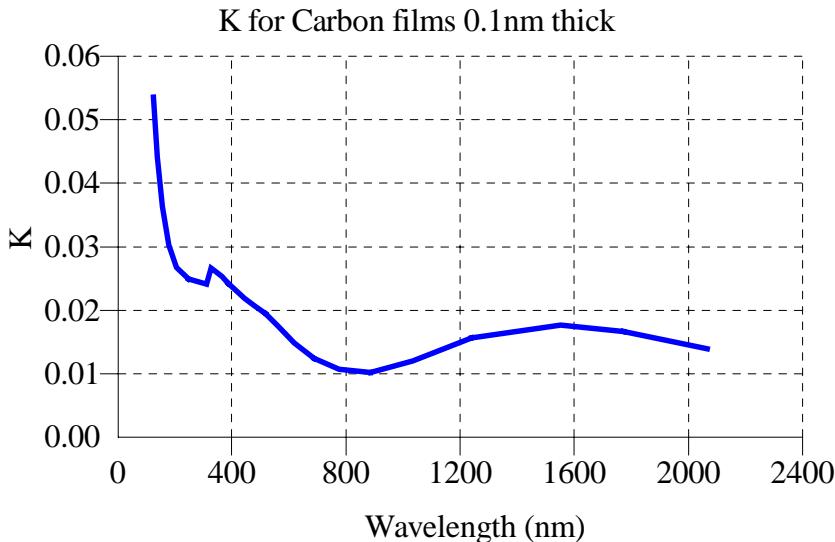


Figure 10.10. Plot of K against wavelength for 0.1 nm thickness of carbon film.

$$\begin{aligned}
 &= \left(\frac{4\pi nkd}{\lambda} \right) \left(\frac{1}{\text{Re}(Y)} \right) \left\{ 1 - \frac{[y_0 - \text{Re}(Y)]^2 + [\text{Im}(Y)]^2}{[y_0 + \text{Re}(Y)]^2 + [\text{Im}(Y)]^2} \right\} \\
 &= \left(\frac{4\pi nkd}{\lambda} \right) \left(\frac{4y_0}{[y_0 + \text{Re}(Y)]^2 + [\text{Im}(Y)]^2} \right)
 \end{aligned} \tag{10.4}$$

and equation (10.4) permits us to put on the admittance diagram contours of absorption due to contamination on the outer surface. Before we draw actual lines we need to define some of the quantities. It is simplest to use numbers that allow us to scale the diagram easily. We therefore simplify the expression by defining

$$\frac{16\pi nkd}{\lambda} = K. \tag{10.5}$$

Then

$$A = K \frac{y_0}{[y_0 + \text{Re}(Y)]^2 + [\text{Im}(Y)]^2}. \tag{10.6}$$

And if we replace Y by $x + iz$ then the equation giving the contours of constant A/K is

$$(y_0 + x)^2 + z^2 = y_0 \frac{K}{A} \tag{10.7}$$

that is, a circle with centre at the point $(-y_0, 0)$ on the negative branch of the real axis.

As an example of the magnitude of K we can take the values of amorphous carbon given by Palik [42, 43], that is optical constants of $2.26 - i1.025$ at

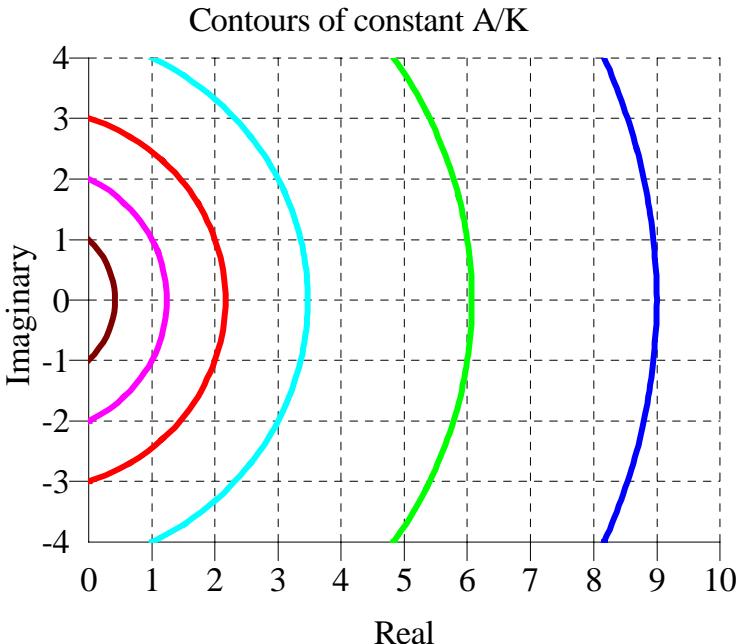


Figure 10.11. The contour lines of constant A/K in the admittance diagram assuming that y_0 is 1.00. From left to right (inner to outer circle) the values of A/K are 0.5, 0.2, 0.1, 0.05, 0.02, 0.01. The origin corresponds to a value of A/K of 1.00.

1000 nm, and assume a thickness of 0.1 nm. A plot of K is shown in figure 10.10 and over most of the wavelength region shown it is between 0.01 and 0.02.

To simplify matters still further we take the value of y_0 as 1.00. The contour lines for this case are then as shown in figure 10.11.

Antireflection coatings all attempt to terminate their loci at the point $(y_0, 0)$. This implies a value of A/K of $1/(4y_0)$, that is 0.25 for y_0 of unity, and, from figure 10.10, this gives, for a perfect antireflection coating, a range of absorptance across the visible region from around 0.25% to 0.7% with a film of carbon 0.1 nm thick. A slightly less than perfect coating may exhibit figures greater or less than these. It all depends on the admittance at termination. Typical results for a four-layer antireflection coating over the visible region are shown in figure 10.12. The design of the coating has little influence on this result and all coatings that have precisely zero reflectance will have exactly the same level of sensitivity.

Reflectors exhibit much greater variation. A dielectric reflector that is made up of quarter-wave layers and terminates with a final high-admittance layer will end its locus to the far right of the diagram and the sensitivity to contamination will be much reduced. This, however, is not so for extended-zone high-reflectance coatings. In such coatings at least part of the high-reflectance zone involves the

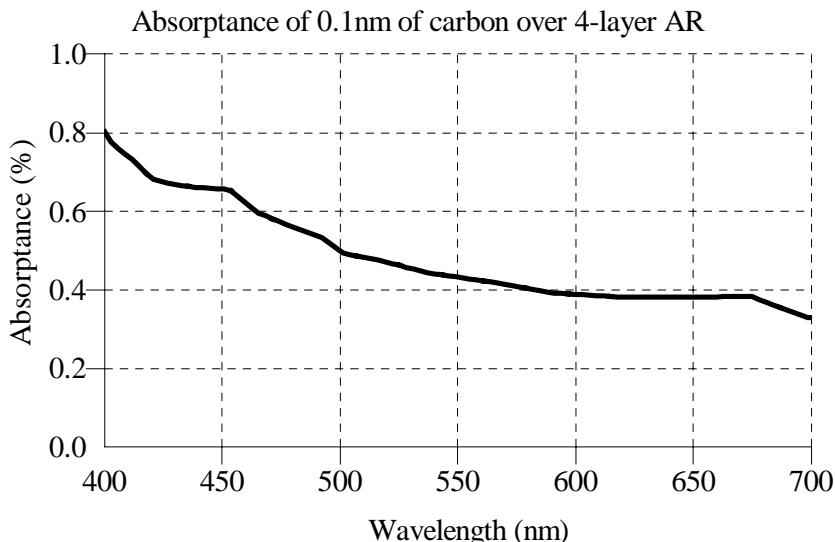


Figure 10.12. The absorptance produced by a layer of carbon of thickness 0.1 nm in front of a four-layer antireflection (AR) coating for the visible region.

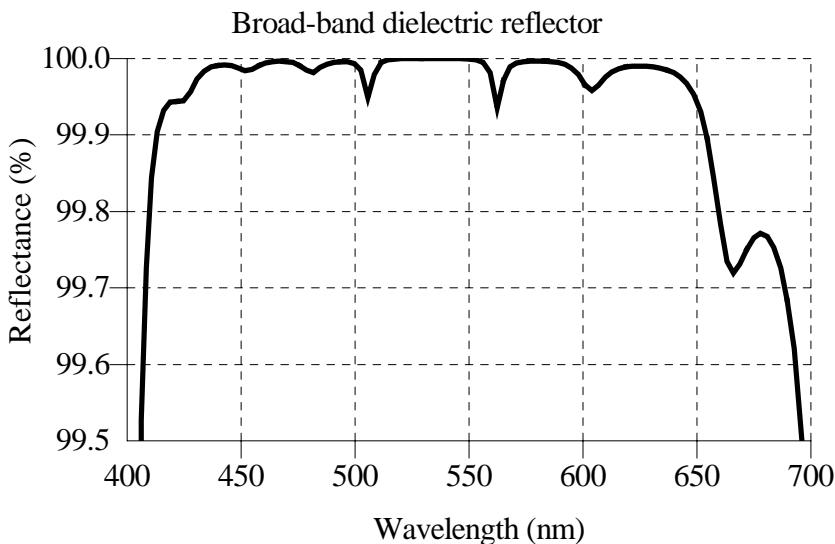


Figure 10.13. The reflectance of an extended-zone high-reflectance coating for the visible region. The coating consists of two mutually displaced quarter-wave stacks making up a total of 39 layers.

Design1: Absorptance

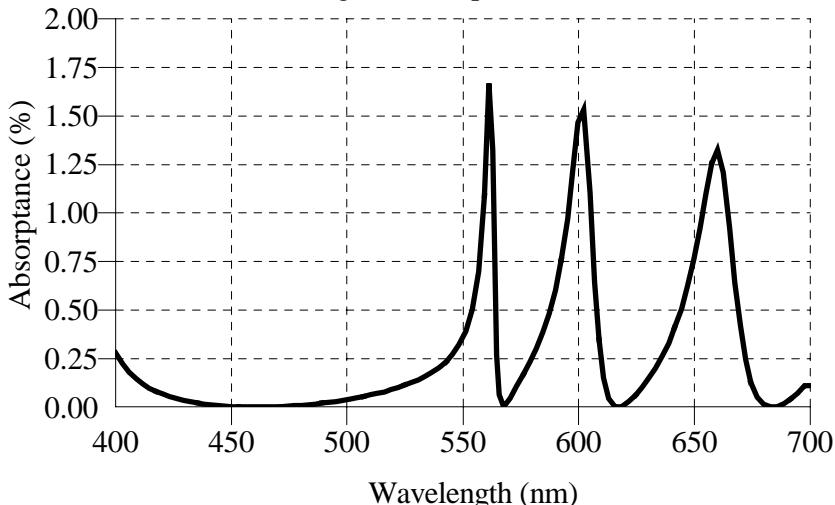


Figure 10.14. The absorptance produced by 0.1 nm of carbon deposited over the outer surface of the reflector of figure 10.13.

Absorptance: Upper quarterwave. Lower halfwave

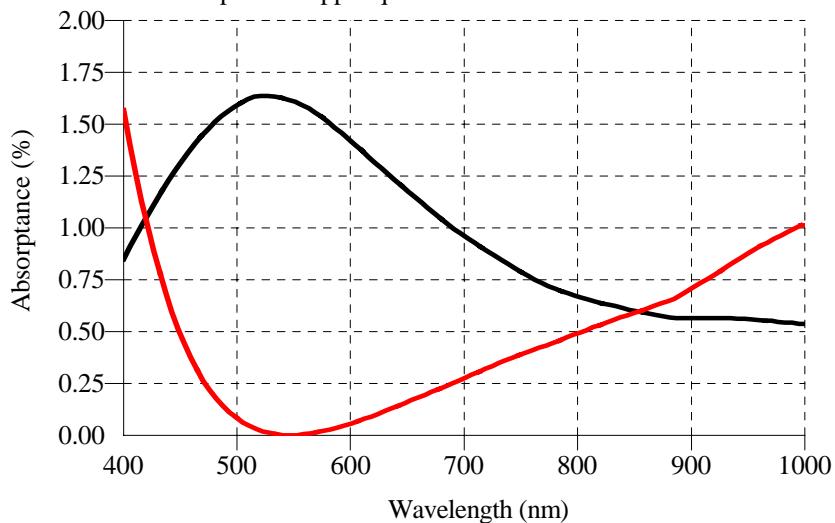


Figure 10.15. Effect of contamination by 0.1 nm thick film of carbon on aluminium reflector with quarter-wave of silica protecting layer (upper curve) and half-wave of silica (lower curve).

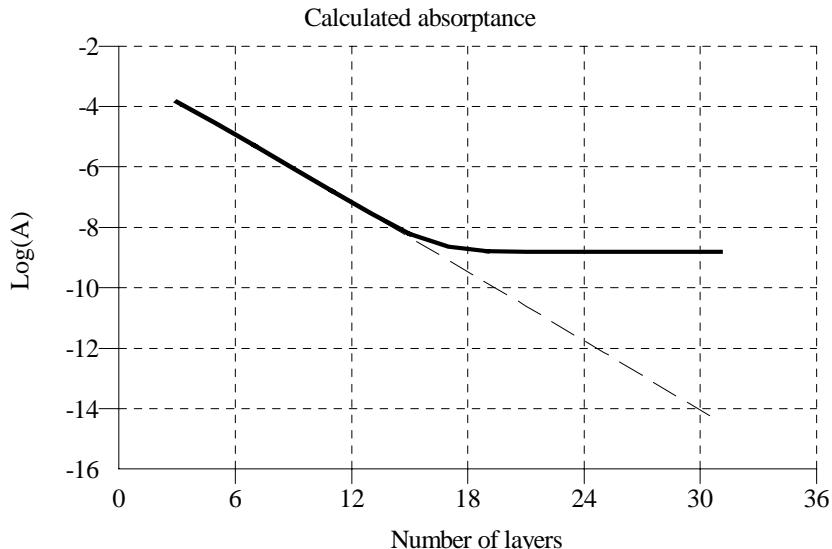


Figure 10.16. The predicted absorptance, plotted as $\log(A)$, of a quarter-wave stack as a function of the odd number of layers. The dashed line is the simple theory. The full line is calculated using the full matrix theory.

inner part of the coating and the outer part exhibits an admittance that circles around from far to the right to very near the imaginary axis. The value of A/K can then be almost as large as 1.0 so that over parts of the visible region the absorptance due to the 0.1 nm thickness of carbon can rise to between 1.0% and 2.0%. This is illustrated by a 39-layer extended zone reflector with performance as in figure 10.13 and absorptance behaviour as in figure 10.14.

Aluminium reflectors are normally protected by a thin layer of low index, most often a half-wave in thickness, although a quarter-wave may also be used. The quarter-wave thickness gives a greater fall in reflectance at the reference wavelength and also a higher electric field. The sensitivity to contamination of the two coatings is quite different and shown in figure 10.15

The simple quarter-wave stack is of enormous importance as the most common high-performance reflector. We have seen how poor the extended-zone high reflector is. What can we deduce about the quarter-wave stack? We can take the contamination figures as at 1000 nm. At the centre wavelength, where all layers are quarter-waves, the admittance presented by a quarter-wave stack, Y , is real. The absorptance of the layer, using the 1000 nm figures and assuming air as incident medium, is therefore given from equation (10.6), by

$$A = \frac{0.0116}{(1 + Y)^2}. \quad (10.8)$$

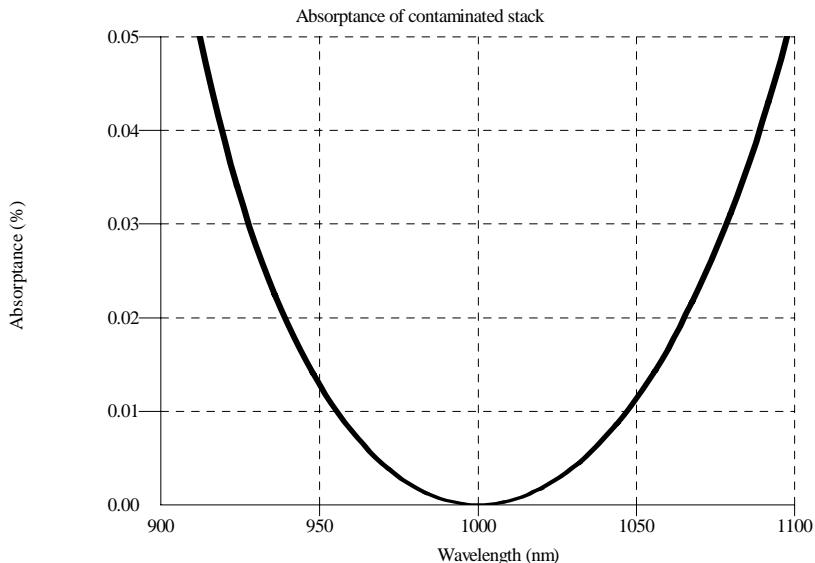


Figure 10.17. Absorptance of the quarter-wave stack with contamination layer as a function of wavelength.

We take a quarter-wave stack of silica and titania and calculate the absorptance as a function of the (odd) number of layers assuming titania outermost. The result is shown as the dashed line in figure 10.16. The results were also calculated using the full matrix theory. Agreement is excellent up to around 15 layers and then the full calculation shows a levelling off. The effect is due to the failure of the thin-layer approximation. The admittance locus of the very thin contamination layer is shifted to the extreme right and now, even though it is exceedingly thin, it swings round towards the imaginary axis. The potential absorptance rises and, when multiplied by the decreasing $(1 - R)$ factor, a constant is obtained. This constant level is very small, less than ten parts per billion. Equation (10.8) shows that for a quarter-wave stack terminated by a low-admittance layer, where Y would be very small, that the limiting absorptance would be 0.0116 or 1.16%. Accurate calculation confirms this.

As the wavelength changes, however, the admittance locus for the quarter-wave stack begins to unwind. The major effect is that the value of $\text{Re}(Y)$ decreases. This is accompanied by a slight decrease also in reflectance. The result is a considerable increase in the level of absorption associated with the contamination layer. Figure 10.17 shows the rapid increase in absorptance up to 500 parts per million from the less than ten parts per billion at the centre wavelength.

Thermally evaporated coatings are known to be affected by moisture. The

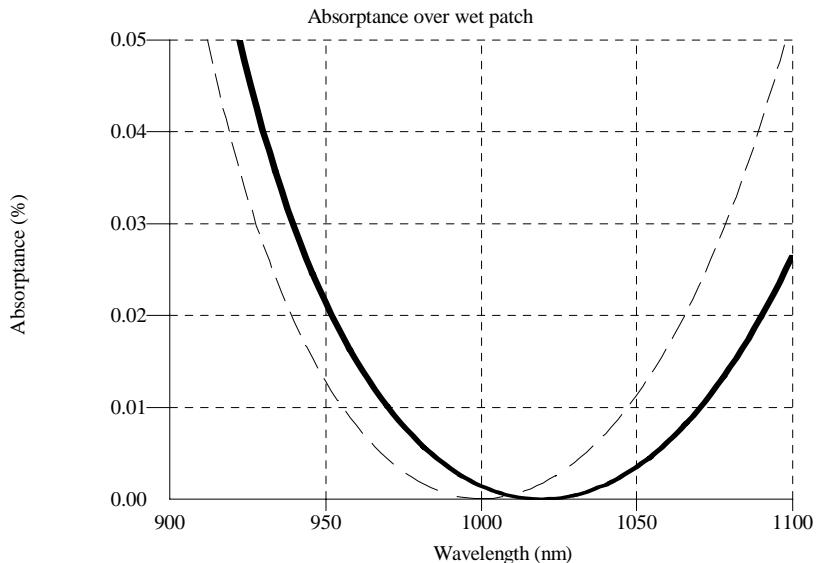


Figure 10.18. The bold line shows absorptance of a contamination layer over a wet patch in a quarter-wave stack. The dashed line shows the absorptance when deposited over a dry area.

moisture enters in localised spots and spreads out in the form of circular patches of increasing diameter. This changes the field distribution in a coating and therefore alters the absorptance associated with a contamination layer (figure 10.18).

Monitoring errors that have no perceptible effect on the reflectance of a quarter-wave stack can have quite major effects on the sensitivity to contamination.

Some additional information on contamination sensitivity at interfaces within the coating are included in the article by Macleod and Clark [40].

References

- [1] Heitmann W 1968 The influence of various parameters on the refractive index of evaporated dielectric thin films *Appl. Opt.* **7** 1541–3
- [2] Pearson J M 1970 Electron microscopy of multilayer thin films *Thin Solid Films* **6** 349–58
- [3] Lissberger P H and Pearson J M 1976 The performance and structural properties of multilayer optical filters *Thin Solid Films* **34** 349–55
- [4] Pulker H K and Jung E 1971 Correlation between film structure and sorption behaviour of vapour deposited ZnS, cryolite and MgF₂ films *Thin Solid Films* **9** 57–66

- [5] Pulker H K and Guenther K H 1972 Electron optical investigation of cross-sectional structure of vacuum-deposited multilayer systems *Vakuum-Technik* **21** 201–7
- [6] Movchan B A and Demchishin A V 1969 Study of the structure and properties of thick vacuum condensates of nickel, titanium, tungsten, aluminium oxide and zirconium dioxide *Fiz Metal Metalloved* **28** 653–60
- [7] Thornton J A 1974 Influence of apparatus geometry and deposition conditions on the structure and topography of thick sputtered coatings *J. Vacuum Sci. Technol.* **11** 666–70
- [8] Thornton J A 1986 The microstructure of sputter-deposited coatings *J. Vacuum Sci. Technol. A* **4** 3059–65
- [9] Reid I M, Macleod H A, Henderson E and Carter M J 1979 The ion plating of optical thin films for the infrared *Proc. Int. Conf. on Ion Plating and Allied Techniques (IPAT 79) (London, July 1979)* (Edinburgh: CEP Consultants) pp 55–62
- [10] Kinoshita K and Nishibori M 1969 Porosity of MgF₂ films—evaluation based on changes in refractive index due to adsorption of vapors *J. Vacuum Sci. Technol.* **6** 730–3
- [11] Harris M, Macleod H A, Ogura S, Pelletier E and Vidal B 1979 The relationship between optical inhomogeneity and film structure *Thin Solid Films* **57** 173–8
- [12] Müller K-H 1986 Model for ion-assisted thin-film densification *J. Appl. Phys.* **59** 2803–7
- [13] Müller K-H 1988 Models for microstructure evolution during optical thin film growth *Proc. Soc. Photo-Opt. Instrumentation Eng.* **821** 36–44
- [14] Sargent R B 1990 Effects of surface diffusion on thin-film morphology: a computer study *Proc. Soc. Photo-Opt. Instrumentation Eng.* **1324** 13–31
- [15] Sargent R B 1989 Surface diffusion: a computer study of its effects on thin film morphology *PhD Dissertation* (University of Arizona)
- [16] Placido F 1997 *RF Sputtering of Aluminium Oxynitride Rugates. Micrographs of Rugate Structures* Private communication (Department of Physics, University of Paisley)
- [17] Lingg L J 1990 Lanthanide trifluoride thin films: structure, composition and optical properties *PhD Dissertation* (University of Arizona)
- [18] Messier R, Takamori T and Roy R 1975 Observations on the ‘explosive’ crystallisation of non-crystalline Ge *Solid State Commun.* **16** 311–14
- [19] Klinger R E and Carniglia C K 1985 Optical and crystalline inhomogeneity in evaporated zirconia films *Appl. Opt.* **24** 3184–7
- [20] Targove J D 1987 The ion-assisted deposition of optical thin films *PhD Dissertation* (University of Arizona)
- [21] Boulesteix C and Lottiaux M 1987 *Behavior of Zirconia Film in Electron Microscope* Private communication (University of Aix-Marseille III, Marseille, France)
- [22] Ogura S 1975 Some features of the behaviour of optical thin films *PhD Thesis* (Newcastle upon Tyne Polytechnic)
- [23] Ogura S and Macleod H A 1976 Water sorption phenomena in optical thin films *Thin Solid Films* **34** 371–5
- [24] Schildknecht J, Steudel A and Walther H 1967 The variation of the transmission wavelength of interference filters by the influence of water vapour *J. Phys.* **28** C2/276–9
- [25] Koch H 1965 Optische Untersuchungen zur Wasserdampfsorption in Aufdampfschichten (insbesondere in MgF₂ Schichten) *Phys. Status Solidi* **12** 533–43
- [26] Koch H 1967 Über Sorptionsvorgänge beim Belüften von MgF₂ Schichten *Proc.*

- Coll. on Thin Films (Budapest, 1965)* (Budapest: Verlag: Kultura) pp 199–203
- [27] Macleod H A and Richmond D 1976 Moisture penetration patterns in thin films *Thin Solid Films* **37** 163–9
 - [28] Richmond D 1976 Thin film narrow band optical filters *PhD Thesis* (Newcastle upon Tyne Polytechnic)
 - [29] Lee C C 1983 Moisture adsorption and optical instability in thin film coatings *PhD Dissertation* (University of Arizona)
 - [30] Pulker H K 1982 Stress, adherence, hardness and density of optical thin films *Proc. Soc. Photo-Opt. Instrumentation Eng.* **325** 84–92
 - [31] Seeley J S, Hunneman R and Whatley A 1981 Far infrared filters for the Galileo-Jupiter and other missions *Appl. Opt.* **20** 31–9
 - [32] Roche P, Bertrand L and Pelletier E 1974 Influence of temperature on the optical properties of narrowband optical filters *Opt. Acta* **21** 927–46
 - [33] Takashashi H 1995 Temperature stability of thin-film narrow-band pass filters produced by ion-assisted deposition *Appl. Opt.* **34** 667–75
 - [34] Duparré A and Kassam S 1993 Relation between light scattering and microstructure of optical thin films *Appl. Opt.* **32** 5475–80
 - [35] Duparré A 1995 Light scattering of thin dielectric films *Handbook of Optical Properties. Volume 1. Thin Films for Optical Coatings* ed R E Hummel and K H Guenther (Boca Raton: CRC) pp 273–303
 - [36] Duparré A and Kaiser N 1998 AFM helps engineer low-scatter films *Laser Focus World* (Tulsa, OK: PennWell) pp 147–52
 - [37] Amra C 1993 From light scattering to the microstructure of thin-film multilayers *Appl. Opt.* **32** 5481–91
 - [38] Amra C 1995 Introduction to light scattering in multilayer optics *Thin Films for Optical Systems* ed F R Flory (New York: Marcel Dekker) pp 367–91
 - [39] Koslowski M 1995 Damage-resistant laser coatings *Thin Films for Optical Systems* ed F R Flory (New York: Marcel Dekker) pp 521–49
 - [40] Macleod A and Clark C 1997 How sensitive are coatings to contamination? *11th International Conference on Vacuum Web Coatings (Miami, FL)* (New Jersey: Bakish Materials Corporation) pp 176–86
 - [41] Macleod H A and Clark C 1997 Electric field distribution as a tool in optical coating design *40th Annual Technical Conf. Proc. (New Orleans)* (Society of Vacuum Coaters) pp 221–6
 - [42] Palik E D (ed) 1985 *Handbook of Optical Constants of Solids* (San Diego: Academic)
 - [43] Palik E D 1991 *Handbook of Optical Constants of Solids II* (San Diego: Academic)

Chapter 11

Layer uniformity and thickness monitoring

In the previous chapter we considered what is probably the most difficult aspect of thin-film coating and filter production, that of materials. As we saw, these are not always satisfactory, and there are still problems associated with their stability. Once the materials have been chosen, and their properties are known, the thin-film designer, using the methods discussed in chapters 3–7, can usually produce a design to meet a given specification. Given suitable materials and an acceptable design, however, there are still further difficulties to be overcome in the construction of a practical filter. The two most important remaining factors are, first, controlling the uniformity of layer thickness over the area of the substrate, and second, controlling the overall thickness of each layer. Lack of uniformity causes a shift of characteristic wavelength over the surface of the filter, without necessarily affecting the performance in other ways, while thickness errors usually cause a reduction in performance. The magnitude of the errors which can be tolerated will vary from one design to another and the estimation of this is dealt with briefly. The bulk of this chapter is concerned with the general problem of minimising these two sources of error. One other important topic is substrate preparation, and that is considered on pages 497–9.

11.1 Uniformity

In the evaporation process, it is usual to maintain the pressure within the chamber sufficiently low to ensure that the molecules in the stream of evaporant will travel in straight lines until they collide with a surface. In order to calculate the thickness distribution in a plant, the assumption is usually made that every molecule of evaporant sticks where it lands. This assumption is not strictly correct, but it does allow uniformity calculations that are sufficiently accurate for most purposes. The distribution of thickness is then calculated in exactly the same way as intensity of illumination in an optical calculation. All that is required to enable the thickness to be estimated is a knowledge of the distribution of evaporant from the source.

Holland and Steckelmacher [1] published an early and detailed account of techniques for the prediction of layer thickness and uniformity and established the theory that is essentially that still used in uniformity predictions. Their expressions were later extended by Berndt [2]. Holland and Steckelmacher divided sources into two broad types: those which have even distribution in all directions and can be likened to a point source, and those which have a distribution similar to that from a flat surface, the intensity falling off as the cosine of the angle between the direction concerned and the normal to the surface. The expressions for the distribution of material emitted from the two types of source are as follows.

For the point source:

$$dM = [m/(4\pi)]d\omega$$

and for the directed surface source:

$$dM = [m/\pi] \cos \varphi d\omega$$

where m is the total mass of material emitted from the source in all directions and dM is the amount passing through solid angle $d\omega$ (at angle φ to the normal to the surface in the case of the second type of source).

If the material is being deposited on a surface element dS of the substrate which has its normal at angle ϑ to the direction of the source from the element, then the amount which will condense on the surface will be given by:

for the point source:

$$dM = \left(\frac{m}{4\pi}\right) \left(\frac{\cos \vartheta}{r^2}\right) dS$$

and for the directed surface source:

$$dM = \left(\frac{m}{\pi}\right) \left(\frac{\cos \varphi \cos \vartheta}{r^2}\right) dS.$$

In order to estimate the thickness of the deposit we need to know the density of the film. If this is denoted by μ then the thickness will be:

for the point source:

$$dM = \left(\frac{m}{4\pi\mu}\right) \left(\frac{\cos \vartheta}{r^2}\right)$$

and for the directed surface source:

$$t = \left(\frac{m}{\pi\mu}\right) \left(\frac{\cos \varphi \cos \vartheta}{r^2}\right).$$

These are the basic equations used by Holland and Steckelmacher for estimating the thickness in uniformity calculations.

11.1.1 Flat plate

The simplest case is that of a flat plate held directly above and parallel to the source. Here the angle φ is equal to the angle ϑ and the thickness is as follows.

For the point source:

$$t = \left(\frac{m}{4\pi\mu} \right) \left(\frac{\cos \vartheta}{r^2} \right) = \frac{mh}{4\pi\mu (h^2 + \rho^2)^{3/2}}$$

and for the directed surface source:

$$t = \left(\frac{m}{\pi\mu} \right) \left(\frac{\cos^2 \vartheta}{r^2} \right) = \frac{mh^2}{\pi\mu (h^2 + \rho^2)^2}$$

with notation as in figure 11.1. These expressions simplify to:

for the point source:

$$t/t_0 = [1 + (\rho/h)^2]^{-3/2}$$

and for the directed surface source:

$$t/t_0 = [1 + (\rho/h)^2]^{-2}$$

and are plotted in figure 11.2. t_0 is the thickness immediately above the source where $\rho = 0$. In neither case is the uniformity at all good. Clearly the geometry is not suitable for any very accurate work unless the substrate is extremely small and in the centre of the plant.

11.1.2 Spherical surface

A slightly better arrangement that can sometimes be used is a spherical geometry where the substrates lie on the surface of a sphere. A point source will give uniform thickness of deposit on the inside surface of a sphere when the source is situated at the centre. It can be shown that the directed surface source will give uniform distribution similarly when it is itself made part of the surface. In fact, it was the evenness of the coating within a sphere which led Knudsen [3] first to propose the cosine law for thin-film deposition. The method is often used in plants for simple blooming of components such as lenses where the uniformity need not be better than, say, 10% of the layer thickness at the centre of the component. However, for precise work, this uniformity is still not adequate.

A higher degree of uniformity involves rotation of the substrate carrier, which we shall now consider.

11.1.3 Rotating substrates

The situation here is as if, in figure 11.1, the surface for coating were rotated about a normal at distance R away from the source. As the surface rotates, the

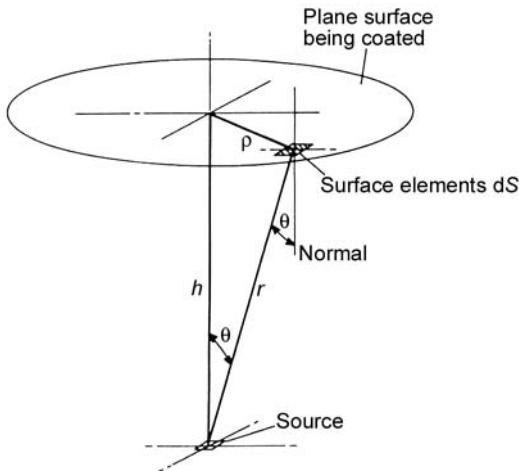


Figure 11.1. Diagram showing the geometry of the evaporation from a central source on to a parallel plane surface.

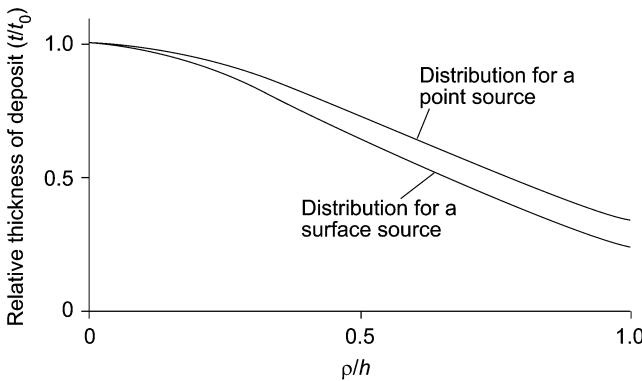


Figure 11.2. Film thickness distribution on a stationary substrate from a central source.

thickness deposited at any point will be equal to the average of the thickness which would be deposited on a stationary substrate around a ring centred on the axis of rotation, provided always that the number of revolutions during the deposition is sufficiently great to make the amount deposited in an incomplete revolution a very small proportion of the total thickness. By choosing the correct distance between source and axis of rotation, the uniformity can be made vastly superior to that for stationary substrates.

We shall consider first the directed surface source. Figure 11.3 shows the situation. The calculation is basically similar to that for the flat plate with a central

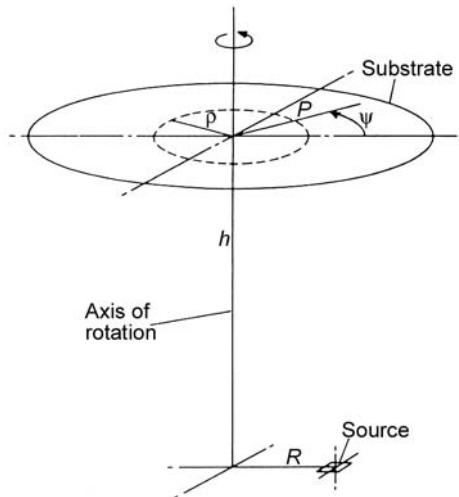


Figure 11.3. Diagram showing the geometry of the evaporation from a stationary offset source onto a rotating substrate.

source. Here we stop the plate and calculate the mean thickness around the circle containing the point in question and centred on the axis of rotation. The radius of the circle is ρ , and if we define any point P on the circle by the angle ψ , then the thickness at the point is given by

$$t = \left(\frac{m}{\pi \mu} \right) \left(\frac{h^2}{(h^2 + \rho^2 + R^2 - 2\rho R \cos \psi)^2} \right)$$

where r , the distance from the source to the point, is given by

$$r^2 = h^2 + \rho^2 + R^2 - 2\rho R \cos \psi.$$

Then, taking the mean of the thickness around the circle, we have for the thickness of the deposit in the rotating case

$$t = \left(\frac{m}{\pi \mu} \right) \left(\frac{1}{2\pi} \right) \int_0^{2\pi} \frac{h^2 d\psi}{(h^2 + \rho^2 + R^2 - 2\rho R \cos \psi)^2}.$$

Now the integral $\int_0^{2\pi} d\psi / (1 - a \cos \psi)^2$ can be evaluated by contour integration giving

$$t = \left(\frac{m}{\pi \mu} \right) \left(\frac{h^2}{(h^2 + \rho^2 + R^2)^2} \right) \left(\frac{1}{\{1 - [2\rho R / (h^2 + \rho^2 + R^2)]^2\}^{3/2}} \right)$$

$$\frac{t}{t_0} = [(1 + R^2/h^2)^2 (1 + \rho^2/h^2 + R^2/h^2)]$$

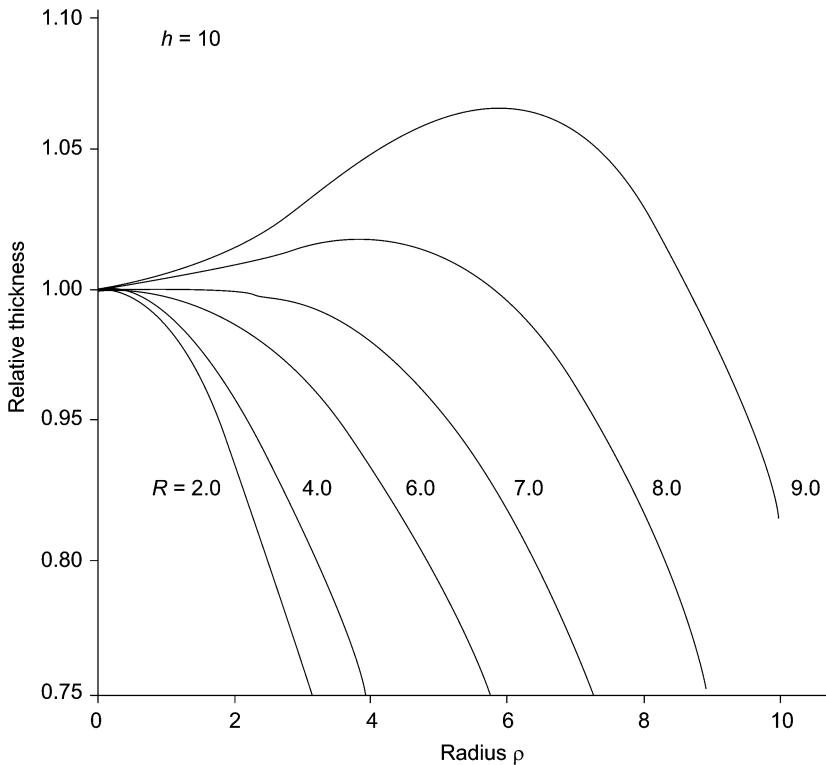


Figure 11.4. Theoretical film thickness distribution on substrates rotated about the centre of the plant for various source radii and substrate heights. The sources are assumed to be small directed surfaces parallel to the substrates.

$$\times \{[1 + \rho^2/h^2 + R^2/h^2 - 2(\rho/h)(R/h)]^{3/2} \\ \times [1 + \rho^2/h^2 + R^2/h^2 + 2(\rho/h)(R/h)]^{3/2}\}^{-1}$$

where t/t_0 is, as before, the ratio of the thickness at the radius in question to that at the centre of the substrate holder.

Figure 11.4 shows this function plotted for several different dimensions which are typical of medium-sized coating plants. The distribution can immediately be seen to be vastly superior to that when the substrates are stationary. For one particular combination of dimensions, that corresponding to $R = 7$, the distribution is extremely even over the central part (radius 3.75) of the plant. This is the arrangement used in the production of narrowband filters where the uniformity must necessarily be very good. If the uniformity is not quite so important, where rather broader filters or perhaps antireflection coatings are concerned, then the sources can be moved outwards, allowing a larger area to be coated at the expense of a slight decline in uniformity.

A similar expression is found for a point source but this time involving elliptic integrals. The thickness at the point P , assuming that the substrate does not rotate, is given by

$$t = \left(\frac{m}{4\pi\mu} \right) \left(\frac{h}{(h^2 + \rho^2 + R^2 - 2\rho R \cos \psi)^{3/2}} \right)$$

and in the presence of rotation the thickness at any point around the ring of radius ρ will be the mean of the expression, i.e.

$$\begin{aligned} t &= \left(\frac{m}{4\pi\mu} \right) \left(\frac{1}{2\pi} \right) \int_0^{2\pi} \frac{h d\psi}{(h^2 + \rho^2 + R^2 - 2\rho R \cos \psi)^{3/2}} \\ t &= \frac{m}{4\pi^2\mu} \int_0^\pi \frac{h d\psi}{(h^2 + \rho^2 + R^2 - 2\rho R \cos \psi)^{3/2}}. \end{aligned}$$

Now let $(\pi - \psi)/2 = \gamma$, then $d\psi = -2d\gamma$, and the expression for thickness becomes

$$t = \frac{m}{4\pi^2\mu} \int_{\pi/2}^0 \frac{-h d\gamma}{[h^2 + (R + \rho)^2 - 4\rho R \sin^2 \gamma]^{3/2}}$$

which can be written

$$\begin{aligned} t &= \left(\frac{m}{4\pi^2\mu} \right) \left(\frac{h}{[h^2 + (R + \rho)^2]^{3/2}} \right) \\ &\quad \times \int_0^{\pi/2} \frac{d\gamma}{\{1 - [4\rho R / (h^2 + (R + \rho)^2)] \sin^2 \gamma\}^{3/2}}. \end{aligned}$$

Now the integral in this expression is a standard form

$$\frac{1}{(1 - k^2)} E(k, \alpha) = \int_0^\alpha \frac{d\gamma}{(1 - k^2 \sin^2 \gamma)^{3/2}}$$

where $E(k, \alpha)$ is an elliptic integral of the second kind, and is a tabulated function [4]. The expression for thickness then becomes:

$$T = \left(\frac{hm}{4\pi^2\mu} \right) \left(\frac{E(k, \pi/2)}{[h^2 + (R + \rho)^2]^{1/2} [h^2 + (R - \rho)^2]} \right)$$

where

$$k = 4\rho R / \left[h^2 + (R + \rho)^2 \right].$$

Curves of this expression are given by Holland and Steckelmacher [1], and the shape is very similar to that for the directed surface source.

Almost all the sources used in the production of thin-film filters, especially the boat type, give distributions similar to the directed surface source. Holland and Steckelmacher also describe some experiments which they carried out to determine this point. Keay and Lissberger [5] have studied the distribution from a howitzer source loaded with zinc sulphide, and it appears that this is somewhere in between the point source and the directed surface source, probably due to scattering in the evaporant stream immediately above the heater where the pressure is high. The cloud of vapour that forms seems to act to some extent as a secondary point source. This behaviour of the howitzer probably depends to a considerable extent on the material which is being evaporated. Graper [6] has studied the distribution of evaporant from an electron gun and has found that this is somewhat more directional than the directed surface source. Its distribution can be described by a $\cos^x \vartheta$ law where x is somewhere between 1 and 3, and depends on the power input and on the amount of material in the hearth. Using zinc sulphide and cryolite, Richmond [7] found that the distribution from an electron gun source was best represented by a law of the form $\cos \vartheta$.

Normally, in calculating the distribution to be expected from a particular geometry, we assume that we are using directed surface sources, and then, when setting up a plant for the first time, the sources are placed at the theoretically best positions. The first few runs soon show whether or not any further adjustments are necessary, and if they are, they are usually very slight and can be made by trial and error. Once the best positions are found, it is important to ensure that the sources are always accurately set to reproduce them. Care should be taken to make sure that the angular alignment is correct. A source at the correct geometrical position but tilted away from the correct direction will give uniformity errors just as much as if it were laterally displaced. The frontispiece shows a plant that is being fitted with a flat plate work holder for the manufacture of narrowband filters.

Where uniformity must be good over as large an area as possible but where the ultimate is not required, it is possible to use a combination of a spherical surface and rotating plate. A domed work holder, or calotte, is rotated about its centre with the sources offset beneath it so that they are approximately on the surface of the sphere, with slight adjustments made during setting up. This gives very good results over a much larger area than would be possible with the simple rotating flat plate. Figure 11.5 shows the interior of a machine that uses this arrangement.

When still improved uniformity is required, it is possible to achieve it by what is known as a planetary jig. In this arrangement, the substrates not only rotate about the centre of the jig, but also about their own individual centres at much greater speed, so that they execute many revolutions for each single revolution of the jig as a whole. This carries a stage further the averaging process that occurs with the simple rotating jig.



Figure 11.5. Photograph showing the interior of a machine with a domed calotte.

11.1.3.1 Use of masks

It is possible to make corrections to distribution by careful use of masks. In their simplest form they are stationary and are placed just in front of the substrates that rotate on a single carrier about a single axis. The masks are cut so that they modify the radial distribution of thickness. Theoretical calculations give dimensions for masks of approximately the correct shape, which can then be trimmed according to experimental results to arrive at the final form. For a number of reasons, it is normal to leave the central monitor glass uncorrected. It is difficult to correct the central part of the chamber where the mask width tends to zero, and, in any case, the monitor is usually stationary. Furthermore, in some monitoring arrangements, there is an advantage in having more material on the monitor than on the batch.

A further degree of freedom was introduced by Ramsay *et al* [8] in the form of a rotating mask. For a large flat substrate which is approaching the dimensions of the plant there is little other than simple rotation that can be done, in terms

of the carrier jig, to improve uniformity. Planetary arrangements require much more room. Stationary masks are of some help but they are somewhat sensitive to the characteristic of the sources and are not therefore sufficiently stable for a very high degree of uniformity. A much more stable arrangement, that has been shown capable of uniformities of the order of 0.1% over areas of around 200 mm diameter, involves rotating the mask about a vertical axis at a rotational speed considerably in excess of that of the substrate carrier. This effectively corrects the angular distribution of the source that can be positioned at the centre of the plant. The mask rotation axis is usually placed very near the source and positioned so that the line drawn from the source through the mask centre intersects the perimeter of the substrate carrier. In practice the axis of rotation and the rotating shutter are close to the source position and slight adjustment of the axis can be made for trimming purposes. It has been found to be an exceptionally stable arrangement.

11.2 Substrate preparation

Before a substrate can be coated, it must be cleaned. The forces which hold films together and to the substrate are all short-range interatomic and intermolecular forces. These forces are extremely powerful, but their short range means that we can think of each atomic layer as being bound to the neighbouring layers only, and being little affected by material which is further removed from it. Thus, the adhesion of a thin film to the substrate depends critically on conditions at the substrate surface. Even a monomolecular layer of a contaminant on the surface can change the force of adhesion by orders of magnitude. Condensation of evaporant, too, is just as sensitive to surface conditions that can alter completely the characteristics of the subsequent layers. Substrate cleaning so that the condensing material attaches itself to the substrate and not an intervening layer of contaminant is therefore of paramount importance.

The typical symptoms of an inadequately cleaned substrate are a mottled, oily appearance of the coating, coupled usually with poor adhesion and optical performance. This can be caused also by such defects in the plant as backstreaming of oil from the pumps. When these symptoms appear it is usually advisable to extend any subsequent improvements in cleaning techniques to the plant as well.

A good account of various cleaning methods is given by Holland [9]. A more recent account is that of Mattox [10]. The best cleaning process will depend very much on the nature of the contamination that must be removed and, although it may seem self-evident, in all cleaning operations it is essential to avoid contaminating the surface rather than cleaning it. For laboratory work, when the substrates are reasonably clean to start with (microscope slide glass is usually in this condition), then for most purposes it will be found sufficient to wash the substrates thoroughly in detergent and warm water (not household detergent that

sometimes has additives which cause smears to appear on the finished films), to rinse them thoroughly in running warm water (in areas where tap water is fairly pure, hot tap water will often be found adequate), and then to dry them thoroughly and immediately with a clean towel or soft paper tissue, or, better still, to blow them dry with a jet of clean dry nitrogen. The substrates should never be allowed to dry themselves or stains will certainly occur which are usually impossible to remove. Substrates should be handled as little as possible after cleaning and, since they never remain clean for long, placed immediately in the coating plant and the coating operation started. Wax or grease will probably require treatment with an alcohol such as isopropyl, perhaps rubbing the surface with a clean fresh cotton swab soaked in the alcohol and then flooding the surface with the liquid. Care must be taken to ensure that the alcohol is really clean. A bottle of alcohol available to all in a laboratory seldom remains clean for long and a better arrangement is to keep it under lock and key and to allow the alcohol into the laboratory in wash bottles that emit the alcohol when squeezed.

This basic cleaning procedure can be modified and supplemented in various ways, especially if large numbers of substrates are to be handled automatically. Ultrasonic scrubbing in detergent solution or in alcohol is a very useful technique, although prolonged ultrasonic exposure is to be avoided since it can eventually cause surface damage. It is important that the substrates should be kept wet right through the cleaning procedure until they are dried as the final stage. Vapour cleaning is frequently used for this. The substrates are exposed to the vapour of alcohol or other degreasing agents so that initially it condenses and runs off, taking any residual contamination or the remains of the agent from the previous cleaning stage with it. The substrates gradually reach the temperature of the vapour and then no further condensation takes place, when the substrates can be withdrawn perfectly dry. Since the agent is condensing from the vapour phase, it is in an extremely pure form. An alternative end to the cleaning process is a rinse in deionised water followed by drying in a blast of dry, filtered nitrogen.

It is very difficult to see marks on the surface of the substrate with the naked eye. Dust can be picked up by oblique illumination, but wax and grease cannot.

An old and common test for assessing the quality of a cleaning process is to breathe on one of the substrates so that moisture condenses on it in a thin layer. This tends to magnify the effects of any residue. The moisture acts in almost exactly the same way as a condensing film since the condensation pattern depends on the surface conditions. A surface examined in this way is said to exhibit a good or bad 'breath figure'. A contaminated surface gives a smeared pattern, while a clean surface is completely even. Since even this step can introduce slight residual contamination, it is better used only on a sample as an indication of the condition of the batch.

Once the substrates are in the chamber, and they should always be loaded as soon as possible after cleaning, they can be given a final clean by a glow discharge. The equipment for this, which consists of a high-voltage supply, preferably DC, together with the necessary lead-in electrodes, is fitted as standard in most

plants. At a suitable pressure, which will vary with the particular geometry of the electrodes but which will usually be around 0.06 mb, a glow discharge is struck and, provided the geometry is correct, the surface of the substrates is bombarded with positive ions. This effectively removes any light residual contamination, although gross contamination will persist. It is not certain whether the cleaning action actually arises from a form of sputtering or whether the glow discharge is merely a convenient way of raising the temperature of the surfaces so that contaminants are baked off. Generally the glow discharge is limited in duration to five or perhaps ten minutes. It has been suggested that, although glow discharge cleaning does remove grease, it does encourage dust particles; for coatings where minimum dust is required, such as high-performance laser mirrors, glow discharge cleaning is frequently omitted. Lee [11] found that the omission of glow discharge cleaning led to a very great increase in the incidence of moisture penetration patches in his films and consequently to a fall in the performance of his filters.

The evaporation of the first layer should begin as soon as possible after the glow discharge has stopped. Cox and Hass [12] used a discharge current of 80 mA and a voltage of 5000 V for 5 min to clean substrates before coating them with zinc sulphide, and found that the time between finishing the discharge and starting the evaporation should be not greater than three minutes. If the time was allowed to exceed five minutes, then the quality of the films, especially their adhesion, deteriorated.

If, as sometimes happens, a filter is left for a period, say overnight, in an uncompleted state, it will often be found advisable to carry out a short period of glow discharge cleaning before starting to evaporate the remaining layers.

11.3 Thickness monitoring

Given suitable materials, clean substrates, and a machine with substrate-holder geometry to give the required distribution accuracy, the main problem which remains is that of controlling the deposition of the layers so that they have the characteristics required by the coating or filter design. Of course, many properties are required, but refractive index and optical thickness are the most important. There is no satisfactory way, at present, of measuring the refractive index of that portion of a film which is actually being deposited. Such measurements can be made later but for closed loop control, dynamic measurements are required. Normal practice, therefore, is simply to control, as far as possible, those deposition parameters that would affect refractive index so that the index produced for any given material is consistent. Measurements are made of the index and the value usually obtained is used in the coating design. This procedure, while it usually gives satisfactory results, is far from ideal and is used simply because, at the present time, there is no better way.

Film thickness can more readily be measured and, therefore, controlled.

The simplest systems display a signal to a plant operator who is responsible for interpreting it and assessing the correct instant to terminate deposition. At the other end of the scale, there are completely automatic systems in which operator judgement plays no part and in which even operator intervention is rarely required.

There are many ways in which the thickness can be measured. All that is necessary is to find a parameter that varies in a suitable fashion with thickness and to devise a way of monitoring this parameter during deposition. Thus, parameters such as mass, electrical resistance, optical density, reflectance and transmittance have all been used. Of all the methods, those most frequently used involve either optical measurements of reflectance or transmittance or the measurement of total deposited mass by the quartz-crystal microbalance.

The question of the best method for the monitoring of thin films is, of course, inseparable from that of how accurately the layers must be controlled. This second question is a surprisingly difficult one to answer. Indeed, it is impossible to separate the two questions: the tolerances which can be allowed and the method used for monitoring are closely related and one cannot be considered in depth independently of the other.

For convenience, however, we will consider some of the more common arrangements for monitoring, including only the most rudimentary ideas of accuracy and then, at a later stage, consider the question of tolerances along with some of the more advanced ideas of monitoring and its various classifications.

11.3.1 Optical monitoring techniques

Optical monitoring systems consists of some sort of light source illuminating a test substrate which may or may not be one of the filters in the batch, and a detector analysing the reflected or transmitted light. From the results of that analysis, the evaporation of the layer is stopped as far as possible at the correct point. Usually, so that the layer may be stopped as sharply as possible, the plant is fitted with a shutter which can be inserted in front of the evaporation sources. This is a much more satisfactory method than merely turning off the supply to the boats, which always take a finite time to stop emitting. Such a shutter can be seen in figure 9.4.

Almost all the early workers in the field used the eye as the detector, and the thicknesses of the films were determined by assessing their colour appearance in white light. In many cases they were concerned with simple single-layer coatings such as single-layer blooming, which are not at all susceptible to errors. When the blooming layer is of the correct thickness for visible light, the colour reflected from the surface in white light has a magenta tint, owing to the reduction of the reflectance in the green. The visual method is quite adequate for this purpose and is still being widely used. A very clear account of the method is given by Mary Banning [13], who compiled table 11.1.

In the production of other types of filter where the errors of the visual method

Table 11.1. (After Banning [13].)

ZnS	Colour change for Na_3AlF_6	Optical thickness for green light
Bluish white	Yellow	
↓	↓	
White	Magenta	$\lambda/4$, first-order maximum
↓	↓	
Yellow	Blue	
↓	↓	
Magenta	White	$\lambda/2$, first-order minimum
↓	↓	
Blue	Yellow	
↓	↓	
Greenish white	Magenta	$3\lambda/4$, second-order maximum
↓	↓	
Yellow	Blue	
↓	↓	
Magenta	Greenish white	λ , second-order minimum
↓	↓	
Blue	Yellow	
↓	↓	
Green	Magenta	$5\lambda/4$, third-order maximum

would be too large, other methods must be used. An early paper by Polster [14] describes a photoelectric method which is basically the same as that used most often today. We saw in chapter 2 that if the film is without absorption, then its reflectance and transmittance measured at any one wavelength will vary with thickness in a cyclic manner, similar to a sine wave, although, for the higher indices, the waves will be flattened at their tops. The turning values correspond to those wavelengths for which the optical thickness of the film is an integral number of quarter wavelengths, the reflectance being equal to that of the substrate when the number is even and a maximum amount removed from the reflectance of the substrate when the number is odd. Figure 11.6 illustrates the behaviour of films of different values of refractive index. This affords the means for measurement. If the detector in the system is made highly selective, for example by putting a narrow filter in front of it, then the measured reflectance or transmittance will vary in this cyclic way, and the film may be monitored to an integral number of quarter-waves by counting the number of turning points passed through in the course of the deposition. A typical arrangement to perform this operation is shown in figure 11.7. The filter may be an interference filter or, more flexible, an adjustable prism or grating monochromator.

Consider the deposition of a high-reflectance multilayer stack where all the

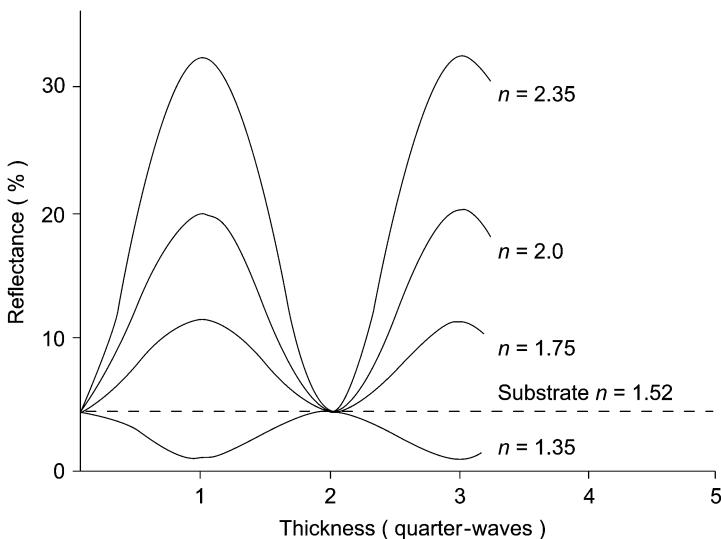


Figure 11.6. Curves showing the variation with thickness of the reflectance of several films with different refractive indices.

layers are quarter-waves. Let the monitoring wavelength be the wavelength for which all the layers are one quarter-wavelength thick. The reflectance of the test piece will vary as shown in figure 11.8 [15]. The example shown is typical of a reflecting stack for the visible region. The reflectance can be seen to increase during the deposition of the first layer, which is of high index, to a maximum where the deposition is terminated. During the second layer the reflectance falls to a minimum where the second layer is terminated. The third layer increases the reflectance once again and the fourth layer reduces it. This behaviour is superimposed on a trend towards a reflectance of unity so that the variable part of the signal becomes a gradually smaller part of the total. This puts a limit on the number of layers which can be monitored in reflectance in this way to around four, when a fresh monitoring substrate must be inserted. In transmission monitoring, this effect does not exist and the variable part of the signal remains a sufficiently large part of the whole. The only problem is that the overall trend of the signal is towards zero, so that eventually it will become too small in comparison with the noise in the system. With reasonable optics and a photomultiplier detector the number of layers which may be dealt with in this way is around 21. At this stage the noise usually becomes too great.

Frequently, automatic methods of detection of the layer end point are used. Automatic methods, however, are not universally employed and machine operator control is still an important technique. For the greatest accuracy, the output of the detector should be displayed on a chart recorder making it easier to determine the turning values. With such an arrangement, a trained operator can usually

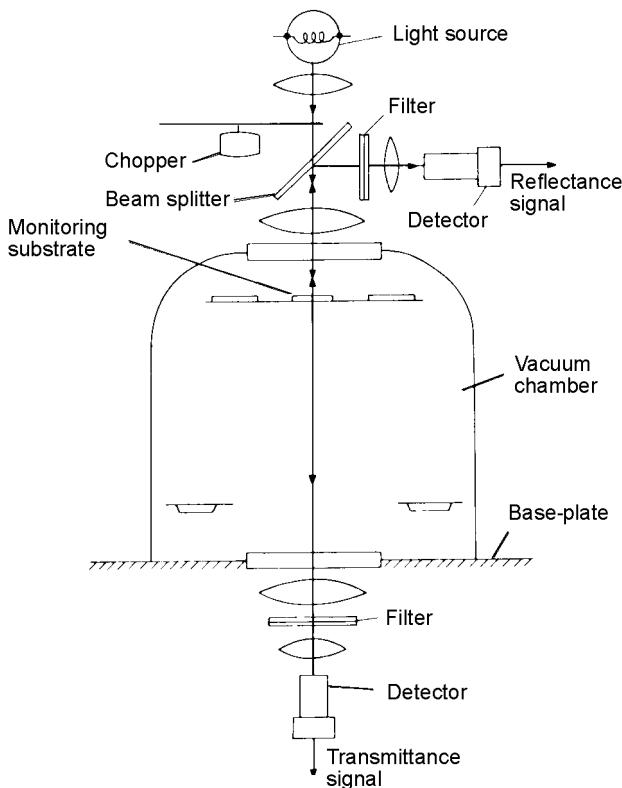


Figure 11.7. A possible arrangement of a monitoring system for reflectance and transmittance measurements.

terminate the layers to an accuracy on the monitoring substrate of around 5% or so, depending on the index of the film, although with great care and attention it may be possible to achieve nearer 2%. Of course, as we shall see, this does not necessarily mean that the actual thickness of the filters in the batch will be as accurate. Other sources of error operate to introduce differences between the monitor and the batch.

To improve the signal-to-noise ratio it is usual to chop the light before it enters the plant, partly because the evaporation process produces a great deal of light during the heating of the boats, but mainly because, at the signal levels encountered, the electronic noise without some filtering would be impossibly great. The chopper should be placed immediately after the source of light but before the plant, and the filter should be inserted after the plant. This arrangement reduces the stray light to a greater extent than would placing either the filter before the plant or the chopper after it. It is, of course, always advisable to limit as far as possible the total light incident on the detector, partly because unchopped

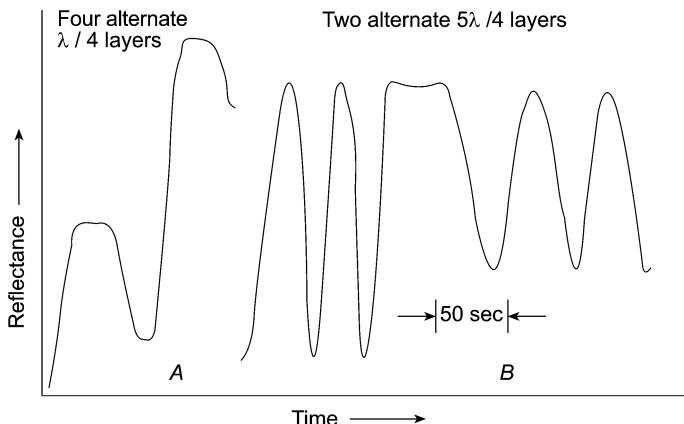


Figure 11.8. Record taken from a pen recorder of the reflectance of a monitor glass during film deposition. (After Perry [15].)

radiation can push the detector into a nonlinear region and partly because it can cause damage to the device especially if it is a photomultiplier. If a filter rather than a monochromator is used, then great care should be taken to ensure that the sidebands are particularly well suppressed. Photomultipliers and other detectors have characteristics that can vary considerably with wavelength, and if the monitoring wavelength lies in a rather insensitive region compared with the peak sensitivity, then small leaks in the more sensitive region, which might not be very noticeable in the characteristic curve of the filter, can cause considerable difficulties from stray light, even giving spurious signals of similar or greater magnitude than the true signal. Prism or grating monochromators are often safer for this work, besides being considerably more flexible.

The technique in which the layer termination is at an extremum of the signal is sometimes called turning-value monitoring. We can investigate the errors likely to arise in this type of monitoring as follows. Suppose that in the monitoring of a single quarter-wave layer there is an error γ in the value of reflectance at the termination point. This will give rise to a corresponding error φ in the phase thickness of the layer δ where

$$\delta = (\pi/2) - \varphi.$$

Because of the nature of the characteristic reflectance curve of the single layer, the error in phase thickness will be rather greater in proportion than the original error in reflectance. The admittance of the layer will be given by the characteristic matrix:

$$\begin{bmatrix} \cos \delta & (i \sin \delta)/y \\ iy \sin \delta & \cos \delta \end{bmatrix} \begin{bmatrix} 1 \\ y_m \end{bmatrix}$$

where

$$\cos \delta = \sin \varphi \quad \text{and} \quad \sin \delta = \cos \varphi.$$

This gives

$$Y = \frac{\sin \varphi + i(y_m \cos \varphi) / y}{iy \cos \varphi + y_m \sin \varphi}$$

where the symbols have their usual meaning. Introducing the approximations for $\sin \varphi$ and $\cos \varphi$ up to and including powers of the second order, we have

$$Y = \frac{\varphi + i(y_m/y)(1 - \varphi^2/2)}{iy(1 - \varphi^2/2) + y_m\varphi}$$

and the reflectance of the monitor *in vacuo* will be given by

$$R = \left| \frac{(y_m - 1)\varphi + i(y - y_m/y)(1 - \varphi^2/2)}{(y_m + 1)\varphi + i(y + y_m/y)(1 - \varphi^2/2)} \right|^2$$

which simplifies to

$$R = \frac{(y - y_m/y)^2}{(y + y_m/y)^2} \left(1 + \frac{4y_m(y_m^2 + 1 - y^2 - y_m^2/y^2)}{(y^2 - y_m^2/y^2)^2} \varphi^2 \right). \quad (11.1)$$

The values of y and φ are related as follows:

$$\gamma = \frac{4y_m(y_m^2 + 1 - y^2 - y_m^2/y^2)}{(y^2 - y_m^2/y^2)^2} \varphi^2 = \sigma \varphi^2 \quad (11.2)$$

since the first factor in equation (11.1) is just the reflectance when γ and φ are both zero.

Now, in most cases, it will not be possible to determine the reflectance at the turning value to better than 1% of the true value. In many cases, especially where there is noise, it will not be possible even to do as well as this. However, assuming this value for γ , the expression for the error in the layer thickness becomes

$$\pm 0.01 = \sigma \varphi^2$$

where the sign \pm is taken to agree with $\sigma \varphi^2$ and depends on whether or not the turning value is a maximum or a minimum. If the error is expressed in terms of a quarter-wave thickness which is equivalent to $\pi/2$ radians, the expression becomes

$$\text{Error} = \frac{\varphi}{\pi/2} = \frac{0.1}{(\pi/2) |\sigma|^{1/2}}. \quad (11.3)$$

A typical case is the monitoring of a quarter-wave of zinc sulphide on a glass substrate where $y = 2.35$ and $y_m = 1.52$. Substituting these values in expression (11.2) and using it in (11.3), the fractional error in the quarter-wave becomes 0.08. This is a colossal error compared with the original error in reflectance, and illustrates the basic lack of accuracy inherent in this method.

In the infrared, it is often possible to use wavelengths for monitoring which are shorter than the wavelengths of the desired filter peaks by a factor of perhaps two or even four. This improves the basic accuracy by the same factor. For layers similar to that considered above, the errors would then be 0.04 or 0.02. These errors are on the limit of permissible errors, and it is clear that this simple system of monitoring is not really adequate for any but the simplest of designs.

What makes the method particularly difficult to apply is that it is only the portion of the signal before the turning point that is available to the operator, who has therefore to anticipate the turning value, and the fact that trained plant operators can achieve the theoretical figures for accuracy says much for their skill.

An alternative method, inherently more accurate, involves the termination of the layer at a point remote from a turning value where the signal changes much more rapidly. This consists of the prediction of the reflectance of the monitoring substrate when the layer is of the correct thickness and then the termination of the deposition at that point. One disadvantage is that the reflectance of the monitor, or the transmittance, is not an easy quantity to measure absolutely, because of calibration drifts during the process, due partly to such causes as the gradual coating of the plant windows—almost impossible to avoid. Another is that whereas with turning value monitoring it is often possible to use just one single monitor, on which all the layers can be deposited, so that it becomes an exact replica of the other filters in the batch, in this alternative method the prediction of the reflectances used as termination values is very difficult if only one monitor is used, because small errors in early layers affect the shape of the curve for later layers.

Some of these difficulties may be avoided by using a separate monitor for each and every layer. To avoid the errors due to any shift in calibration which may occur in changing from one monitor to the next or in the coating of the plant windows, it is wise if at all possible to choose the parameters of the system so that the layer is thicker than a quarter-wave at the monitoring wavelength. This ensures that the termination point of the layer is beyond at least the first turning value, which can therefore be used as a calibration check. It will also be found necessary to set up the reflectance scale for each fresh monitoring substrate and the initial uncoated reflectance which will be known accurately can be used for this. Because a large number of monitor glasses is required, special monitor changers have been designed and are commercially available, which will accommodate stacks of 40 or so glasses. The low-index material may have rather poor contrast on the monitor substrates and a frequent variant of this method is the deposition of two layers, high index followed by low index, on each monitor substrate.

The principal objection which most workers almost instinctively feel towards this system is that no longer is the monitor an exact replica of the batch of filters. This is to some extent a valid objection. The layer which is being deposited on an otherwise uncoated substrate is condensing on top of what may be quite a different structure from the partially finished filters of the batch. Behrndt and Doughty [16] have noticed a definite measurable difference between layers which are deposited on top of an already existing structure and those deposited on fresh substrates. They compared the deposition of zinc sulphide shown by a crystal monitor (this special type of monitor will be discussed shortly), which already had a number of layers on it, with the layer going down on a fresh glass substrate, and found that the layer began to grow on the crystal immediately the source was uncovered, but that the optical monitor took some time to register any deposition. The difference could amount to several tens of nanometres before the rates became equal. This, they decided, was due to the finite time for nuclei to form on the fresh glass surface and the rather small probability of sticking of the zinc sulphide until the nuclei were well and truly formed. Once the film started to grow, all the molecules reaching the surface would stick. On the crystal where a film already existed, not necessarily of zinc sulphide, nucleation sites were already there and the film started to grow immediately. The sticking coefficient of a material on a fresh monitor surface falls with rising vapour pressure, and zinc sulphide has a particularly large vapour pressure. Similar trouble was not experienced with thorium fluoride, which has a much lower vapour pressure. Behrndt and Doughty found that the problem could be solved by providing nucleation sites on the clean monitor slides by precoating them with thorium fluoride, which has a refractive index very close to that of glass. Some 20 nm or so of thorium fluoride was found to be sufficient and did not affect the monitoring of zinc sulphide deposited on top. (Since thorium fluoride is radioactive and somewhat out of favour a different low-index fluoride would be advisable.) This effect becomes greater the greater the surface temperature of the monitor. By changing the type of evaporation source to an electron-beam unit, which produced less radiant heat for the same evaporation rate, it was found possible to operate at monitor temperatures low enough to cause the effect to disappear.

The authors also remarked on an effect which is well known in thin-film optics. Thick substrates tend to have layers condensing on them which are thicker than those on thin substrates in the same or similar positions in the plant. In the case cited by the authors, the thin substrates were around 0.040 in, while the thick ones were around half an inch thick. The difference in coating thickness was sufficient to shift the reflectance turning values by some 40–50 nm at 632.8 nm. This was shown, qualitatively, to be due to the difference in temperature between the two substrates. The thicker substrates took longer to heat up than the thin ones. The heating in this particular case was almost entirely due to radiation from the sources and, again when electron-beam sources were introduced, the effect was considerably reduced.

The accuracy of the monitoring process can be improved greatly if a system

devised by Giacomo and Jacquinot [17], and known usually as the ‘maximètre’, is employed. This involves the measurement of the derivative of the reflectance versus wavelength curve of the monitor. At points where the reflectance is a turning value, the derivative of the reflectance with respect to wavelength is zero and is rapidly changing from a positive to a negative value in the case of a maximum and vice versa in the case of a minimum. The original apparatus consisted of a monochromator with a small vibrating mirror before the slits on the exit side so that a small spectral interval was scanned sinusoidally. The output signal from the detector consisted of a steady DC component, representing the mean reflectance, or transmittance, over the interval, a component of the same frequency as the scanning mirror representing the first derivative of the reflectance against wavelength, a component of twice the scanning frequency, representing the second derivative of the reflectance, and so on. A slight complication is the variation in sensitivity of the system with wavelength that appears as a change in the reflectance signal and hence the derivative, unless it is compensated. In their arrangement, Giacomo and Jacquinot produced an intermediate image of the spectrum within the monochromator, and a razor blade positioned along it made a linear correction to the intensity over a sufficiently wide region and was found to be accurate enough. A more usual technique today would be to make a correction electronically. The accuracy claimed for this system is a few tenths of a nanometre, typically 0.2–0.3 nm, and this is certainly achieved. A problem, as we have seen in chapter 9, is that the layers are frequently insufficiently stable themselves to retain optical thicknesses to this accuracy, especially when exposed to the atmosphere.

A method, similar in some respects, but with some definite advantages in interpretation, was devised by Ring and Lissberger [18, 19]. It consists of measuring the reflectance or transmittance at two wavelengths and finding the difference. In the original system, a monochromator was used, containing a chopping system that switched the output of the monochromator from one wavelength to another and back again. The AC signal from the detector was a measure of the difference. Since the two wavelengths could be placed virtually anywhere within the region of sensitivity of the detector, the method had greater flexibility than the Giacomo and Jacquinot system. Greatest contrast in the two reflectance signals as a layer was being deposited could be obtained by placing the two wavelengths at the points of greatest opposite slope in the characteristic of the thin-film structure at the appropriate stage. When the signals at the two wavelengths were equal, the output of the system passed through a null, and, if displayed on a chart recorder, made detection of the terminal point of a particular layer, usually indicated by the null, particularly easy to detect.

More recently, the ideas inherent in these systems have been extended to broad spectral regions. Although the principles of these more modern methods are not new, it is the advances in detectors and in electronics and data analysis that have made them practical. Many of the systems have been developed in industry and frequently have not been published. In the cases of those that have

been written up, detailed descriptions of the precise way in which they are used have often been lacking. Usually the technique involves a comparison between the spectral characteristic which is actually obtained at any instant, and that which is required at the instant of termination of the particular layer. In the earlier systems this was carried out visually by displaying both curves on a cathode-ray tube. This works well when there is a close match between predicted and measured performance but frequently errors in earlier layers, and changes in the characteristics of layers from what is expected, cause the actual curves to differ to a greater or lesser extent from the predictions. In these circumstances, there can be great difficulty in assessing visually the correct moment to terminate a layer. The most recent systems, therefore, are usually linked to a computer which calculates a figure of merit which can either be displayed to a plant operator or, better still, used in the completely automatic termination of layers.

Details of scanning monochromator systems have been published by a number of authors. An early description of such a system is that of Hiraga *et al* [20], where the scanning was carried out by a rotating helical slit assembly.

Pelletier and his colleagues in Marseilles [21, 22] have developed two such systems. The first uses a stepping motor to rotate a grating and scan the system over a wide wavelength region, the second uses a holographic grating with a flat spectrum plane in which is situated a silicon photodiode array detector which can be scanned electronically. Sullivan and his colleagues [23–25] have had great success in implementing a completely automatic system of monitoring including error compensation.

11.3.2 The quartz-crystal monitor

The normal modes of mechanical vibration of a quartz crystal have very high Q and can be transformed into electric signals by the piezoelectric properties of the quartz and vice versa. The crystal acts, therefore, as a very efficient tuned circuit that can be coupled into an electrical oscillator by adding appropriate electrodes. Any disturbance of its mechanical properties will cause a change in its resonant frequency. Such a disturbance might be an alteration of the temperature of the crystal or its mass. The principle of monitoring by the quartz-crystal microbalance (as it is called) is to expose the crystal to the evaporant stream and to measure the change in frequency as the film deposits on its face and changes the total mass. In some arrangements the resonant frequency of the crystal is compared with that of a standard outside the plant and the difference in frequency is measured, in others the number of vibrations in a given time interval is measured digitally. Usually the frequency shift will be converted internally into a measure of film thickness using film constants fed in by the operator. Since the signal from the quartz-crystal monitor changes constantly in the same direction it can be used more easily in automatic systems than optical signals.

The mechanical vibrational modes of a slice of quartz crystal are very complicated. It has been found possible to limit the possible modes and the

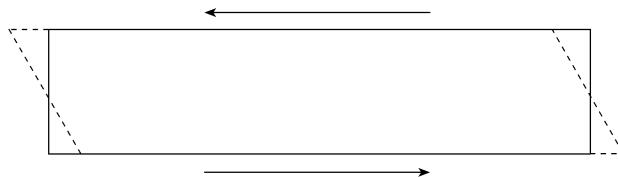


Figure 11.9. Quartz crystal operating in shear.

coupling between them by cutting the slice with respect to the axes of the crystal in a particular way, by proportioning the dimensions of the slice correctly and by supporting the crystal in its holder in the correct way. Quartz-crystal vibrational modes also vary with temperature, some having positive temperature coefficient and some negative, and it has been found possible to cut the slice in such a way that modes which have opposite temperature dependence are intentionally coupled so that the combined effect is a resonant frequency independent of temperature over a limited temperature range. The usual cut of crystal which is used in thin-film monitors is the *AT* cut. This is cut from a slice which was oriented so that it contained the *x* axis of the crystal and was at an angle of $35^\circ 15'$ to the *z* axis. The mode of vibration is a high-frequency shear mode (figure 11.9) and the temperature coefficient is small over the range $-40\text{ }^\circ\text{C}$ to $+90\text{ }^\circ\text{C}$, of the order of $\pm 10^{-6}\text{ }^\circ\text{C}^{-1}$ or slightly greater. The coefficient changes sign several times throughout the range so that the total fractional change in frequency over the complete range is only around 5×10^{-5} . Usually the frequency chosen is around 5 MHz although the range could be anything from 0.5 MHz to 50 or 100 MHz.

As the thickness of the evaporant builds up, the frequency of the crystal falls and the reduction in frequency is proportional both to the square of the resonant frequency and to the mass of the film deposited. In a typical arrangement the measurement of mass thickness can be carried out to an accuracy of around 2%, which should be adequate for most optical filters. Unfortunately, the sensitivity of the crystal decreases with increasing build up of mass and the total amount of material which can be deposited before the crystal must be cleaned is limited. With existing crystals this makes them less useful for multilayer work, especially in the infrared, where in most cases a single crystal could not accommodate a complete filter. One way round this problem is to place a screen over the filter which cuts down the material reaching it to a fraction of that reaching the substrates in the batch. This, of course, reduces the accuracy of the system. Because the crystal measures mass and not optical thickness, it must be calibrated separately for each material used. One further difficulty, important only in some applications, is that the temperature of the crystal must be limited to below $120\text{ }^\circ\text{C}$ (otherwise the temperature coefficient becomes excessively large), so it may not always be possible to keep it at the same temperature as the other substrates in the plant.

There are, however, considerable advantages in the use of quartz-crystal

monitors. Since the output moves in a constant direction and does not reverse it is more readily accommodated by automatic control systems. Further the crystal does not need optical windows with their attendant difficulties of maintenance and screening from the evaporant. Alignment is much simpler than for optical monitors although the requirements for dimensional stability are just as severe. In recent years there have been developments in the use of multiple-crystal sensors distributed around the chamber able to sense changes in the plume of material from the sources and make appropriate corrections to the monitoring calculations. The deposition of only one material on a crystal gives much more stable calibration than if more than one material is involved. This is because the shear modulus of the material as well as the mass determines the shift in frequency and hence the calibration. The common practice, therefore, is now to employ one dedicated set of crystals for each material. With such improvements the results that can be achieved by pure-crystal monitoring are excellent.

In the case of narrowband filters, the optical monitoring is successful because of a built-in error compensation process. This makes it difficult for the crystal monitor to achieve the same yield if peak wavelength is the most important parameter. For processes where error compensation is necessary to achieve the optical performance, optical monitoring is preferred. Then the crystal monitoring is usually still employed, but for source and rate control sensing rather than primary monitoring.

A useful set of instructions and tips on the quartz-crystal monitor will be found in a paper by Riegert [26] which deals much more fully with the topics mentioned above. Manufacturers' manuals include good information also.

11.4 Tolerances

The question of how accurately we must control the thickness of layers in the deposition of a given multilayer is surprisingly difficult to answer and has attracted a great deal of attention over the years.

One of the earliest approaches to the assessment of errors permissible in multilayers was devised by Heavens [27] who used an approximate method based on the alternative matrix formulation in equation (2.146). His method, useful mainly when calculations must be performed manually, consisted of a technique for recalculating fairly simply the performance of a multilayer with a small error in thickness in one of the layers. He showed that the final reflectance of a quarter-wave stack is scarcely affected by a 5% error in any one of the layers.

Lissberger [28, 29] developed a method for calculating the performance of a multilayer involving the reflectances at the interfaces. In multilayers made up of quarter-waves, the expressions took on a fairly simple form which permitted the effects of small errors, in any or all of the layers, on the phase change caused in the light reflected by the multilayer to be estimated. Lissberger's results, applied to the all-dielectric Fabry-Perot filter, show that the most critical layer is the

spacer. The layers on either side of the spacer layer are next most sensitive and the remainder of the layers progressively less sensitive the further they are from the spacer.

We have already mentioned in chapter 7 the paper by Giacomo *et al* [30] where they examined the effects on the performance of narrowband filters of local variations in thickness, or ‘roughness’, of the films. This involved the study of the influence of thickness variations in any layer on the peak frequency of the complete filter. The treatment was similar in some respects to that of Lissberger. For the conventional Fabry–Perot filter, layers at the centre had the greatest effect. If all layers were assumed equally rough, the design least affected by roughness would have all the layers of equal sensitivity and attempts were made to find such a design. A phase-dispersion filter gave rather better results than the simple Fabry–Perot, but still fell short of ideal.

Baumeister [31] introduced the concept of sensitivity of filter performance to changes in the thickness of any particular layer. The method involved the plotting of sensitivity curves over the whole range of useful performance of a filter, curves which indicated the magnitude of performance changes due to errors in any one layer. His conclusions concerning a quarter-wave stack were that the central layer is the most sensitive and the outermost layers least sensitive. An interesting feature of these sensitivity curves for the quarter-wave stack is that the sensitivity is greatest nearest the edge wavelength. This is confirmed in practice with edge filters, where errors usually produce more pronounced dips near the edge of the transmission zone than appear in the theoretical design.

Smiley and Stuart [32] adopted a different approach using an analogue computer. There were some difficulties involved in devising an analogue computer, but, once constructed, it possessed the advantage at the time that any of the parameters of the thin-film assembly could be easily varied. A particular filter, which they examined, was:

$$\text{Air}|4H\ L\ 4H|\text{Air}$$

with $n_H = 5.00$ and $n_L = 1.54$. Errors in one of the $4H$ layers and in the L layer were investigated separately. They found that errors greater than 1% in one $4H$ layer had a serious effect, errors of 5%, for example, caused a drop in peak transmittance to 70% and errors of 10% a drop to 50%, together with considerable degradation in the shape of the pass band. Errors of up to 10% in the L layer had virtually no effect on either the shape of the pass band or on the peak transmittance.

An investigation was performed by Heather Liddell as part of a study reported by Smith and Seeley [33] into some effects of errors in the monitoring of infrared Fabry–Perot filters of designs:

$$\text{Air}|H\ L\ H\ L\ H\ H\ L\ H\ L\ H\ L|\text{Substrate}$$

and

$$\text{Air}|H\ L\ H\ H\ L\ H\ L|\text{Substrate}.$$

A computer program to calculate the reflectance of a multilayer at any stage during deposition was used. Monitoring was assumed to be at or near a frequency of four times the peak frequency (i.e. a quarter of the desired peak wavelength) of the completed filter. It was shown that, if all layers were monitored on one single substrate, then, provided the form of the reflectance curve during deposition was predicted, and it was possible to terminate layers at reflectances other than turning values, there could be an advantage in choosing a monitoring frequency slightly removed from four times peak frequency. If no corrections were made for previous errors, then a distinct tendency for errors to accumulate in even-order monitoring (that is monitoring frequency an even integer times peak frequency) was noted.

The major problem in tolerancing is that real errors cannot be treated as small, that is to say that first-order approximations are unrealistic. The error in one layer interacts nonlinearly with the errors in other layers and it is not possible to treat them separately.

In recent years the most satisfactory approach for dealing with the effects of errors and the magnitude of permissible tolerances has been found to be the use of Monte Carlo techniques. In this method, the performance of the filter is calculated, first with no errors and then a number of times with errors introduced in all the layers. In the original form of the technique, introduced by Ritchie [34], the errors are thickness errors and completely random and uncorrelated. They belong to the same infinite population, taken as normal with prescribed mean and standard deviation. The performance curves of the filter without errors and of the various runs with errors are calculated. Although statistical analyses of the results can be made, it is almost always sufficient simply to plot the various performance curves together, when visual assessment of the effects of errors of the appropriate magnitude can be made. The method really provides a set of traces which reproduce, as far as possible, what would actually be achieved in a succession of real production batches. The characteristics of the infinite normal population can be varied and the procedure repeated. It is sufficient to calculate some eight or perhaps ten curves for a set of error parameters. The level of error at which a satisfactory process yield would be achieved can then readily be determined. In the earliest version of the technique, the various errors were drawn manually from random number tables and converted into members of a normal population using a table of area under the error curve. (The procedure is described in textbooks of statistics—see Yule and Kendall [35], for example.) Later versions of the technique simply generate the random errors by computer. Although the errors are usually drawn from a normal population, the type of population has little effect on the order of the results. Normal distributions are convenient to program, and since there is no strong reason for not using them and because errors made up of a number of uncorrelated effects are well represented by normal distributions, most error analyses do make use of them.

Figure 11.10 shows some examples of plots where the errors are simple independent thickness errors of zero mean. From these and similar results we

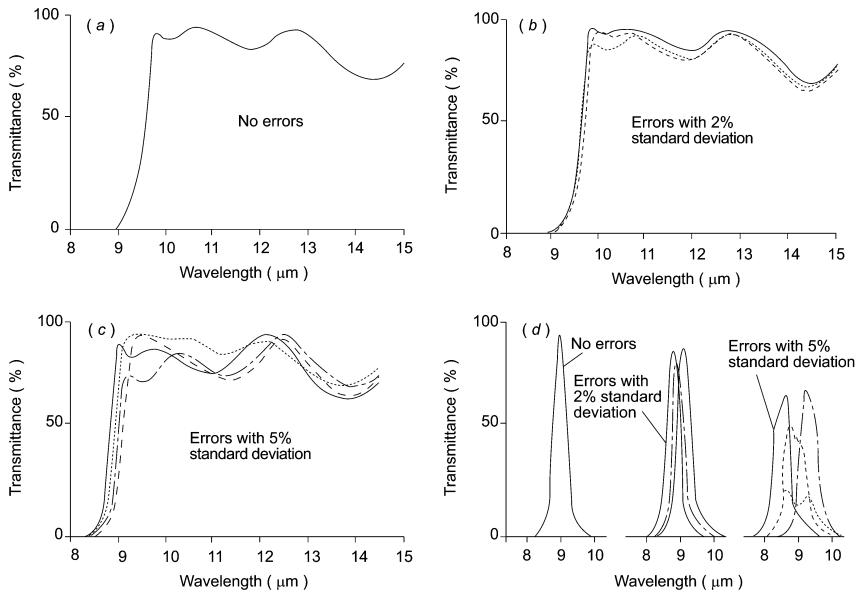


Figure 11.10. The effects of random errors in layer thickness on the performance of thin-film filters. (a), (b) and (c) A typical longwave pass filter of design Air | $L(0.5LH0.5L)^71.49H|$ Ge where $H = \text{PbTe}$ ($n = 5.30$) and $L = \text{ZnS}$ ($n = 2.35$). (d) A DHW or two-cavity filter. Design: Air | $HLLHLHLLHLH$ | Ge where $L = \text{ZnS}$, $H = \text{PbTe}$, $\lambda_0 = 9 \mu\text{m}$. (Some of the curves have been broken for clarity.) (Courtesy of F S Ritchie and Sir Howard Grubb, Parsons & Co. Ltd.)

find that the errors which can be tolerated in a longwave pass filter are normally of standard deviation 5%, in a shortwave pass filter around 2.5%, and in an antireflection coating such as the quarter–half–quarter around 3%.

In a two-cavity filter of the type in figure 11.10, the permissible errors are not greater than 2% while, for narrower filters or filters with greater number of cavities, the tolerances must be tighter. In fact, a rough guide is that the permissible standard deviation is not greater than the halfwidth of the filter. In a Fabry–Perot filter the main effect of random errors is a peak wavelength shift, the shape of the pass band being scarcely affected even by errors as large as 10%. The standard deviation of the scatter in peak wavelength is slightly less than the standard deviation of the layer thickness errors so that some averaging process is operating, although the orders of magnitude are the same.

A system of monitoring in which the thickness errors in different layers are uncorrelated requires that each layer should be controlled independently of the others. In this type of monitoring, therefore, we cannot expect high precision in the centring of narrowband Fabry–Perot filters and we foresee great difficulties in being able to produce narrowband multiple-cavity filters at all.

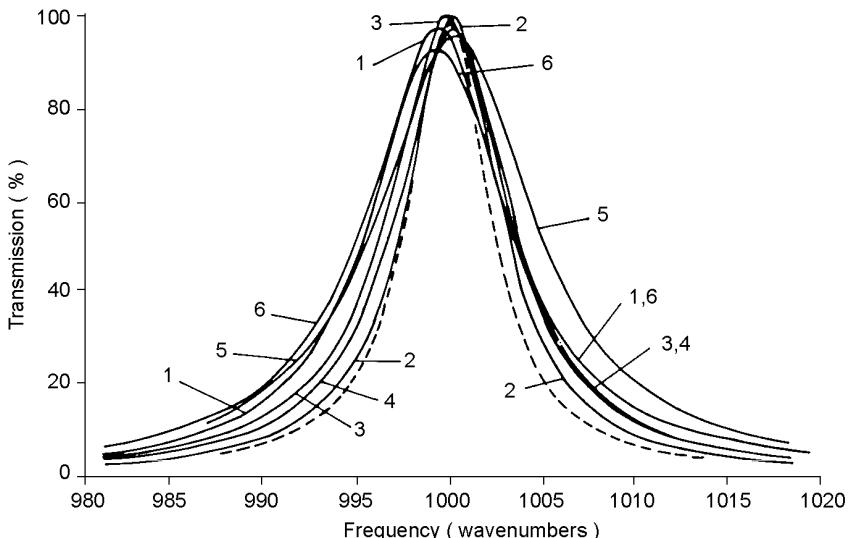


Figure 11.11. The effect of 1% standard deviation reflectance error on the performance of the Fabry–Perot filter: Air |*HLHL HH LHLH*| Ge. The substrate is germanium ($n = 4.0$), L represents a quarter-wave of ZnS ($n = 2.3$) and H a quarter-wave of PbTe ($n = 5.4$). The monitoring is in first order. The dashed curve is the performance with no errors. (After Macleod [36].)

This monitoring arrangement is what we have called indirect. Systems where each layer is controlled on a separate monitoring chip are of this type. There are difficulties with monitoring of low-index layers on a fresh glass substrate because of the small changes in transmittance or reflectance, and so the monitoring chips are usually changed after a low-index layer and before a high index, two or four layers per chip being normal. Sometimes these layers will be monitored to turning values. More frequently what is sometimes called level monitoring will be used. Here the layer reflectance or transmittance signal is terminated at a point removed from the turning value where the signal is still changing, leading to an inherently greater accuracy. This approach involves what is really an absolute measurement of reflectance or transmittance, and so the termination point is frequently chosen to be after a turning value rather than before, so that the extremum can be used as a calibration. This usually implies a shorter wavelength for monitoring or the introduction of a geometrical difference between batch and monitor, placing the monitor nearer the source or placing masks in front of the batch.

Narrowband filters are not normally monitored in this way. Instead, all the layers are monitored on the same substrate, usually the actual filter being produced, a system known as direct monitoring. At the peak wavelength of the filter, the layers should all be quarter-waves or half-waves, and so we can

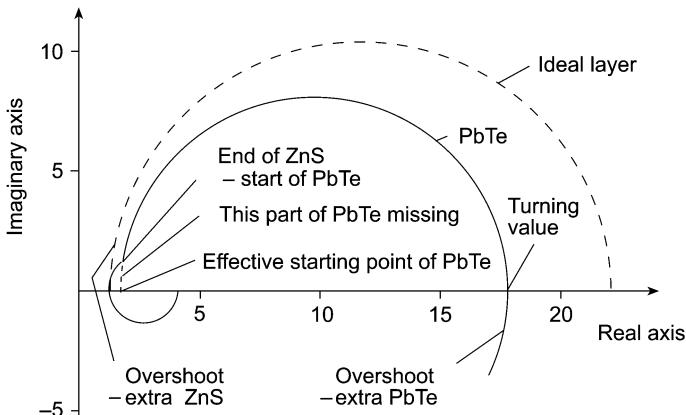


Figure 11.12. The admittance locus of the first two layers of the filter in figure 11.10 when there is an overshoot in the first layer of around one-eighth wave optical thickness. (After Macleod [36].)

expect a signal which reaches an extremum at each termination point. The accuracy cannot therefore be particularly high for any individual layer and, at first sight, it would appear that the achievable accuracy should be far short of what must be required. Since each layer is being deposited over all previous layers on the monitor substrate, then there is an interaction between the errors in any layer and those in the previous layers not included in the tolerancing calculation described above. We really require a technique which models the actual process as far as possible and this is a quite straightforward piece of computing. Each layer is simply considered to be deposited on a surface of optical admittance corresponding to that of the multilayer which precedes it, rather than on a completely fresh substrate. The results of such a simulation are shown in figure 11.11, taken from Macleod [36], which demonstrates the powerful error compensation mechanism that has been found to exist. The compensation has also been independently and simultaneously confirmed by Pelletier and his colleagues [37]. Its nature is perhaps best explained by the use of an admittance diagram.

Figure 11.12 shows such a diagram drawn for several quarter-waves. Since both the isoreflectance contours (see chapter 2) and the individual layer loci are circles centred on the real axis, the turning values must always occur at the intersections of the loci with the real axis, regardless of what has been deposited earlier. At the termination point of each layer there is the possibility of restoring the phase to zero or to π . As far as any individual layer is concerned, it is principally the over- or undershoot of the previous layer that affects it. If the previous layer is too thick, the current one will tend to be thinner to compensate, and vice versa. Of course it is impossible to cancel completely all effects of an error in a layer. The process is actually transforming the thickness errors into

errors in reflectance at each stage since the loci will be slightly displaced from their theoretical position. This is not a serious error. As can be guessed from the shape of the diagram, the reflectance error is a second-order effect. Since the phase is self-corrected each time a layer is deposited, the peak wavelength of the filter will remain at the desired value, that of the monitoring wavelength. The remaining error, the residual one in reflectance, is then translated into changes in peak transmittance and halfwidth. Since the reflectance change is always a reduction, the bandwidth of an actual filter is invariably wider than theoretical. The peak transmittance falls to the extent that the reflectances on either side of the spacer layer are unbalanced. This is usually quite small and the reduction in peak transmittance is generally much less important than the increase in bandwidth.

In this monitoring arrangement, thickness errors in any individual layer are a combination of a compensation of the error in the previous layer together with the error committed in the layer itself. The magnitude of the thickness errors can be quite misleading in interpreting whether or not the filter can be made successfully. In figure 11.10, for example, thickness errors of the order of 50% occur in some layers and yet the filter characteristics are all useful ones.

The important characteristic is actually the error in reflectance or transmittance in determining the turning values, and it is possible to develop theoretical expressions which relate the reflectance or transmittance errors to the reduction in performance of the final filter [36]. This analysis includes an assessment of the sensitivity of each layer to errors which indicate those layers where the greatest care in monitoring should be exercised. These can be different from the thickness sensitivity of Lissberger [28, 29] already mentioned. With high-index spacer layers, greatest sensitivity is found in the low-index layers following the spacer, while with low-index spacers, the spacer itself has the highest sensitivity. A feature of this analysis is that it demonstrates that for any particular error magnitude, there is a point where improved halfwidth does not result from an increase in the number of layers because the effect of errors is increasing more rapidly than the theoretical decrease in bandwidth. Then it is necessary to move to second- and higher-order spacers if decreased bandwidth is to result. This corresponds to what is found in practice. The error analysis also demonstrates that high-index spacers are to be preferred over low-index. We have already seen in chapter 7 that high-index spacers give decreased angular sensitivity and greater tuning range.

Formulae which permit the calculation of the errors in reflectance, in halfwidth and in peak transmittance as a function of the magnitude of the random errors in determining the turning values exist [36], but for most purposes a computer simulation will suffice. It should be noted that the compensation is effective only for the first order. Second-order monitoring, that is monitoring at the wavelength for which the layers are all half-waves, is not effective in preserving the peak wavelength. We can understand this because the admittance diagram is quite different and so the compensation is of a different nature. Likewise, third-order monitoring is not as effective as first-order, and, although

the scatter in peak wavelength is less than that obtained with second-order monitoring, it is, nevertheless, quite large.

Multiple-cavity filters are similar in behaviour but there are some complications. The coupling layers in between the various Fabry–Perot sections of the filter turn out to be particularly sensitive to errors in a rather peculiar way. Preliminary examination of the admittance diagram for the various layers of a multiple-cavity filter and even the standard error analysis do not immediately reveal any marked difference in terms of error sensitivity between these layers and those of Fabry–Perot filters. Closer investigation shows that there is always one transition from one layer to the next occurring at or near to the central coupling layer where a thickness error is compensated by an error of the same rather than the opposite sense [38]. The condition is sketched in figure 11.13. An increase in thickness in the first layer results in an increase in thickness of the subsequent layer and vice versa. This condition must occur once between each pair of cavities. The net result is an increase or decrease in the relative spacing of the cavities causing the appearance of a multiple-peaked characteristic curve. The peaks become more pronounced, the greater the relative error in spacing. One of the peaks always corresponds to the normal control wavelength and is close to the theoretical transmittance. The other peaks (one for a two-cavity, two for a three-cavity, and so on) can appear on either side of the main peak depending on the nature of the particular errors. This false compensation can be destroyed if the second of the two layers concerned can be controlled independently of the others, either on a separate monitor plate or by a quartz-crystal monitor, or even by simple timing. It is essential that it should also be deposited on the regular monitor as well, so that the compensation of the full filter should not be destroyed [38].

Pelletier and his colleagues [39] have studied theoretically the behaviour of the ‘maximètre’ types of monitoring systems in the production of narrowband filters. They conclude that, as we would expect, the accuracy of the system in the production of single layers is very much better than a single-wavelength system. In the monitoring of narrowband filters all on one substrate there is a compensation process operating like the turning value method but it is more complex in operation. For very small errors in most layers the system works adequately, but for large errors in most layers or small errors in certain critical layers, the errors accumulate in such a way as to cause a drastic broadening of the bandwidth of a Fabry–Perot filter or complete collapse of a multiple-cavity filter. Pelletier has introduced two concepts to describe this behaviour. Accuracy represents the error that will be committed in any particular layer without reference to the multilayer system as a whole. Stability represents the way in which the errors accumulate as the multilayer deposition proceeds. The accuracy of the ‘maximètre’ is excellent and greater than in the turning value method, but the stability in the control of narrowband filters is very poor and it can easily become completely unstable. Subsidiary measurements are therefore required to ensure stability if advantage is to be taken of the very great accuracy

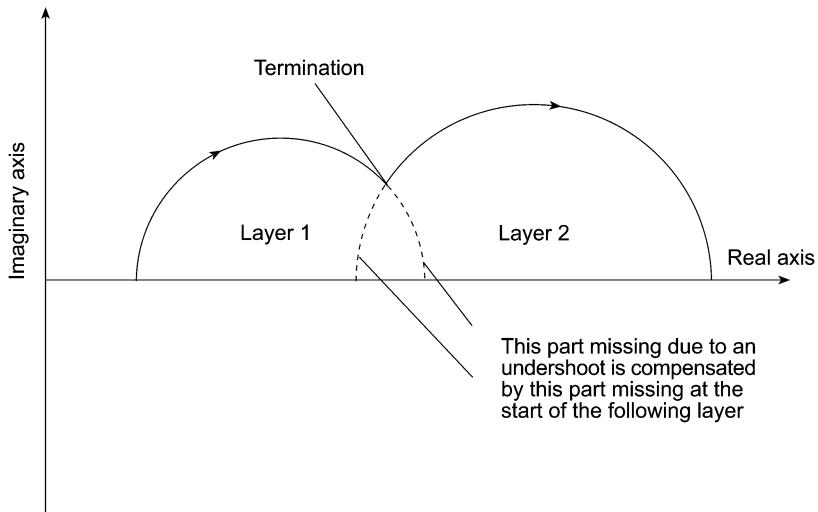


Figure 11.13. Error compensation when the admittance circles are on the same side of the real axis. (After Macleod and Richmond [38].)

that is possible. Narrowband filters and their monitoring systems have been surveyed by Macleod [36].

The concepts of accuracy and stability and the discovery that the one does not ensure the other imply that different measurements may be necessary to ensure that both are simultaneously assured. This leads to the idea of broadband monitoring in which simultaneous measurements are made at a large number of wavelengths over a wide spectral region and a merit function representing the difference between actual and desired signals is computed. The merit function can then be used as a monitoring signal and layer deposition terminated when the merit function reaches a minimum. Although perfect deposition should ensure a minimum of zero in the figure of merit, inevitable errors in layer index and homogeneity will perturb the result. The accuracy and stability of such a broadband system in the monitoring of certain components such as beam splitters has been investigated by computer simulation [41] and evidence found for useful error compensation. Apart from the very qualitative justification discussed above no theory for such compensation yet exists and it may operate only in quite specific cases. Extensions of broadband monitoring to a system that would re-optimise those layers of a design yet to be deposited on the basis of errors measured in earlier layers appear possible and are under investigation in a number of laboratories. Even if successfully developed they are never likely to be able to reduce the need for stable reproducible materials.

Quartz-crystal monitoring, in which the mass rather than optical thickness is measured, seems unlikely to possess powerful compensation. Yet simulation

of a simple broadband system for antireflection coatings comparing optical monitoring with quartz crystal gave results which indicate that the quartz crystal is in no way inferior [42]. The relative merits of quartz crystal and optical monitoring form a subject of constant debate and published results for quartz crystal are impressive [43, 44]. It is clear that narrowband filters, if they are to be controlled in peak wavelength, do require direct optical monitoring, but quartz crystal monitoring is suitable for most other filter types. The general opinion, based to some extent on instinct, is that quartz-crystal monitoring is most suitable for production of successive batches of identical components. For single runs of varying coating types, optical monitoring appears normally to be preferred. Optical monitoring is also preferred in applications such as filters for the far infrared, where very large thicknesses of materials are deposited in each coating run.

References

- [1] Holland L and Steckelmacher W 1952 The distribution of thin films condensed on surfaces by the vacuum evaporation method *Vacuum* **2** 346–64
- [2] Behrndt K H 1963 Thickness uniformity on rotating substrates *Transactions of the 10th AVS National Vacuum Symposium* (London: McMillan) pp 379–84
- [3] Knudsen M 1915 Das Cosinusgesetz in der kinetischen Gastheorie *Ann. Phys.* **48** 1113–21
- [4] Jancke E and Emde F 1952 *Tables of Higher Functions* 5th edn (Leipzig: Teubner)
- [5] Keay D and Lissberger P H 1967 Application of the concept of effective refractive index to the measurement of thickness distributions of dielectric films *Appl. Opt.* **6** 727–30
- [6] Graper E B 1973 Distribution and apparent source geometry of electron-beam heated evaporation sources *J. Vacuum Sci. Technol.* **10** 100–3
- [7] Richmond D 1976 Thin film narrow band optical filters *PhD Thesis* (Newcastle upon Tyne Polytechnic)
- [8] Ramsay J V, Netterfield R P and Mugridge E G V 1974 Large-area uniform evaporated thin films *Vacuum* **24** 337–40
- [9] Holland L 1956 *Vacuum Deposition of Thin Films* (London: Chapman and Hall)
- [10] Mattox D M 1978 Surface cleaning in thin film technology *Thin Solid Films* **53** 81–96
- [11] Lee C C 1983 Moisture adsorption and optical instability in thin film coatings *PhD Dissertation* (University of Arizona)
- [12] Cox J T and Hass G 1958 Antireflection coatings for germanium and silicon in the infrared *J. Opt. Soc. Am.* **48** 677–80
- [13] Banning M 1947 Practical methods of making and using multilayer filters *J. Opt. Soc. Am.* **37** 792
- [14] Polster H D 1952 A symmetrical all-dielectric interference filter *J. Opt. Soc. Am.* **42** 21–5
- [15] Perry D L 1965 Low loss multilayer dielectric mirrors *Appl. Opt.* **4** 987–91
- [16] Behrndt K H and Doughty D W 1966 Fabrication of multilayer dielectric films *J. Vacuum Sci. Technol.* **3** 264–72
- [17] Giacomo P and Jacquinot P 1952 Localisation précise d'un maximum ou d'un

- minimum de transmission en fonction de la longeur d'onde. Application à la préparation des couches minces *J. Phys. Rad.* **13** 59A–64A
- [18] Ring J 1957 *PhD Thesis* (University of Manchester)
 - [19] Lissberger P H and Ring J 1955 Improved methods for producing interference filters *Opt. Acta* **2** 42–6
 - [20] Hiraga R, Sugawara N, Ogura S and Amano S 1974 Measurement of spectral characteristics of optical thin film by rapid scanning spectrophotometer *Japan. J. Appl. Phys. (Suppl. 2, Part 1)* 689–92
 - [21] Borgogno J P, Bousquet P, Flory F, Lazarides B, Pelletier E and Roche P 1981 Inhomogeneity in films: limitation of the accuracy of optical monitoring of thin films *Appl. Opt.* **20** 90–4
 - [22] Flory F, Schmitt B, Pelletier E and Macleod H A 1983 Interpretation of wide band scans of growing optical thin films in terms of layer microstructure *Proc. Soc. Photo-Opt. Instrumentation Eng.* **401** 109–16
 - [23] Sullivan B T and Dobrowolski J A 1992 Optical multilayer coatings produced with automatic deposition error compensation *Optical Interference Coatings (Tucson, AZ)* (Optical Society of America) pp 278–9
 - [24] Sullivan B T and Dobrowolski J A 1992 Deposition error compensation for optical multilayer coatings. I. Theoretical description *Appl. Opt.* **31** 3821–35
 - [25] Sullivan B T and Dobrowolski J A 1993 Deposition error compensation for optical multilayer coatings. II. Experimental results—sputtering system *Appl. Opt.* **32** 2351–60
 - [26] Riegert R P 1968 Optimum usage of quartz crystal monitor based devices *IVth International Vacuum Congress (Manchester)* (Bristol: Institute of Physics and the Physical Society) pp 527–30
 - [27] Heavens O S 1954 All-dielectric high-reflecting layers *J. Opt. Soc. Am.* **44** 371–3
 - [28] Lissberger P H 1959 Properties of all-dielectric filters. I. A new method of calculation *J. Opt. Soc. Am.* **49** 121–5
 - [29] Lissberger P H and Wilcock W L 1959 Properties of all-dielectric interference filters. II. Filters in parallel beams of light incident obliquely and in convergent beams *J. Opt. Soc. Am.* **49** 126–30
 - [30] Giacomo P, Baumeister P W and Jenkins F A 1959 On the limiting bandwidth of interference filters *Proc. Phys. Soc.* **73** 480–9
 - [31] Baumeister P W 1962 Methods of altering the characteristics of a multilayer stack *J. Opt. Soc. Am.* **52** 1149–52
 - [32] Smiley V N and Stuart F E 1963 Analysis of infrared interference filters by means of an analog computer *J. Opt. Soc. Am.* **53** 1078–83
 - [33] Smith S D and Seeley J S 1968 *Multilayer Filters for the Region 0.8 to 100 Microns* (Air Force Cambridge Research Laboratories)
 - [34] Ritchie F S 1970 Multilayer filters for the infrared region 10–100 microns *PhD Thesis* (University of Reading)
 - [35] Yule G U and Kendall M G 1958 *An Introduction to the Theory of Statistics* 14th edn (London: Charles Griffin)
 - [36] Macleod H A 1972 Turning value monitoring of narrow-band all-dielectric thin-film optical filters *Opt. Acta* **19** 1–28
 - [37] Bousquet P, Fournier A, Kowalczyk R, Pelletier E and Roche P 1972 Optical filters: monitoring process allowing the auto-correction of thickness errors *Thin Solid Films* **13** 285–90

- [38] Macleod H A and Richmond D 1974 The effect of errors in the optical monitoring of narrow-band all-dielectric thin film optical filters *Opt. Acta* **21** 429–43
- [39] Pelletier E, Kowalczyk R and Fournier A 1973 Influence du procédé de contrôle sur les tolérances de réalisation des filtres interférentiels à bande étroite *Opt. Acta* **20** 509–26
- [40] Macleod H A 1976 Thin film narrow band optical filters *Thin Solid Films* **34** 335–42
- [41] Vidal B, Fournier A and Pelletier E 1979 Wideband optical monitoring of nonquarterwave multilayer filters *Appl. Opt.* **18** 3851–6
- [42] Macleod H A 1981 Monitoring of optical coatings *Appl. Opt.* **20** 82–9
- [43] Pulker H K 1978 Coating production: new ideas at a time of demand *Opt. Spectra* **12** 43–6
- [44] Laan C J v d and Frankena H J 1977 Monitoring of optical thin films using a quartz crystal monitor *Vacuum* **27** 391–7

Chapter 12

Specification of filters and environmental effects

Ideally, if a filter is to be manufactured for a customer for a given application, then the performance required by the customer, and the design, manufacturing and test methods, should all be defined, even if only implicitly. These details form different aspects of the specification of the filter.

There is no standard method for setting up the specification of an optical filter or coating, the problem being much the same as for any other device. There are three main aspects to be considered: the performance specification which lists the details of the performance required from the filter and is usually the customer's specification, the manufacturing specification which defines the design and details the steps involved in the manufacture of the filter, and the test specification laying down the tests which must be carried out on the filter to ensure that it meets the performance requirements, these latter aspects being mainly the concern of the manufacturer. In the following notes a few of the more important points are mentioned, but they do not form a complete guide to the writing of specifications, which is a complete subject in its own right.

Optical filter specifications can conveniently be divided into two sections, one concerned with optical properties and the other with physical or environmental properties. We shall first of all consider the optical properties.

12.1 Optical properties

12.1.1 Performance specification

The performance specification of a filter is really a statement of the capabilities of the filter in a language that can readily be interpreted by both system designer, and customer, and filter manufacturer alike. It can sometimes be prepared by a filter manufacturer from a knowledge of the performance which he knows he can achieve, either for a customer or possibly without having a particular application

in mind, as in the case of a standard product in a catalogue about which little need be said here. Probably more often, the performance specification will be written by the system designer and will state a level of performance required from a filter in order to achieve a desired level of performance from a system. In writing such a specification, an answer must first of all be given to the question: what is the filter for? The purpose of the filter must be set down as clearly and concisely as possible and this will form the basis for the work on the performance specification. There is really no systematic method for specifying the details of performance. Sometimes it happens that the performance of the system in which the filter is to be used must be of a certain definite level, otherwise there will be no point in proceeding further. The filter performance requirements can then be quite readily set down. Often, however, it will not be quite so simple. No absolute requirement for performance may exist, only that the performance should be as high as possible within allowable limits of complexity or perhaps price. In such a case, the performance of the system with different levels of filter performance must be balanced against cost and system complexity, and a decision made as to what is reasonable. The final specification will be a compromise between what is desirable and what is achievable. This will often need the input of much design and manufacturing information and close contact between customer and manufacturer. It should always be remembered in this that specifications that cannot be met in practice can be of only academic interest.

By way of an example let us briefly consider the case where a spectral line must be picked out against a continuum. Clearly a narrowband filter will be required, but what will be the required bandwidth and type of filter? The energy from the line to be transmitted by the filter will depend on the peak transmittance (assuming that the peak of the filter can always be tuned to the line in question), while the energy from the continuum will depend on the total area under the transmission curve, including the rejection region at wavelengths far removed from the peak. The narrower the pass band, the higher the contrast between the line and the continuum, especially as narrowing the pass band generally also improves the rejection. However, the narrower the pass band, because of the increased difficulty of manufacture, the higher the price, and, further, because of the increased sensitivity to lack of collimation, the larger the tolerable focal ratio. This latter point implies that for the same field of view, a filter with a narrower bandwidth must be made larger to permit the use of the larger focal ratio, which in turn will increase still further the difficulties of manufacture and, possibly, the complexity of the entire system. Another way of improving the performance of the filter is by increasing the steepness of edge of the pass band while still retaining the same bandwidth. A rectangular pass-band shape gives higher contrast than a simple Fabry-Perot of identical halfwidth and usually possesses the additional advantage that the rejection remote from the peak of the filter is also rather greater. This edge steepness can be specified by quoting the necessary tenth peak bandwidth or even the hundredth peak bandwidth. Again, inevitably, the steeper the edges, the more difficult the manufacture and the higher

the price.

Because filters, as with any manufactured product, cannot be made exactly to a specification in absolute terms, some tolerances must always be stated. For a narrowband filter, the principal parameters that should be given tolerances are peak wavelength, peak transmittance and bandwidth. Since in almost all applications the higher the peak transmittance the better, it is usually sufficient to state a lower limit for it. There are two aspects of peak wavelength tolerance. The first is uniformity of peak wavelength over the surface of the filter. There will always be some grading of the films, although perhaps small, and a limit must be put on this. The effect is similar to that of an incident cone of illumination (which has been discussed on pp 288–92) and it is usually best to limit the uniformity errors in the specification to not more than one-third of the halfwidth. The second aspect is error in the mean peak wavelength measured over the whole area of the filter. The tolerance for this is usually made positive so that the filter can always be tuned to the correct wavelength by tilting. For a given bandwidth the amount of tilt that can be tolerated in any application will be determined to a great extent by the aperture and field of the system, since the total range of angles of incidence that can be accepted by a filter falls as the tilt angle is increased.

The bandwidth of the filter should also be specified and a tolerance put on it, but, because of the difficulty of controlling bandwidth very accurately, it is not usually desirable to tie it up too tightly and the tolerance should be kept as wide as possible, not normally less than 0.2 times the nominal figure unless there is a very good reason for it.

One other important parameter involved in the optical performance specification, is rejection in the stopping zones, which may be defined in a number of different ways. Either the average transmittance over a range, or absolute transmittance at any wavelength in the range, can be given an upper limit. The first would usually apply where the interfering source is a continuum and the second where it is a line source, in which case the wavelengths involved should be stated, if known.

Yet another entirely different method of specifying filter performance is by drawing maximum and minimum envelopes of transmittance against wavelength. The performance of the filter must not fall outside the region laid down by the envelopes. It is important that the acceptance angle of the filter also be stated. This type of specification is rather more definite than the first type mentioned above. A disadvantage, however, is that it may be rather too severe since everything is stated in absolute terms when average values may be just as good. A further point is that it is impossible to devise a test to determine whether or not a filter meets an absolute specification of this type. Finite bandwidth of the measuring apparatus will ultimately be involved. It is advisable, therefore, if specifying a filter in this way, to include a note to the effect that the performance specified at each wavelength is the average over a certain definite interval.

There is little else that can be said in general terms about the optical performance specification. In any one application these factors will assume

different relative importance and each case must to a very great extent be considered on its own merits. Clearly this is an area where it is of prime importance that the system designer works very closely with the filter designer.

12.1.2 Manufacturing Specification

We shall now consider briefly the manufacturing specification containing the filter design together with details of the manufacturing method. In most cases, this will be intended for the use of the plant operator.

First, the filter design, including the materials, will be given. Most filters contain not more than three different thin-film materials having relatively low, medium and high refractive index. Designs are usually written in terms of quarter-wave optical thicknesses at a reference wavelength λ_0 using the symbols L , M and H . Typical designs may be written:

$$\begin{aligned} L|Ge|LHLHHLH \quad L = ZnS \quad H = Ge \\ M|Si|MHLHHLH \quad L = CaF_2 \quad M = ZnS \quad H = Ge \end{aligned}$$

the substrates being indicated by the symbols | Ge | and | Si |. Next the constructional details should be written down. These consist of the monitoring method to be used, including the wavelengths, and the form of the signals together with other important details such as substrate temperature, special types of evaporation sources, and so on. It will be found useful to arrange the whole manufacturing specification in the form of a table that can be issued to the plant operators for use as a checklist. Operators should always be encouraged to observe critically the operation of the plant so that faults or anomalies can be spotted at an early stage, and it is a help in this if they are expected to list comments in appropriate places on the form. It will also be found convenient to give each filter production batch a different reference number. Once the filters are produced, the completed specification form can then be filed by the plant operator to form the plant logbook. Additional information such as pumping performance can also be recorded on the sheets, useful from the maintenance point of view. For calculation purposes there is no consensus on whether the incident medium should be at the top or the foot of a table of design. For manufacture, however, the first layer to be deposited is necessarily next to the substrate and it is usual to list the layers in tables of manufacturing instructions from innermost, that is next to the substrate, to outermost.

Software products can assist in setting up the manufacturing specification, especially the sequence of monitoring signals. In some cases these can be automatically fed into the deposition controller so that the printed copy can be simply for reference and record keeping.

12.1.3 Test Specification

Probably the most important specification of all is the test specification. This lays down the complete set of tests that will be carried out on the filters to measure the performance. It should always be remembered that, although the filter will have been designed to meet a particular performance specification, it is only the performance laid down in the test specification that can actually be guaranteed, and, although it may seem obvious, the test specification must be written with the requirements of the performance specification always in mind. In fact it is possible simply to specify the performance of a filter as that which will pass the appropriate test specification. It will sometimes be found that the test specification, if it exists at all, is a rather loose document or that sometimes the customer's performance specification will serve both roles. If so, then someone somewhere along the line will be interpreting the performance specification in order to decide on the tests which have to be applied, and it is always better to have the tests and the method of interpretation in writing.

The first essential in any test specification is a definite statement of the performance or the make and type of the test equipment to be used. This ensures that results can be repeated if necessary, even if remote from the original testing site. Next, the various tests together with the appropriate acceptance levels can be set down.

It is in the measurement of such factors as uniformity where the tests and the method of interpretation are particularly important. Absolute uniformity is impossible to measure in the ordinary way. The peak wavelength would have to be measured at every point on the filter with an infinitesimally small measuring beam. A simpler and usually satisfactory method is to check the peak wavelength at the centre of the filter and at four approximately equally spaced areas around the circumference, using a specified area of measuring beam. The spread over the filter is taken to be the spread in the values of peak wavelength over the five separate measurements. The spectrometer used for the measurement will also have a finite bandwidth and features of the filter which are rather less than this will, in general, not be picked up. This applies particularly to the measurement of rejection. Rejection must be measured over a very wide region, and for the test to be completed in a reasonable time, a fast scanning speed must be used, which in turn requires a broad bandwidth. This averages the measurement over a finite region and is one of the reasons for stating the actual wavelengths of the lines if the energy that is to be rejected has a line rather than a continuous spectrum.

A technique for measuring the rejection of films using a Fourier transform spectrometer has been suggested by Bousquet and Richier [1]. While this is difficult to apply in the visible region, the availability of commercial Fourier transform spectrometers for the infrared makes it a feasible technique for infrared filters.

Of course, inevitably, the more extensive the testing which must be carried out on each individual filter, the more expensive that filter is going to be.

Performance testing of low-price standard filters is, in the main, carried out on a batch basis, with only a few details being checked on each individual filter. This is a point which should be borne in mind by a prospective customer buying a standard filter from a catalogue, that a superlative level of performance cannot be absolutely guaranteed from a single given filter, which, by its price, cannot have had more than the basic testing carried out on it.

So far we have dealt with the directly measurable optical performance of the filter, but there are additional properties which are of a subjective nature and rather more difficult to measure. These are connected with the quality and finish of the films and substrates. Substrates are specified as for any optically worked component, details such as flatness or curvature of surface, degree of polish and allowable blemishes, sleeks and the like can all be stated. We shall not consider substrates further here. There is a specification, used particularly in the USA, MIL-E13830 A, which gives a useful set of standards for optical components including substrates.

The quality of the coating can be measured by the presence or absence of defects such as pinholes, stains, spatter marks and uncoated areas.

Pinholes are important for two reasons. First they are actually small uncoated, or partially uncoated, areas and as such will allow extra light to be transmitted in the rejection regions, reducing the overall performance of the filter. Second, and this is especially so for filters for the visible region, they are unsightly and detract from the appearance. In fact, they usually look worse to the eye than the effect they actually have on performance. Apart from the purely subjective appearance, the permissible level of pinholes can be defined on the basis of a given maximum number of a certain size per unit area, calculated to reduce the rejection in the stop bands by not more than a given amount. To calculate this figure, a minimum area of filter that will be used at any one time must be assumed. This will depend on the application, but in the absence of any definite information on this a suitable figure is 5 mm × 5 mm. Obviously the smaller this area, the lower the size of the largest pinhole. Of course, the actual counting of pinholes in any filter would involve a prohibitive amount of labour and in practice, with visible filters, the measurement is usually carried out visually, comparing the filter with limit samples. A simple fixture consisting of a light box with sets of filters laid out on it, some just inside, some on, and some just outside the limit, can be easily constructed. For infrared filters on transparent substrates this method can also be applied, but for filters on opaque substrates it is easier to measure actual rejection performance.

Spatter marks are caused by fragments of material ejected from the sources and, unless gigantic, do not affect the optical performance, the danger being that the fragments may be removed later, leaving pinholes. The incidence can be tied down just as with pinholes, but, as the optical performance is not affected unless the number of marks is enormous, the basis for deciding what is permissible is entirely subjective—although usually if the spatter is particularly bad it will be accompanied by pinholes. Often specifications will state that there must

be no spatter marks visible to the naked eye, but this is vague, particularly when dealing with inspectors with no optical experience. Disagreements can arise between manufacturer and customer especially when, as can happen, the customer's inspectors use an eyeglass to assist the naked eye. The best course is probably to relate the test to agreed limit samples when it can be carried out in exactly the same way as for pinholes, or else to omit it altogether.

Stains can be caused in a number of ways. The most common reason is a faulty substrate. One type of mark that is often seen, especially when antireflection coatings are involved, is due to a defect in the optical working. The polishing process consists partly of a smoothing out of irregularities in the surface by a movement of material. If the grinding, which always precedes the polishing, has been too coarse, then the deeper pits during the polishing are filled in with material which is only loosely bonded to the surface, although the polish appears satisfactory to the eye. In the heating and then coating of the surface, this poorly bonded material breaks away, leaving a patch of surface that is etched in appearance and often possesses well-defined boundaries. The only remedy for this type of blemish is improved polishing techniques. Other stains that may appear can be caused by faulty substrate cleaning. If water or even alcohol is allowed to dry on a surface without wiping, water marks appear. Droplets should always be removed from the surface by a final vapour cleaning stage, or by blowing with clean air (great care must be taken to make sure the air is clean and does not carry oil vapour with it), or by wiping with a clean tissue or cloth during the cleaning process. Water should never be allowed to dry on the surface by itself. Stains, unless particularly bad, do not usually affect the optical performance to anything like the extent their appearance would suggest (except in the case of very high performance components such as Fabry–Perot plates or laser mirrors), and the basis for judging them is again subjective.

Finally, the filter must be held in a jig during coating so that at least some uncoated areas must exist. These usually take the form of a ring around the periphery of the filter, perhaps around 0.5 mm wide. There will be a slight taper in the coating at the very edge which must also be allowed for, the combined taper and uncoated area forming a strip perhaps 1.0 mm in width. The uncoated area actually serves a useful purpose because mechanical mounts can grip the component at this point without damaging the coating. Damage near the edge is dangerous because it is there that delamination is frequently initiated. Jigs that allow the substrates to chatter as they rotate can cause such defects. Uncoated areas should not occur within the boundary of the filter proper; when they do it is a sign of adhesion failures that may recur. They may be due to substrate contamination or to moisture penetration with weakening of adhesion, as described in chapter 9, but they are always cause for rejection of the component. Blisters, too, which are a slightly different version of the same fault, are also cause for immediate rejection.

12.2 Physical properties

As far as the physical properties of the filter are concerned, there are two primary aspects. First, the dimensions of the filter must meet the requirements laid down. This is purely a matter of mechanical tolerances that we need not go into any further here. Second, the filter must be capable of withstanding, as far as possible, the handling it will receive in service and also of resisting any attack from the environment. The assessment of the robustness of the coating will now be considered in greater detail.

The approach almost invariably used in defining and testing the robustness of a coating is to combine the performance and test specifications. A series of tests reproducing typical conditions likely to be met in practice is set up, and then performance is defined as being a measure of the ability to pass the particular tests. This avoids the difficulty in setting up a more general performance specification.

There is one basic difference between the tests of optical performance and those we are about to discuss. Optical tests are all nondestructive in nature while tests of robustness are, in the main, destructive. The filters are tested deliberately to cause damage, and the extent of the damage, if it can be measured, used as a measure of the robustness of the filter. It is thus not possible to carry out the whole series of tests on the actual filter that is to be supplied to the customer and it is normal to use a system of batch testing. A number of filters is made in a batch and either one or perhaps two chosen at random for testing. Provided these test filters are found acceptable then the complete batch is assumed satisfactory. This arrangement is, of course, not peculiar to thin-film devices. Another aspect of this batch testing is involved in what is known as a type test. Often if a large number of filters, all of the same type and characteristic, are involved, a series of very extensive and severe tests will be carried out on a sample of filters from a number of production batches. The test results will then be assumed to apply to the entire production of this type of filter. Once the filters have passed this type test, normal production testing is carried out on a reduced scale. It is imperative that once the type test has been successful there are no subsequent changes, even of a minor nature, in the production process, otherwise the type test would be invalidated.

12.2.1 Abrasion Resistance

Coatings on exposed surfaces, such as the antireflection coating on a lens, will probably require cleaning from time to time. Cleaning usually consists of some sort of rubbing action with a cloth or perhaps lens tissue. Often there may be dust or grit on the surface of the lens, which may not be removed before rubbing. The result of such treatment is abrasion and it is important to have the abrasion resistance of exposed coatings as high as possible. An absolute measure of abrasion resistance is not at all easy to establish because of the difficulty of defining it in absolute terms, and the approach which has been adopted has been to reproduce, under controlled conditions, abrasion similar to that likely to be met in

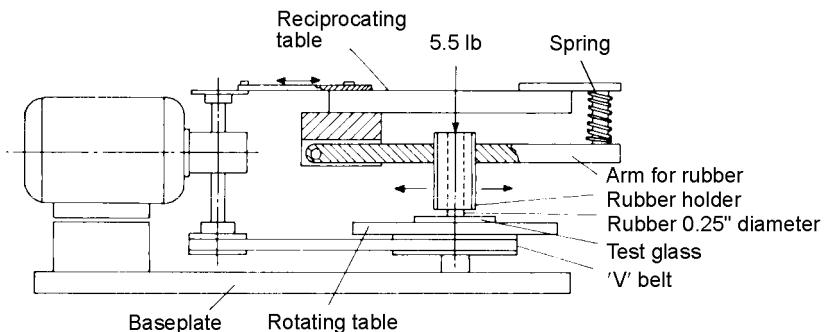


Figure 12.1. Schematic arrangement of an abrasion machine. The reciprocating table is supported by two horizontal bars not shown in the diagram. (After Holland and van Dam [2].)

practice only rather more severe. The degree to which the coating withstands the treatment is then a guide to its performance in actual use. In the UK a great deal of work was carried out on standardising this test by the Sira Institute (formerly the British Scientific Instrument Research Association). Their method involved a standard pad made from rubber loaded with emery powder, which, with a precise load, is drawn across the surface under test a given number of times—typically 20 times with a loading of 5 lb in^{-2} . Their work was directed mainly towards the assessment of the performance of magnesium fluoride single-layer antireflection coatings for the visible. It has been established that sufficiently robust coatings of this type do not show signs of damage under the normal test conditions given above. Abrasion resistance, however, has been found to be not just a function of the film material but also of the thickness. Multilayer coatings are generally much more prone to damage than either of the component materials in single-layer form. It is therefore necessary to establish fresh standards for each and every type of coating. There are also difficulties in achieving exactly the same abrading performance from different batches of abrading pad. Similar tests using pads that may or may not include abrading particles are widely used. In the spectacle industry it is not uncommon to find similar tests using rough cloth and even steel wool.

Unfortunately such tests do not normally produce an actual measure of the abrasion resistance, but merely decide whether or not a given coating is acceptable. Because of this, some investigations into a better arrangement were carried out by Holland and van Dam [2]. Their test is based on the principle that a measurement of abrasion resistance must involve actual damage to the films. The measure of the damage can then be taken as a measure of the abrasion resistance. Their method was to subject the films to abrasive action that varied in intensity over the surface and that was, at its most intense point, sufficiently severe to remove completely the coating. The point at which the coating just stopped being

completely removed was then found. Of course the method is still relative in that a different standard must be set up for every thin-film combination, but it does permit comparison of the abrasion resistance of similar coatings, impossible with the previous method. The apparatus is shown in figure 12.1. It consists of a reciprocating arm carrying the abrasive pad of the Sira type, and is 0.25 in in diameter, loaded with 5.5 lb. The table carrying the sample under test rotates approximately once for every three strokes of the pad. The pad traces out a series of spirals on the surface of the sample and the geometry is arranged so that the diameter of the abraded area is approximately 1.25 in. The abrasion takes the form of a gradual fall off in intensity towards the outside of the circle, and the test is arranged to carry on for such a time that the central area of the coating is completely removed while the outside not at all. Holland and van Dam found that some 200 strokes were sufficient to do this with single layers of magnesium fluoride. They then defined the abrasion resistance measure of the coating by the formula

$$w = \left(\frac{d^2}{D^2} \right) \times 100\%$$

where d is the diameter of the circle where the coating has been completely removed and D is the diameter of the area that has been subjected to abrasion. Holland and van Dam studied particularly the case, as had Sira, of the single-layer magnesium fluoride antireflection coating for the visible region and they quote a wide range of most interesting results.

They investigated many different conditions of evaporation including angle of incidence and substrate temperature. A common value for the abrasion resistance of a typical magnesium fluoride layer of thickness to give antireflection in the green is between two and five, depending on the exact conditions of deposition. Best results were obtained when the substrate temperature during evaporation was 300 °C and the glow-discharge cleaning before coating lasted for 10 min. There was a significant reduction in abrasion resistance if either the temperature were allowed to drop to 260 °C or if there were only 5 min of glow-discharge cleaning. They also found that the abrasion resistance of the film is increased considerably by burnishing with a Selvyt cloth or by baking further at 400 °C in air after deposition. Another significant result obtained concerns the occurrence of a critical angle of vapour incidence during film deposition, beyond which the abrasion resistance falls off extremely rapidly. This critical angle varies slightly with film thickness but is approximately 40° for thicknesses in excess of 300 nm and rises as the thickness decreases.

The test appears never to have received general recognition in specifications. It should be extremely useful as a quality-control test in manufacture, especially as a reduction in quality can be detected long before it drops below the level of the normal abrasion test, and remedial action can be taken before any coatings are even rejected.

12.2.2 Adhesion

Adhesion has already been discussed in chapter 9. In the simplest type of adhesion test, a piece of adhesive tape is stuck down on the surface of the coating and pulled off. Whether or not this removes the film is taken as an indication of whether the adhesion of the film to the substrate is less than or greater than that of the tape to the film. The test is again of the go–no-go type.

It is important if consistent results are to be obtained that some precautions are taken in carrying out the test. The first is that the tape should have a consistent peel adhesion rating, which should be stated in the specification. Peel adhesion is measured by sticking a freshly cut piece of tape on a clean surface, usually metal, and then steadily pulling it off, normal to the surface. The tension per unit tape width, usually expressed in grams per inch, is the measure of the peel adhesion rating of the tape. The rating obtained in this way is usually virtually the same as the rating obtained when the tape is removed from a thin-film coating. Some precautions in applying the test are necessary. Fresh tape should always be used. The tape should be stuck firmly to the coating, exerting a little pressure and smoothing it down. It should be removed steadily, pulling it at right angles to the surface, and never snatched off, which would put an uncontrolled impulsive load on the film and would certainly lead to inconsistent results. The same thickness of tape should be used for all testing. With thicker tape of the same peel adhesion rating, the test would be slightly less severe. The width of the tape, however, does not matter. A rating which is often used is 1200 g in⁻¹ width. If necessary, the adhesion rating of any tape can easily be checked using a spring balance. For obvious reasons the test is often called the ‘Scotch Tape test’.

Attempts have been made to devise quantitative techniques for adhesion measurement and a number of these have also been discussed in greater detail in chapter 9. The simplest and most straightforward is the direct-pull test, involving the attachment of the flat end of a cylindrical pin to the coating, followed by measurement of the force necessary to pull it off. Provided the coating is detached with the pin, the force required divided by the area of the pin is then the measure of adhesion. An alternative test that has some advantages as well as disadvantages is the scratch test, in which a loaded stylus is drawn across the coating with gradually increasing load. At each stroke the coating is examined under a microscope for signs of damage. The load at which the coating is completely removed is taken as the measure of adhesion. The Goldstein and DeLong [3] technique involving the use of a microhardness tester as a scratch tester has also been mentioned in chapter 9.

12.2.3 Environmental Resistance

One further aspect of thin-film performance is also of very great importance. This is the resistance that the film assembly offers to environmental attack. Probably the most important aspect of the environmental performance of the filter is the

resistance to the effects of humidity but the resistance to other agents, such as temperature, vibration, shock, and corrosive fluids such as salt water, may all be important.

There are two possible approaches. Either the filter may be expected to operate satisfactorily while actually undergoing the test or it may only be expected to withstand the test conditions without suffering any permanent damage, although the performance need not be adequate during the actual application of the test. The latter is usual as far as interference filters are concerned, and in such a case the specification is known as a 'derangement specification' because it is sufficient that the performance is not permanently deranged by the application of the test conditions. In what follows we shall assume that the type of specification is the derangement type. Derangement specifications are easier to apply than the other type because the normal performance measuring equipment can be used remote from the environmental test chamber.

Of all the agents which are likely to cause damage, atmospheric moisture is probably the most dangerous. For most applications, particularly where severe environments are excluded, it will be found sufficient for the filter to be tested by exposing it for 24 h to an atmosphere of relative humidity $98\% \pm 2\%$ at a temperature of $50^\circ\text{C} \pm 2^\circ\text{C}$. It is often found that although the coatings are not removed by this test they are softened, and it is useful to carry out this test before the adhesion or abrasion-resistance tests, which can follow on immediately after.

A great deal of work has been carried out by government bodies on the environmental testing of equipment and components for the Services. This has resulted in specifications that are equivalent to the most severe conditions ever likely to be met in both tropical and polar climates. These specifications include in the UK DEF133 and DTD1085 for aircraft equipment. Relevant specifications in the USA include MIL-C-675, MIL-C-14806, MIL-C-48497 and MIL-M-13508. The tests vary from one specification to another but can include exposure to the effects of high humidity and temperature cycling over periods of 28 days, exposure conditions equivalent to dust storms, exposure to fungus attack, vibration and shock, exposure to salt, fog and rain, and immersion in salt water. It is not always possible for coatings to meet all tests in these specifications and concessions are often given if the coatings are to be enclosed within an instrument. Humidity and exposure to salt fog and water are particularly severe tests. Fungus does not normally represent as severe a problem to the coatings as it does to the substrates. Certain types of glass can be damaged by fungus and in such cases coatings, even if they themselves are not attacked, will suffer along with the substrates. Most instruments likely to be exposed to sand or dust are adequately sealed since their performance is likely to suffer if dust or sand is permitted to enter. Thus dust storms are usually a danger only to those elements with surfaces on the outside of an instrument.

References

- [1] Bousquet P and Richier R 1972 Etude du flux parasite transmis par un filtre optique à partir de la détermination de sa fonction de transfert *Opt. Commun.* **5** 27–30
- [2] Holland L and Dam E W v 1956 Wear resistance of magnesium fluoride films on glass *J. Opt. Soc. Am.* **46** 773–7
- [3] Goldstein I S and DeLong R 1982 Evaluation of microhardness and scratch testing for optical coatings *J. Vacuum Sci. Technol.* **20** 327–30

Chapter 13

System considerations: applications of filters and coatings

It is only rarely that thin-film filters or coatings are used by themselves. They usually form part of an optical system and it is in integrating coatings into such systems where many problems appear. There is an unfortunate tendency to leave coatings until late in the design process and some of the most severe problems occur during the attempted integration of coatings once the remainder of the design has been frozen. Such problems could frequently have been avoided had the incorporation of coatings been studied at a time when there was still some design flexibility.

Coatings cannot automatically be deposited with equal ease on any surface. Furthermore some tolerances must be permitted on coating performance. Then there is the shift in coating characteristics with angle of incidence, with temperature and with atmospheric humidity. Coatings often possess considerable intrinsic strain and the resulting stress can cause distortion that is significant in substrates of interferometric quality if they are not sufficiently thick. Lack of uniformity in coatings can also cause problems. Some of these difficulties arise from coating characteristics that show rapid change of phase with wavelength, characteristics frequently possessed by broadband reflectors. A lack of uniformity in the coating, if it is dielectric, is equivalent to a wavelength variation over the surface and if the phase dispersion is high then the resulting phase errors can be out of all proportion to the errors in thickness. The net result is an apparent loss of figure of the coated component that may show surprisingly large variations with wavelength. Extended-zone reflectors frequently exhibit rapid phase dispersion and so should be used with caution in applications where interferometric quality is required. All of these points have been discussed elsewhere in this book and the intention of repeating them here is simply to reinforce the point that coatings are like any other component and must be designed into the system as an integral part and not simply added at a later stage.

Coatings rarely stretch right to the edge of a substrate. Substrates must be held in jigs during coating and it is normal to do this by a lip that obscures the rim

of the substrate leaving an uncoated ring. This is not entirely a disadvantage. Delamination is always most likely to start at the edge of a coating and the uncoated rim around the coating gives it a much more regular edge and reduces the risk of delamination. Further, the mount for the component need not make contact with the coating where it could damage it and increase the chance of spontaneous delamination. The uncoated ring can, however, be a disadvantage if the component is a filter that rejects certain wavelength regions because stray light can leak through the uncoated part unless precautions to baffle it are taken. The uncoated area can be considerably reduced by the use of wire clips to hold the substrates by the edges during deposition, a technique frequently used with components such as sunglasses, but problems with stray light leakage can sometimes lead to the requirement that there should be no uncoated area whatsoever. The normal method for achieving this is to cut the component after coating. This should be carried out only if absolutely necessary. It increases the cost considerably because of the risk of failure involved in the cutting operation and it inevitably leaves a coating edge that is uneven on a microscopic scale and more likely to include stress concentrators that can initiate delamination.

It is always more difficult to coat a curved surface than a plane one and the difficulties increase with the curvature. Difficult coatings with tight tolerances should wherever possible be deposited on plane surfaces. Narrowband filters can be tuned to shorter wavelengths by tilting. If small tilts can be permitted (by the use of wedged holders for example) then the tolerances on peak wavelength can be relaxed.

Standard size components are always to be preferred. The manufacturer already has the necessary jigs and fixtures and the substrates are available in quantity. Fewer test runs are required and there are fewer unexpected difficulties. When something goes wrong with the process an entire batch of components is usually lost. Such failures are more likely with components of unusual shape or size, and so a greater number of uncoated components must be produced to ensure the correct number of final coated components. All of this means that the cost of nonstandard components is considerably greater than standard.

Most filters will consist of a series of components some of which are designed to reject radiation in regions outside the pass bands. Surprisingly disappointing performance can be achieved in cases where the rejected light is reflected rather than absorbed. We can illustrate this by considering two surfaces having reflectances and transmittances of R_1 , T_1 , R_2 and T_2 . Light can be considered as being reflected backwards and forwards between the surfaces and being combined incoherently. The net transmittance is then given by the expressions in section 2.14 (p 70) as:

$$T = \frac{T_1 T_2}{1 - R_1 R_2}.$$

If R_1 and R_2 are zero, that is, what is not transmitted is absorbed, then we have

the expected result

$$T = T_1 T_2.$$

However, if $R_1 = 1 - T_1$ and $R_2 = 1 - T_2$ then the result becomes similar to equation (2.140):

$$T = \frac{1}{(1/T_1) + (1/T_2) - 1}.$$

Consider the case where $T_1 = T_2 = 0.01$. The first expression gives $T = (0.01)^2 = 0.0001$, a very satisfactory figure, while the second expression gives

$$T = \frac{1}{100 + 100 - 1} = \frac{1}{199} = 0.005$$

very disappointing from the point of view of rejection. The solution is somehow to reduce the effect of R_1 and R_2 either by ensuring that the reflected beams rapidly walk out of the system aperture, by, for example, tilting the components relative to each other, or by placing absorbing components in between the two surfaces so that the beams are rapidly attenuated.

Sometimes reflecting and absorbing components will be combined in a system. Examples of this might be a heat-reflecting filter coating consisting of an interference shortwave pass filter deposited on a heat-absorbing glass or a narrowband filter consisting of an all-dielectric interference section, a metal–dielectric coating and an absorption glass. It is usually best in such cases to assemble the components such that the low-loss interference section faces the source. This ensures that the maximum amount of energy is rejected by reflection and minimises the temperature rise and possible resulting long-term damage. In the case of the narrowband filter assembly, the overall rejection performance of the filter is assisted by placing the absorbing glass component in between the two interference sections for the reasons discussed above.

Polarisation effects can sometimes be the cause of unexpected performance variation. We can illustrate this with the somewhat extreme case of a simple single-layer dielectric beamsplitter shown in figure 13.1. The performance of such a coating, assuming a quarter-wave (monitored at normal incidence) of zinc sulphide ($n = 2.35$) immersed in glass ($n = 1.52$) at an angle of incidence of 45° , is given by

$R_s = 33.15\%$	$R_p = 4.03\%$	$R_{\text{mean}} = 33.15\%$
$T_s = 66.85\%$	$T_p = 95.97\%$	$T_{\text{mean}} = 81.41\%$

Let us assume that the reflecting surface has a reflectance of 100% and calculate the irradiance of the output beam as a fraction of the input irradiance. A

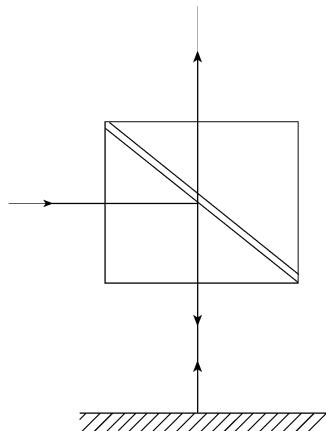


Figure 13.1. Arrangement of a single-layer dielectric beamsplitter used for calculation of efficiency discussed in the text.

simple calculation involves the unpolarised figures for T and R and yields $TR = (18.59\% \times 81.41\%) = 15.13\%$. However, this calculation has taken no account of the polarising effect of the beam splitter itself. The true figure for unpolarised incident light should be $0.5(R_s T_s + R_p T_p) = 13.01\%$ (a difference greater than 10% of the previous figure). Polarisation of the input beam alters the results still further. With s-polarised input light the figure would be $R_s T_s = 22.16\%$ while with p-polarised light it would be as low as $R_p T_p = 3.87\%$. Thus with varying degrees of polarisation of the input light the efficiency of the system can vary from 3.87% to 22.16%. To avoid performance fluctuations resulting from such effects, a quarter-wave plate with axis at 45° to the plane of incidence is often inserted in the input side of a system to convert both s- and p-polarised light to circularly polarised, which makes the overall performance of the system equivalent to unpolarised light. (It is unlikely that the input light should be already circularly polarised, but of course in that case the quarter-wave plate could make the situation worse.) Metal layers suffer less from polarisation effects, but they, too, do still have significant polarisation-sensitive behaviour.

That was an example of an immersed coating. Note that immersed coatings always have very high effective angles of incidence since the important quantity for Snell's law is $n_0 \sin \vartheta_0$ rather than ϑ_0 . Thus, in immersed coatings, angle-of-incidence effects are invariably enhanced. Polarisation effects are particularly pronounced but so also are the simple wavelength shifts associated with a change in angle of incidence.

Even in coatings that are not immersed, the changes in angle of incidence associated with a highly divergent or convergent beam can cause problems, especially if the component is tilted with respect to the axis. Sometimes the problems can be eased by deliberately introducing a variation in coating thickness

over the surface of the component. This can be particularly effective when a point source is used close to a component when the small source dimensions ensure that only a small range of angles of incidence correspond to each point on the component surface.

A point to watch concerns polarisation effects associated with skew rays. p- and s-polarisation performance is calculated with respect to the plane of incidence. A skew ray possesses a plane of incidence that is usually rotated with respect to the principal plane of incidence containing the axial ray of the system. This can cause problems in large aperture polarisers, for example, where, although the s-transmittance for the skew rays can be very low, the corresponding plane of polarisation is actually rotated and can lead to an appreciably large leakage of light which is s-polarised with reference to the plane of incidence of the axial ray. As a rough example we can consider a cone of 1° half-angle incident at 45° on a polarising beam splitter. The plane of incidence of the marginal azimuthal rays will be rotated at an angle of approximately $1^\circ / \sin 45^\circ$, or 1.4° with respect to the plane of incidence of the axial ray. Let us assume that both axial ray and marginal ray have zero transmittance for s-polarised light and unity for p-polarised light. Because of the rotation of the plane of incidence the effective transmittance of the marginal ray in the s-plane of the axial ray will then be $\sin^2(1.4^\circ)$ or 0.06%.

A very useful account of problems associated with the integration of thin-film coatings into optical systems has been written by Matteucci and Baumeister [1].

13.1 Potential energy grasp of interference filters

It is worthwhile considering why interference filters are used in preference to other types of wavelength-selecting devices such as prism and grating monochromators. Of course the size and mechanical stability of the thin-film filter are in themselves powerful arguments in favour of its use, and, especially in cases where space and weight are at a premium, in satellite-borne instruments for example, they are probably sufficient. However, there is an even more compelling reason for adopting thin-film filters and this is the greatly increased potential grasp of energy over dispersive systems.

Compared with a grating monochromator, for instance, the thin-film filter with the same bandwidth is capable, provided the rest of the system is correctly designed round it, of collecting several hundred, and in some cases thousand, times the amount of energy collected by the monochromator. This section, therefore, is devoted to a comparison of the interference filter with the diffraction grating, particularly from the point of view of the potential total energy grasp.

In order to compare the energy-gathering properties of various components, we have to assume that each is used in an ideal system designed to make maximum use of its energy-gathering powers, that the bandwidths of the various systems are equal, and that any dispersive components are used well within their

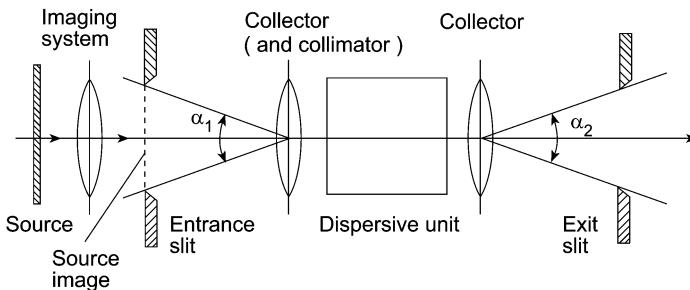


Figure 13.2. An idealised dispersive monochromator. (After Jacquinot [2].)

limiting resolutions so that their response functions are not complicated by large diffraction effects. We shall also assume that the source of illuminations is of equal brightness in all cases and that the collecting condensing optics are such that the entrance apertures of all systems are completely filled. The energy grasp under these conditions is then computed in each case as a function of the appropriate area of the component, and the comparison made on the basis of these figures.

In fact this analysis has been carried out by Jacquinot [2] for a diffraction grating, a prism and a Fabry–Perot interferometer. He has shown first that there is always a clear advantage in using a diffraction grating rather than a prism, the advantage varying from around three to perhaps 100 with the dispersion of the prism materials. Because of this, the comparison that primarily concerns us is between the interference filter and the diffraction grating. Jacquinot has also compared the Fabry–Perot interferometer having an air spacer with the diffraction grating, and showed that there is a clear gain of 300–400 times in the energy grasp of an interferometer over a grating of the same area. The case of an interference filter is similar but the spacer layer has an index appreciably greater than unity, especially in the infrared, which increases its grasp still further. In the analysis below, we shall follow the main lines of Jacquinot's argument, but shall extend the analysis to include a spacer of index other than unity.

Jacquinot considers a spectrometer consisting of an input slit, a collector and collimator of some description, a dispersive element which here is a grating, and an output element imaging the entrance slit on the exit slit, the final element in the system. It is assumed that the resolution is limited by the width of the slits and that the grating is capable of higher resolution if required. This means that we can define the resolution purely in terms of slit width and dispersion. In this condition the maximum luminosity for a given resolution will be achieved when the entrance and exit slit widths, expressed in terms of spectral interval, are equal, when a triangular response function will be obtained from the instrument. It is assumed that the source, which is an extended one, is monochromatic and of uniform brightness.

There will be some sort of imaging system which will produce an image of

the source on the entrance slit. The brightness of the source image will be equal to that of the actual source, except for the transmission of the imaging system, which we can take to be unity without affecting the final result, since all systems to be compared will have a similar arrangement before the entrance aperture. Given that the brightness of the image is identical to that of the source, it only remains for the aperture of the imaging system to be made large enough for the aperture of the collector and collimator before the grating to be completely filled. Again we can assume that this has been carried out in all arrangements without any loss in generality. The situation is sketched in figure 13.2. The notation used here is, as far as possible, exactly that used by Jacquinot in his original paper to make the comparison easier. Let the brightness of the source image be denoted by B . Let the monochromator be adjusted so that the image of the entrance slit falls directly on the exit slit and let both slits have the same width and length. This corresponds to the apex of the triangle. The energy transmitted by the system will be given by

$$E = BS\omega T$$

where ω is the solid angle subtended by either slit at the appropriate collector element and S the area of the beam at the collector. $S\omega$ will be the same for both the entrance and the exit slit since we have arranged for the image of one to coincide with the other. T is the transmittance of the monochromator. If the width of the exit slit is α_2 and the length β_2 , then the expression becomes

$$E = BST\beta_2\alpha_2.$$

If we denote the resolving power of the system by R , then we have that $\alpha_2 = \lambda D_2/R$ where D_2 is the angular dispersion of the system referred to the output slit, i.e.

$$E = BST\beta_2(\lambda D_2/R).$$

For the grating monochromator the angular dispersion is derived from the equation

$$\sigma(\sin i_1 + \sin i_2) = m\lambda$$

where σ is the grating constant, i.e. the interval between grooves, m is the order number, and i_1 and i_2 are the angles of incidence and diffraction, respectively, at the grating.

$$D_2 = \frac{di_2}{d\lambda} = \frac{m}{\sigma \cos i_2} = \frac{\sin i_1 + \sin i_2}{\lambda \cos i_2}$$

i.e.

$$\lambda D_2 = \frac{\sin i_1 + \sin i_2}{\cos i_2}.$$

Now

$$S = A \cos i_2$$

where A is the area of the grating and we assume that it is completely illuminated and that no light is lost, so that

$$S\lambda D_2 = A (\sin i_1 + \sin i_2).$$

Jacquinot shows that $S\lambda D_2$ is a maximum for the Littrow mounting (where i_1 and i_2 are as nearly equal as possible) used on the blaze angle which we denote by φ . For that mounting

$$S\lambda D_2 = 2A \sin \varphi$$

and

$$E = (BT\beta_2/R) 2A \sin \varphi.$$

φ we can take as 30° , say, when $\sin \varphi = \frac{1}{2}$ and

$$E = BT\beta_2 A / R.$$

We shall now consider the interference filter and compare it with the diffraction grating. The case considered by Jacquinot is that of the conventional Fabry–Perot interferometer made up of a pair of plates in an etalon with a spacer of unity refractive index. Here we are more concerned with the interference filter where the spacer layer has an index greater than unity. As on p 284, we introduce the concept of an effective index of refraction which governs the angular behaviour of the filter. We shall use a similar analysis to that of Jacquinot, but recast it in the form of the results of chapter 7.

Jacquinot suggests that the filters be used with an acceptance angle such as to make the effective bandwidth of the filter $\sqrt{2} \times$ the value at normal incidence. Equation (7.40) gives

$$W_\Theta^2 = W_0^2 + (\Delta\nu')^2$$

where W_0 and W_Θ are the halfwidths corresponding to collimated light at normal incidence and to a cone of semiangle Θ . If Θ is measured in air then

$$\Delta\nu' = v_0 \Theta^2 / 2n^{*2}.$$

For $W_\Theta = \sqrt{2}W_0$ we must have $W_0 = \Delta\nu'$, i.e.

$$W_0 = v_0 \Theta^2 / 2n^{*2}$$

and, from equation (7.41),

$$\hat{T}_\Theta = (W_0 / \Delta\nu') \tan^{-1} (\Delta\nu' / W_0) = \tan^{-1}(1) = \pi/4 = 0.78 \quad (13.1)$$

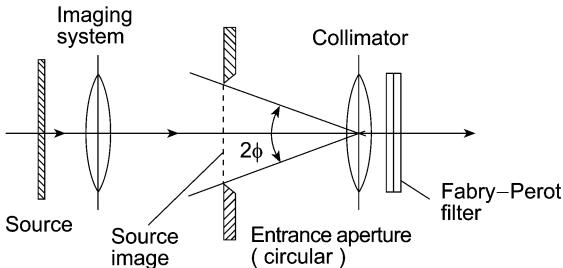


Figure 13.3. An arrangement of a monochromator using an interference filter.

where \hat{T}_Θ is the effective peak transmittance of the filter for a cone of incident light of semiangle Θ referred to the incident medium, which we are assuming is air.

If R_0 is the resolving power for perfectly collimated light at normal incidence and R_Θ that for a cone of semiangle Θ , then

$$R_0 = v_0/W_0$$

and since $\Delta\nu'$ is small compared with v_0

$$R_\Theta = v_0/W_\Theta = R_0/\sqrt{2}.$$

But $W_0 = \Delta\nu'$ so that

$$R_0 = v_0/\Delta\nu' = 2n^{*2}/\Theta^2$$

and so

$$\Theta^2 = \sqrt{2}n^{*2}/R_\Theta. \quad (13.2)$$

If B is again the brightness of the source and A is that area of the filter that is fully illuminated, then the energy collected will be

$$E = BAT(\pi/4)\omega \quad (13.3)$$

where ω is the solid angle subtended by the aperture and T is the normal incidence transmittance. The factor $(\pi/4)$ is included from (13.1). From figure 13.3

$$\omega = 2\pi(1 - \cos\Theta) \approx \pi\Theta^2. \quad (13.4)$$

Then, from equations (13.2), (13.3) and (13.4),

$$E = BAT \frac{\pi^2}{2} \frac{n^{*2}}{\sqrt{2}R_\Theta}.$$

This is similar to the form given by Jacquinot except for the factor n^{*2} which is missing in his expression.

We are now in a position to compare efficiencies. The relative energy grasp of the two systems is

$$\frac{E_{\text{filter}}}{E_{\text{grating}}} = \frac{BAT (\pi^2/2) n^{*2} / (R\sqrt{2})}{BT\beta_2 A/R}. \quad (13.5)$$

We can assume for this comparison that the resolution and areas and transmittances of the two systems are equal (that is transmittance at normal incidence in collimated light for the interference filter). Equation (13.5) then simplifies to

$$\frac{E_{\text{filter}}}{E_{\text{grating}}} = \left(\frac{\pi}{2\sqrt{2}}\right) \left(\frac{n^{*2}}{\beta_2}\right) = 3.4 \frac{n^{*2}}{\beta}.$$

Jacquinot estimates the usual value of β to be 0.01 radian. With extreme care in design, values of 0.1 have been achieved, although this represents the very limit. For an n^* of unity, then, the value of the energy ratio varies between 34 and 340.

However, n^* in the visible region is usually in excess of 1.5, which alters the range to 76–760. For the infrared the advantage of the filter is even greater, for n^* is usually of the order of 3.0, so that the range becomes 306–3060, a massive advantage. This means that we can happily make the interference filter much smaller than the grating and still have a very significant increase in energy grasp over it.

This analysis dealt with the single Fabry–Perot type of filter. The advantage with the DHW type of filter is slightly greater still, since the effective transmittance in a cone of illumination is higher than that of the Fabry–Perot.

13.2 Narrowband filters in astronomy

The problem of detecting faint astronomical objects is rendered even more difficult, than it would otherwise be, by the light of the night sky. This light consists mainly of starlight scattered by dust both in the atmosphere and in interstellar space (including light from our own sun) together with emission from the upper atmosphere, and may be considered to be mainly of a continuous spectral nature although there are a number of emission lines as well. The sky light causes an overall fogging of the photographic plates, which are the most common detectors used in this work (although in recent years increasing use has been made of image tubes). Maximum contrast between the photographic image of a star or other object and the sky background is obtained when the sky fog is just apparent on the plate. The exposure time is chosen to give just this amount of fogging. The efficiency of the photographic detector falls off rapidly on either side of this optimum. The limit of detection of a faint object will be reached when the image is just discernible against the background.

The way in which the limit of detection varies with the parameters of the system has been studied particularly by Baum [3]. A simplified account of the analysis is given by Bowen [4] and it is this latter form that we follow here. The notation used by Bowen, which we also use here, differs slightly from that used by Baum.

The signal which is received from the object will consist of discrete photons arriving at a constant mean rate but randomly spaced. Provided the mean rate is sufficiently small (satisfied for the signals we are considering) we can consider the photons as forming a Poisson distribution (the distribution which deals with sequences of events where the probability of an occurrence in any particular time interval is vanishingly small, but where the total observing time is sufficiently long to ensure a finite number of events). For the Poisson distribution the standard deviation of successive measures of the number of photons N arriving in a certain constant time is simply \sqrt{N} .

Let D be the telescope aperture diameter, f the focal length of the telescope, t the observation time, β the diameter of the image of the object, n the number of photons from the object received per unit area of telescope aperture per second, s the number of background photons received per unit area of telescope aperture per unit solid angle of sky per second, p the limit of linear resolution of the emulsion, q the quantum efficiency of the entire system which includes the photographic emulsion and the transmission of the optical system, and m the number of photons recorded per unit area of photographic plate which will produce the correct level of background fog.

In his paper, Bowen defines the faintness of a star or object as $1/n$. We shall now examine the way in which the limiting detectable faintness varies with the parameters of the system. The fractional error in a measurement is denoted by B and is defined as the standard deviation associated with the measurement divided by the measurement itself. Thus in a measurement of a number of photons N , the fractional error would be $B = (\sqrt{N})/N = 1/\sqrt{N}$.

The number of photons recorded from the object and from an equal area of sky in time t is given by

$$D^2ntq + \beta^2sD^2tq$$

where we are omitting factors of $\pi/4$. The standard deviation in successive measurements will be

$$\left(D^2ntq + \beta^2sD^2tq \right)^{1/2}$$

and the fractional error in the measurement will be

$$\begin{aligned} B &= \frac{\left(D^2ntq + \beta^2sD^2tq \right)^{1/2}}{D^2ntq} \\ &= \frac{\left(n + \beta^2s \right)^{1/2}}{Dnt^{1/2}q^{1/2}}. \end{aligned}$$

For very faint objects, $n \ll \beta^2$ so that

$$B = \frac{\beta s^{1/2}}{Dnt^{1/2}q^{1/2}} \quad (13.6)$$

and the limiting faintness is given by

$$\left(\frac{1}{n}\right)_1 = \frac{B_1 Dt^{1/2} q^{1/2}}{\beta s^{1/2}} \quad (13.7)$$

where B_1 is the highest possible value of B where the object is still just detectable. Bowen suggests that B_1 should be 0.2. This formula applies as it stands to photoelectric detectors and shows how the faintness which can be detected increases with increasing aperture. For the photographic detector, however, the position is not quite the same. Here the time of exposure t must be chosen to give the correct background fog. The efficiency of the plate drops so quickly if the density of the background is incorrect that any other exposure time is of very much less value. This correct exposure time t_0 is given by

$$D^2 t_0 s q = m f^2$$

i.e.

$$t_0 = \frac{m f^2}{D^2 s q}$$

and, substituting in equation (13.7),

$$\left(\frac{1}{n}\right)_1 = \frac{B_1 D q^{1/2}}{\beta s^{1/2}} \sqrt{\left(\frac{m f^2}{D^2 s q}\right)} = \frac{B_1 m^{1/2} f}{\beta s}. \quad (13.8)$$

In the equation we are assuming that β is larger than the resolution limit of the plate. If this is not the case, where f is small for example, then β must be replaced by p/f , giving

$$\left(\frac{1}{n}\right)_1 = \frac{B_1 m^{1/2} f^2}{p s}. \quad (13.9)$$

These results, obtained by Bowen, are not what we might have expected, because they seem to show that the all-important parameter for photographic detection of faint objects is the focal length of the telescope and not the aperture. The longer the focal length, the greater the faintness which may be observed, regardless of the diameter of the aperture of the system. So far, however, we have neglected to notice that observation time is limited to one night. Increasing the focal length without a corresponding increase in aperture increases the necessary exposure time, which varies as the square of the focal length. Let t_m be the longest

allowable exposure time. Then, for any given system, the largest value of focal length f_m will be given by

$$f_m^2 = \frac{t_m D^2 s q}{m} \quad \text{i.e.} \quad f_m = \frac{t_m^{1/2} D s^{1/2} q^{1/2}}{m^{1/2}} \quad (13.10)$$

which when substituted in equation (13.8) and (13.9) gives for f large or β large

$$\left(\frac{1}{n}\right)_1 = \frac{B_1 t_m^{1/2} D q^{1/2}}{\beta s^{1/2}} \quad (13.11)$$

and for f small and β small

$$\left(\frac{1}{n}\right)_1 = \frac{B_1 m^{1/2} t_m D^2 s q}{psm} = \frac{B_1 t_m D^2 q}{pm^{1/2}}. \quad (13.12)$$

These expressions¹ show that, indeed as might be expected, there is a gain in going to larger telescopes.

Given the maximum possible value of D and f , how can the situation be improved by the use of filters? If there is a difference in spectral distribution of the radiation from the object and the sky background, then it is possible that a filter inserted in the system might modify the ratio of photons received from the object to those received from the sky. If this process results in a reduction in n by a factor x to xn , and a reduction in s to ys , then the ratio n/s becomes xn/ys , and if x/y is sufficiently large, then a positive gain in faintness may result. Substituting these values in the expression for the case where the resolution of the emulsion is not the limiting factor, equation (13.11) becomes

$$\left(\frac{1}{n}\right)_1 = \frac{x B_1 t_m^{1/2} D q^{1/2}}{y^{1/2} \beta s^{1/2}}$$

and, assuming we adjust the focal length of the system as before to give the longest exposure time t_m , then the result is obtained that a gain in $1/n$ is achievable provided that $x > \sqrt{y}^2$.

For the case where the emulsion resolution is a limiting factor, the expression (13.12) for $1/n$ shows that there is no possibility of altering the situation by filtering. The filtering will work only when the object is extended, or when the focal length of the telescope is large enough, or when the grain of the plates is fine enough, to ensure that the plate resolution is not a limiting factor.

The great bulk of the sky light is scattered light which has a more or less continuous spectrum. Only the emission from the upper atmosphere has a

¹ Reciprocity failure, which effectively means that q is reduced slightly as t increases, has been neglected in the derivation.

² This, at first sight, odd result follows from the assumption made early in the derivation that the object is faint so that $n \ll \beta^2 s$.

component consisting of discrete lines. Since, for a gain due to filtering, it is not sufficient to ensure that $x > y$ but that $x > \sqrt{y}$, in cases where n has a continuous spectral distribution and there is no great difference between the distributions of n and s , there is probably very little to be gained by filtering. In fact, slight enhancement of the ratio of detected photons accompanied by a drop in transmittance could lead to a loss in performance rather than a gain. However, there are classes of objects which are characterised by line spectra and in these cases it is possible by using filters centred on the lines to retain n only slightly reduced, but to have s greatly reduced. Such a class of objects is the hydrogen emission nebulae. It is now known that hydrogen is one of the elements of interstellar gas—probably the most abundant. Where hydrogen clouds are near bright stars, the atomic hydrogen is ionised by the x-ray and extreme ultraviolet radiation from the stars, and, when the electrons and protons recombine, the characteristic hydrogen spectra are produced. The principal line emitted in the wavelength range detectable at the surface of the Earth is the first line of the Balmer series, H_α at 656.3 nm, which, although not always the brightest line, is the one where contrast can be greatly improved.

The use of an interference filter centred on 656.3 nm greatly increases the contrast between the nebulae and the night sky, and gives a large increase in the faintness of nebulae which can be detected.

Equation (13.10) shows that when the interference filter is installed the focal ratio of the telescope must be adjusted to give the correct level of background fog.

$$\frac{f}{D} = \frac{t_m^{1/2}}{m} (ys)^{1/2} q^{1/2}.$$

Generally, with typical interference filters, the focal ratio should be near unity. Such a focal ratio incident directly on a narrowband interference filter would have a disastrous effect on both the bandwidth and peak transmission. However, the optical arrangement of the big telescopes permits an alternative arrangement. The primary mirror of a large telescope usually produces a pencil of focal ratio around $f/5$. As we have seen in chapter 7, a narrowband filter for the visible region with a bandwidth of around 1% of peak wavelength will accept such a pencil quite satisfactorily, and it is usual to insert the interference filter at or very near the prime focus. Beyond the prime focus a camera is installed which reduces the focal ratio of the system to the desired value. The arrangement is shown in figure 13.4(a). With this layout the variation with field angle of the pass band of the filter (due to angle of incidence variation) is kept very small. If necessary it could be eliminated altogether by use of an extra lens, as in figure 13.4(b).

In figure 13.4, the filter acts as a field stop and may limit the field of view of the instrument. Filters up to 6 in in diameter have been constructed, although 4 in is probably a more usual figure. Filters with a diameter of 2 in are readily available.

Some particularly fine examples of photographs taken with relatively broad combinations of coloured-glass filters and ones with interference filters of

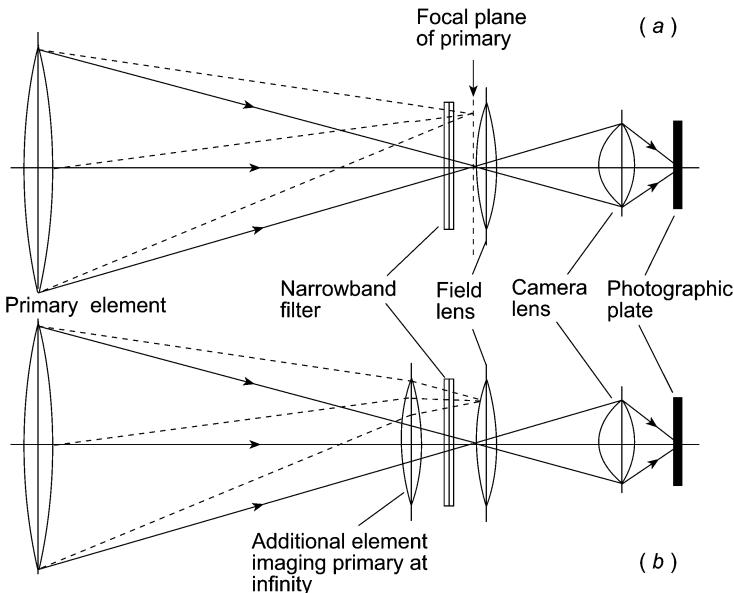
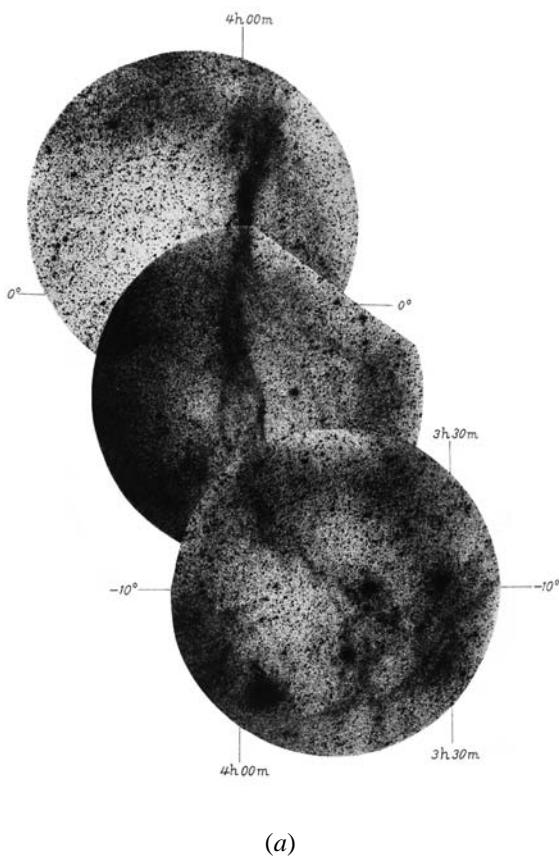


Figure 13.4. A narrowband filter in an astronomical telescope. The primary is shown here as a lens, but in the big telescopes would usually be a mirror. If necessary an additional element can be added as in (b) to alter the inclination or the off-axis pencils so that the effective peak wavelength of the filter is constant over the entire field.

very much narrower bandwidths are given by Courtes [5]. Ring was the first successfully to use all-dielectric filters for this purpose, pioneering the development of these filters in the UK, and a paper by him [6] includes several photographs. A paper by Meaburn [7], who took the excellent photographs in figure 13.5 illustrates extremely well the type of problem solved by interference filters and is well worth reading. Since this section appeared in the first edition, a particularly useful book by Meaburn [8] has been published and should be consulted for further information.

13.3 Atmospheric temperature sounding

In the mid-1960s work began on a series of radiometers to be flown in satellites with the aim of measuring the distribution of temperature in the upper atmosphere. This programme was extremely successful. The first of these radiometers was designed by a joint team from the Universities of Oxford and Reading in the UK, the team at Reading moving to Heriot-Watt University at a late stage of the project. The radiometer was flown in the Nimbus IV spacecraft. The radiometer was known as the selective chopper radiometer (SCR) because of the basic principles

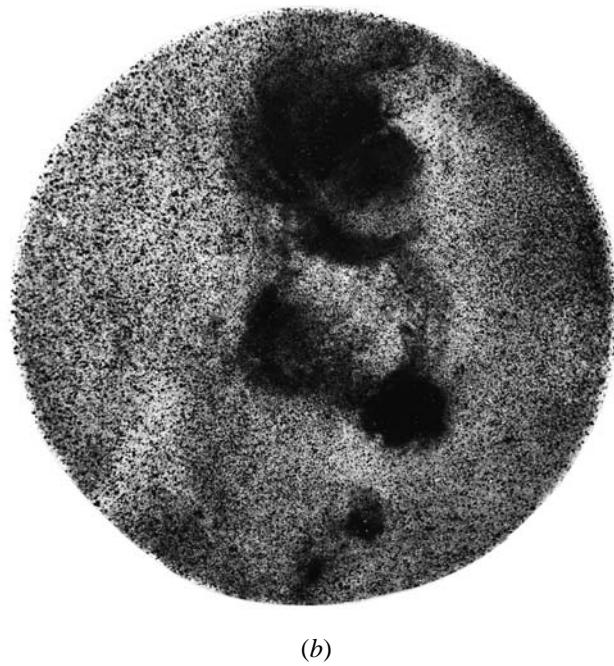


(a)

Figure 13.5. (a) Nebulosities in the Cetus arc. H_{α} photographs of 1-h exposure taken on a 6-in $f/1$ Schmidt camera through a 4-nm bandwidth filter. (After Meaburn [7].) (b) Nebulosities in the galactic anti-centre. Photograph taken through a 4-nm bandwidth filter centred on H_{α} (656.3 nm) with a 6-in aperture Schmidt camera. The exposure was 1.75 h. (Courtesy of Dr J Meaburn.)

of its design and it made extensive use of filters. It made measurements, with a height resolution of 10 km, of the temperature of that part of the atmosphere of height between 15 and 50 km, that is the troposphere and part of the stratosphere. The basic method used in the SCR and in other subsequent radiometers for temperature sounding is the detection and measurement of the radiation from atmospheric carbon dioxide.

Some ideas of the temperature structure of the atmosphere had already been formed, typical temperatures being of the order of 200 K at a height of 10 km rising to 240–280 K at heights of around 50 km. The peak of the black-body curve for a temperature of 200 K lies at a wavelength of 15 μm , while that for



(b)

Figure 13.5. (Continued)

280 K is at 11 μm . The most favourable wavelength region for the measurement of the temperature of the atmosphere by detection of emitted radiation is therefore the band 11–15 μm . Of course the atmosphere will emit radiation only in the regions where it absorbs (the equivalence of absorptance and emittance is a basic physical principle) and this, coupled with the fact that the radiation emitted from a given level must traverse the remainder of the atmosphere above that level to reach the detector in the spacecraft, allows an ingenious method to be used for the deduction of the temperature structure which was first suggested by Kaplan [9].

Carbon dioxide is evenly distributed in the atmosphere and has extensive absorption bands around 15 μm so that it can be used as an indicator of the temperature of the atmosphere as a whole. Fortunately, over most of the important region, carbon dioxide is the only constituent of the atmosphere showing absorption (water vapour would interfere but is important only near the ground, and O₃ at 14 μm in the 25–40 km region can be avoided) which simplifies considerably the calculations. The absorption spectrum of CO₂ consists, at very low pressures, of a number of discrete lines which become gradually broader with increasing pressure. The detector in the spacecraft is arranged so that it responds to only a very narrow band of wavelengths in the CO₂ spectrum. If a waveband is chosen within which the absorption is high, then the radiation emitted at the bottom of the atmosphere will not reach space because the transmission of the

atmosphere above it is low. At greater heights a much greater proportion of the energy emitted will reach the detector.

However, also at greater heights, the energy emitted by the atmosphere will fall, because of decreasing density and pressure of CO₂, and, at a height which will depend on the absorption within the particular waveband chosen, the second process will overtake the first with the result that a major portion of the energy received by the detector will emanate from a narrow range of depths in the atmosphere. The mean depth can be changed by varying the centre wavelength of the band which is being detected, and so altering the variation of absorption with height. The experiment and apparatus are described in various articles [10–14].

The following account is a much simplified version which follows directly work by John Houghton (now Sir John). First we find the emittance of any layer by calculating the absorptance which is equivalent to the emittance. Consider a layer of the atmosphere situated at a depth z below the spacecraft. Let the transmittance of the atmosphere, at frequency ν , above this layer be T_z . In passing through a layer of thickness dz of the atmosphere the fractional intensity lost by unit intensity of radiation will be the absorptance of the layer. Next, consider radiation of initial intensity F at frequency ν at depth z . The fraction of this which appears at the detector in the spacecraft will be either FT_z , or $(F - dF)T_{(z-dz)}$ and as these quantities will be equal we can write

$$(F - dF)T_{(z-dz)} = FT_z.$$

With some adjustment we find

$$A_{dz} = \frac{dF}{F} = \frac{T_z - T_{(z-dz)}}{T_{(z-dz)}} = \frac{-(dT_z/dz)dz}{T_{(z-dz)}}$$

where A_{dz} is the absorptance and hence emittance of the layer. If \mathcal{T} is the mean temperature of the layer, then the black-body emission per unit frequency interval associated with it will be given by $B(\mathcal{T})$ at frequency ν . The energy actually given out by the layer will be given by this expression multiplied by the emittance, i.e.

$$dI_z = K T_{(z-dz)} A_{dz} B_\nu(\mathcal{T})$$

where dI_z is the energy per unit frequency interval received by the radiometer which emanates from a layer of thickness dz at depth z and K is a constant.

Then

$$dI_z = -K \frac{dT_z}{dz} B_\nu(\mathcal{T}) dz dv.$$

If the detector in the spacecraft has a bandwidth of $\Delta\nu$, then the expression for the energy over this band becomes

$$\int_{\Delta\nu} dI_z dv = \int_{\Delta\nu} -K \frac{dT_z}{dz} B_\nu(\mathcal{T}) dz dv$$

and if R_ν/K is the response of the radiometer at frequency ν then the output of the instrument will be given by

$$D_z/dz = \int_{\Delta\nu} -R_\nu \frac{dT_z}{dz} B_\nu(\mathcal{T}) dz d\nu.$$

We can choose the frequency interval $\Delta\nu$ small enough for $B_\nu(\mathcal{T})$ to be a constant over the interval. $B_\nu(\mathcal{T})dz$ can then be moved outside the integral sign. What is left is the function

$$W_z = \int_{\Delta\nu} -R_\nu \frac{dT_z}{dz} d\nu$$

which is known as the weighting function, and represents the response of the system to radiation from depth z . We shall now look a little closer at the form of the weighting function, assuming that a single isolated absorption line is involved.

The absorption coefficient k_ν for radiation of frequency ν is defined by the relationship

$$dI_\nu = -k_z I_z du$$

where dI_ν is the change in intensity I_ν after traversing path length du of the absorbing gas. u is measured in terms of the quantity of gas traversed rather than physical distance and has such units as g cm^{-2} or atmo-cm (the equivalent path length in the gas at normal atmospheric pressure and temperature). The strength of the line S is defined as the absorption coefficient integrated over the whole width of the line.

For radiation of wavenumber ν near the centre of a single gaseous absorption line, k_ν , is given by the Lorentz formula for pressure broadening³:

$$k_\nu = \left(\frac{S}{\pi} \right) \left(\frac{\gamma}{(\nu - \nu_0)^2 + \gamma^2} \right).$$

γ is the halfwidth of the line, which is proportional to pressure and can be written $\gamma = \gamma_0(p/p_0)$. (γ is also inversely proportional to the square root of the absolute temperature, but, as this exhibits much less variation than pressure through the part of the atmosphere which we are considering, we can omit temperature from the calculation.)

If the frequency ν is such that $\gamma^2 \ll (\nu - \nu_0)^2$ then we can write

$$k_\nu = \left(\frac{S}{\pi} \right) \left(\frac{\gamma_0 p}{p_0 (\nu - \nu_0)^2} \right) = \beta p.$$

³ See for example p 47 of Houghton J and Smith S D 1966 *Infra-Red Physics* (Oxford: Oxford University Press).

Now CO₂ is uniformly mixed through the atmosphere so that the mass of CO₂ per unit area between the top of the atmosphere and depth z will be proportional to the atmospheric pressure at that depth, i.e.

$$u = cp$$

where c is a constant. The transmittance of the atmosphere above depth z , at which the pressure is p , will therefore be

$$\begin{aligned} T_z &= \exp\left(-\int_{p=0}^p k_\nu du\right) \\ &= \exp\left(-\int_{p=0}^p ck_\nu dp\right) \\ &= \exp\left(-\frac{1}{2}c\beta p^2\right). \end{aligned}$$

To simplify the analysis we can assume that p varies linearly with z , i.e.

$$p = fz$$

(or alternatively we could use p as the measure of the depth z since it is a single-valued function of z which increases continuously with z). The weighting function for a single monochromatic line of frequency ν , assuming that $R = 1$, is then

$$W_z = -\frac{dT_z}{dz} = \beta cf^2 z \exp\left(-\frac{1}{2}\beta f^2 cz^2\right).$$

The form of this function is shown in figure 13.6. For the purposes of drawing this, a new variable $y = (\frac{1}{2}\beta f^2 c)^{1/2} z$ has been introduced so that

$$-\frac{dT_z}{dz} = \left(2\beta cf^2\right)^{1/2} y e^{-y^2}$$

and the function which is actually plotted in figure 13.6 is ye^{-y^2} .

By choosing the appropriate wavelength, the form of the variation of the absorption coefficient can, to some extent, be controlled and the position of the maximum in terms of the height, or rather depth, varied. The absorption spectrum of CO₂, at 15 μm consists of a large number of separated lines. The teams at Oxford and Reading have made a special study of these, tabulating the positions and strengths, and have been able to choose a series of wavelengths to permit examination of the temperature structure of the atmosphere between 15 and 50 km with a resolution of 10 km.

One of the difficulties which exist is the finite bandwidth of the radiometer. The bandwidths of practical filters cannot be made arbitrarily small and, because

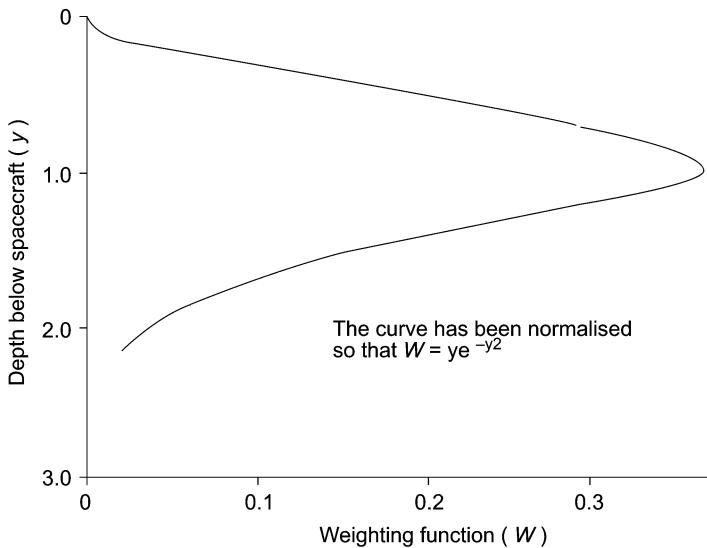


Figure 13.6. The form of the radiometer weighting function.

the CO₂ absorption coefficient varies with wavelength, the bandwidth of the radiometer will cause a reduction in the height resolution. For the channels designed to look deep into the atmosphere, the bandwidth does not affect the result too much and can be 10 cm⁻¹—well within the capabilities of an interference filter. The channels designed to look at the top of the atmosphere, however, must be positioned on the centres of the most intense lines, the Q-branch at 667 cm⁻¹, and the bandwidth must not effectively be greater than 1 cm⁻¹. This is beyond the current state of the art at 15 μm. The ingenious solution that has been adopted and gives the radiometer its name is the use of a chopper filled with CO₂.

To explain the action of this selective chopper we shall first consider the operation of the simpler channels with the acceptable filters. In these channels, partly to ensure that the noise in the electronics is sufficiently low, and partly to ensure that the radiometer registers radiation from the atmosphere only and not from the components of the radiometer itself, which will all be emitting at 15 μm, a chopper is placed in the entrance aperture. Radiation emanating from the atmosphere will be chopped, while radiation from the radiometer itself will not and will escape detection. Of course the chopper will also radiate and so the usual method of alternately inserting and removing a blanking shutter in front of the radiometer entrance aperture would be quite useless, because the radiation from the shutter would also be chopped and detected along with the signal. The method which is used is extremely neat. The entrance aperture of each channel is divided into two equal parts so that one-half of the aperture views, reflected in a fixed mirror, deep space, which can be assumed to be at a temperature of absolute zero

and to represent a reference of zero radiation provided the reflectance of the mirror is sufficiently high, while the other half views the Earth's atmosphere reflected in a second mirror, which can be varied in position for calibration purposes. A chopper consisting of a vibrating black blade is arranged so that it obscures the fixed and variable mirrors alternately and, therefore, effectively chops the incoming signal. The radiation from the chopper blade is not detected because the blade remains within the aperture of the system all the time.

The selective chopper channel of the radiometer is similar to these other channels. However, a narrower filter is used, having a bandwidth of 3.2 cm^{-1} at 667 cm^{-1} , which is the narrowest yet obtainable at this frequency. In addition, a cell containing CO_2 is included in front of each section of the entrance aperture and the black blade of the chopper is exchanged for a mirror which looks at deep space. If the chopper mirror were completely removed, both parts of the entrance aperture would look at the atmosphere, reflected in a mirror, which again can be varied in position. With the chopper mirror in position and vibrating, one section of the aperture will look at deep space while the other section will look at the atmosphere through the appropriate CO_2 cell, and vice versa. The effect is just as if the input radiation were being chopped by alternate cells of CO_2 . The simplest arrangement is to have one cell empty and one filled with CO_2 , when, provided the CO_2 is at the correct pressure, the chopping will be effective only over the line centres. This, together with the narrowband filter, gives an effective bandwidth of around 1 cm^{-1} . Since the cells of CO_2 are within the aperture of the system all the time, the radiation from them will not be chopped and will not be detected. The radiation detected in this way originates from the very top of the atmosphere. The addition of a little CO_2 to the empty cell absorbs out the narrow line centres, leaving an extremely narrow width on either side of centre and giving a still sharper weighting function which allows regions just below the top of the atmosphere to be examined. Various combinations of filter and chopper have been proposed and a set of weighting functions is shown in figure 13.7. Each satellite installation consists of six separate channels.

To maintain the accuracy of the instrument in flight, it is possible to recalibrate it. The principal components in the calibration system are the variable mirrors which are placed in front of each channel and which normally reflect radiation from the atmosphere into the apertures. These mirrors are driven by small stepping motors and can be tilted to view the atmosphere, deep space, or a calibration black body giving a reference for both gain and zero in each channel. The proposed calibration sequence, which will repeat itself indefinitely in flight, is atmospheric radiation for 20 min, space for 2 min and calibration black body for 2 min. The channels having the extra CO_2 cells also have a balance calibration which ensures that the only difference between each half of the aperture is due to the CO_2 in the chopper cells. The narrowband filter which is used in the channel is replaced by a broadband filter at a wavelength outside the CO_2 absorption region which views the Earth's surface. Any signal detected under these circumstances is due to a difference in sensitivity between the two halves of the channel, which

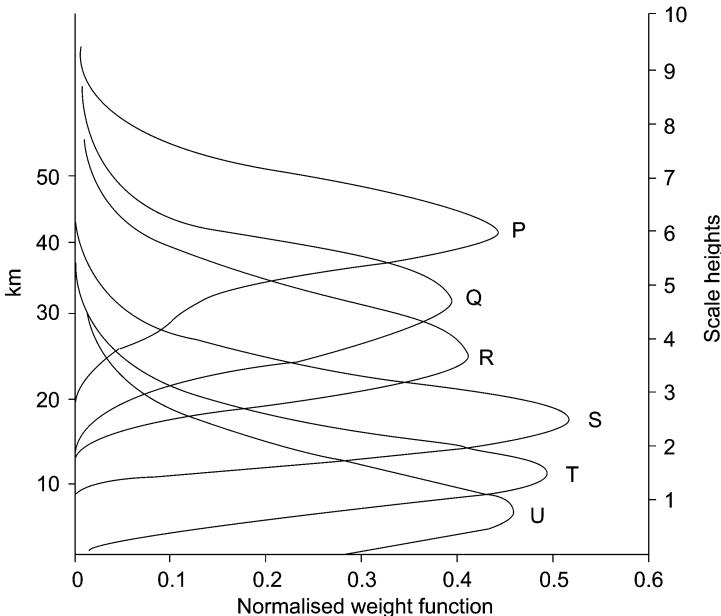


Figure 13.7. Proposed weighting functions for a satellite radiometer. The letters P, Q, R, S, T and U refer to different channels. (Courtesy of Dr S D Smith.)

can be corrected if necessary.

Curves showing the measured transmittance of two of the basic filter elements are reproduced in figure 13.8. The sidebands are suppressed in the instrument by filters of the type shown in figure 6.20. The interference section of the blocking filter is deposited on one of the germanium lenses and an indium antimonide filter is fitted to the end of the light pipe over the detector. In addition, since it was found that the suppression in the wings of the Fabry–Perot filter was not quite high enough, a filter centred on the same wavelength but of the type

$$L|Ge|L\ H\ L\ H\ H\ L\ H\ L\ H\ H\ L\ H|Air$$

which is a rather broader DHW type of around 20 cm^{-1} halfwidth, rather broader than that of figure 13.8(b), is placed in series with each Fabry–Perot. The composite filter possesses the narrow halfwidth of the Fabry–Perot together with the high sideband rejection typical of the DHW.

The construction of the radiometer is shown in figure 13.9. The optical system has been designed to use the full area of the narrowband filters together with the maximum range of angles which can be accommodated without destroying the spectral profile. It was this which prompted the work of Pidgeon and Smith on the angular dependence of filter characteristics discussed on pp 283–92.

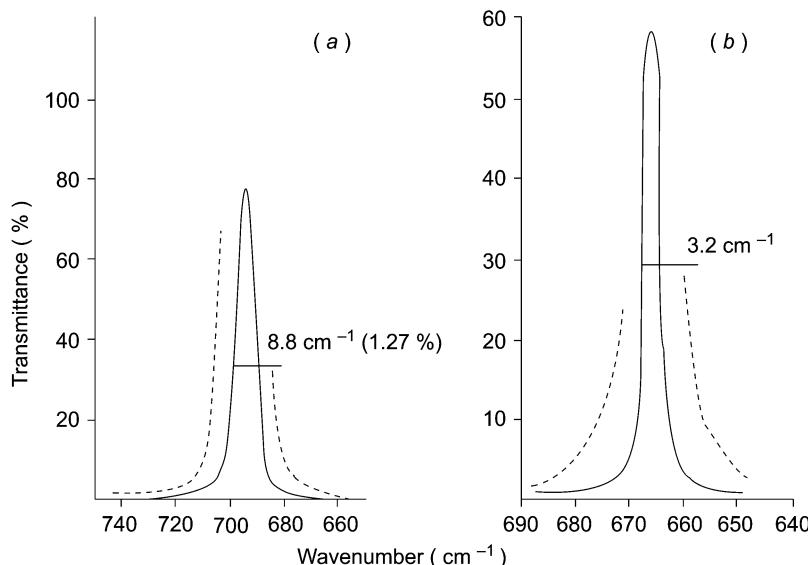


Figure 13.8. Measured transmittance of filters manufactured for the radiometer. The dashed curves are merely the full line curves $\times 10$. (a) Air |HLHHLHL| Ge substrate |L| Air. Peak transmittance 78% at 694.4 cm^{-1} . (b) Air |HLHLHHHLHL| Ge substrate |L| Air. Peak transmittance 58% at 666.4 cm^{-1} . $L = \text{ZnS}; H = \text{PbTe}$. (Courtesy of Dr S D Smith and Sir Howard Grubb, Parsons & Co. Ltd.)

The radiometer was successfully launched in April 1970 and made exceptionally useful temperature surveys of the upper atmosphere revealing much that was novel and unexpected. An early account of the instrument will be found in several articles [15–17].

13.4 Order-sorting filters for grating spectrometers

There is a considerable advantage in using a diffraction grating rather than a prism for the selection of wavelengths in a monochromator or spectrometer because the luminosity is so very much greater for the same resolution. A problem exists, however, with the diffraction grating which does not exist with the prism. This is the appearance of other orders in the spectrum which must be eliminated. The problem is particularly severe in the infrared, and the solution usually adopted has been the use of a low-resolution prism monochromator in series with the higher-resolution grating monochromator. The lower resolution of the prism section, which is all that is necessary since order sorting is its sole function, means that the luminosity can be made as high as the grating section and the advantage associated

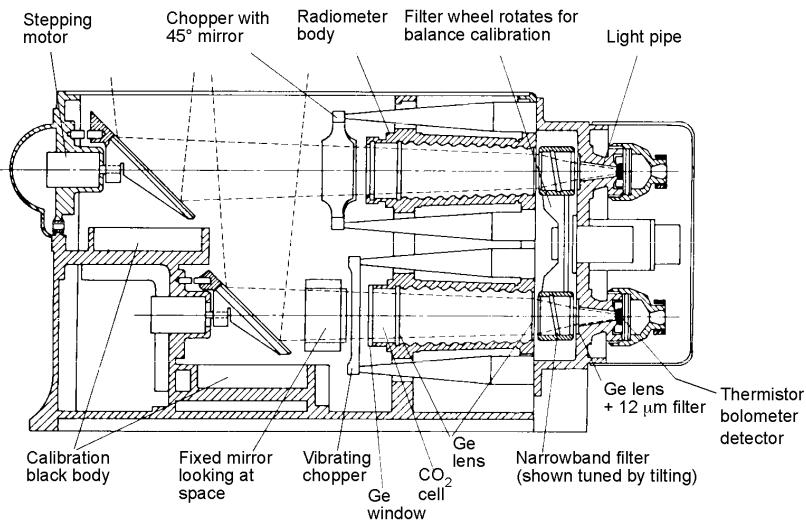


Figure 13.9. Schematic diagram of the selective chopper radiometer. (Courtesy of Dr S D Smith.)

with the grating thereby retained. The grating and prism must be driven so that their respective wavelengths remain in step, a difficulty being that their angular dispersions vary in quite different ways. A simpler and attractive alternative is a longwave pass thin-film filter. Recently several instruments have appeared on the market which use this system rather than the prism.

A paper by Alpert [18] gives an account of the various factors involved in the specification of such filters for infrared instruments. The most important feature is the rejection required in the stop regions. Before we can make an estimate of this rejection, we must first consider the way in which the energy varies in the various grating orders. Included in the assessment must be the characteristics of both the source and the detector.

A simple theory of the diffraction grating is considered in most textbooks on optics. For our present purpose it is sufficient to note two points. The first is that the angles of incidence and diffraction for any particular wavelength are given by the grating equation

$$\sin \vartheta + \sin \varphi = \pm m\lambda/\pi \quad (13.13)$$

where ϑ and φ are the angles of incidence and diffraction, respectively, the sign convention being as shown in figure 13.10(a). σ is the grating constant, that is the spacing of the grooves, and m is the order number. From equation (13.13) we can see immediately the source of our present problem, that the angles corresponding to any wavelength, λ , in the first order also exactly correspond to $\lambda/2$ in the second order, $\lambda/3$ in the third order, and so on. A second point is that the energy distribution in the various diffracted orders of any wavelength

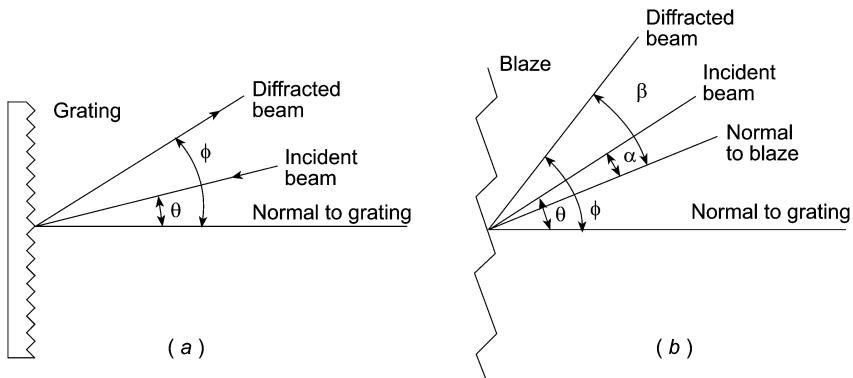


Figure 13.10. (a) Sign convention for θ and ϕ . θ and ϕ have the same sign if they are on the same side of the grating normal. One side is chosen arbitrarily as positive. (b) Sign convention for α and β . α and β have the same sign if they are on the same side of the blaze normal. They are chosen positive when they have the same sense as the positive direction of θ and ϕ .

will be given by the pattern of lines in equation (13.13) modulated by the single-slit diffraction pattern of any one of the grooves at the appropriate wavelength. Modern diffraction gratings are invariably of the reflection type with the grooves ‘blazed’ or tilted, so that the single-slit diffraction pattern has its maximum at a particular wavelength in the first order, known as the blaze wavelength, rather than in the zero order, which increases considerably the efficiency of the grating over a range of wavelengths. In order to estimate the shape of the energy distribution we can assume the form of the grooves to be as in figure 13.10(b), although in practice the form may vary from that shown. α and β are the angles of incidence and diffraction referred to the normal to the groove, instead of the grating normal, but with the same sign convention applying. The intensity of the diffracted beam is given by an expression of the form

$$I = I_0 \frac{\sin^2 [\pi v \sigma \cos \psi (\sin \alpha + \sin \beta)]}{[\pi v \sigma \cos \psi (\sin \alpha + \sin \beta)]^2} \quad (13.14)$$

where it is assumed that the grating will be sufficiently large to intercept the entire incident beam regardless of the angle of incidence. This expression is not strictly accurate over the entire range because at some angles the steps at the ends of the grooves may interfere slightly with the process, but it is good enough for our purpose. ψ is the angle between the grating and the blaze normal.

Most monochromators are of a type where the entrance and exit slits are fixed in position and the grating is rotated to scan the spectrum, and where the angle of incidence is almost equal to the angle of diffraction. Little is lost by assuming that they are equal. With this assumption the curves shown in figure 13.11

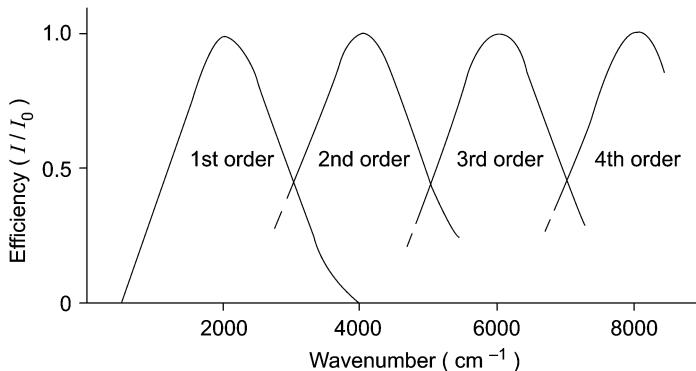


Figure 13.11. Intensity distribution in the various orders of a typical diffraction grating (theoretical) blazed for 2000 cm^{-1} ($5\mu\text{m}$) in the first order.

have been derived for a typical grating and show how the intensities vary in the various orders. An important point is that, for the groove arrangement shown in figure 13.10(b), the dispersion, which is inversely proportional to the groove spacing, balances the alteration in the width of the diffraction pattern as the groove width varies, with the result that the variation of intensity with wavelength in any order depends solely on the blaze wavelength. A useful rule, which is generally used, is that the useful range of a grating which is blazed for a wavelength of λ_0 in the first order is from $2\lambda_0/3$ to $2\lambda_0$ in the first order, from $2\lambda_0/5$ to $2\lambda_0/3$ in the second order, and from $2\lambda_0/(2n+1)$ to $2\lambda_0/(2n-1)$ in the n th order. This is rather simpler in terms of wavenumber, the range being given by $v_0 \pm \frac{1}{2}v_0$ in the first order and $nv_0 \pm \frac{1}{2}v_0$ in the n th order. The bandwidth is more or less constant in terms of wavenumber. Measurements which have been made on gratings confirm the shape of the curves in figure 13.11. Some such measurements are reproduced by Alpert.

Now let us make the assumption that the diffraction grating is to be used in the first order and that the filter problem is the elimination of the second and higher orders. As far as the filter is concerned, the parameter which matters is the ratio of the detector signal in the first order to that in any of the other higher orders. The factors involved are, first of all, the variation of sensitivity of the detector; second, the variation in efficiency of the grating, already dealt with above; third, the dispersion of the grating in the various orders so that the energy in any order which is transmitted by the slits in the monochromator can be calculated; and last, the variation of output of the source. Of course, in some applications there may well be other factors which operate, such as the transmission of some optical components or the variation of reflectance of mirrors.

The detectors commonly used in this part of the infrared are thermal detectors which have reasonably flat response curves. In what follows we assume

that they are perfectly flat. Any variation can be readily included in the analysis if required.

At any wavelength, the slits will pass a small band of wavelengths. If we assume that the slits are narrow enough so that energy variations over the range of wavelengths passed by the slit are negligible, then the energy transmitted in any order will be inversely proportional to the bandwidth of the slits in that order. From equation (13.13), the bandwidth is inversely proportional to the order number, which does help to reduce the requirements for filter performance.

In this part of the infrared, the sources which are generally used are either Nernst filaments or globars. For our present purpose we can assume, without too much error, that the source will be a black body probably peaking at around $2 \mu\text{m}$, although this particular wavelength does not matter very much. The variation of energy with wavelength for a black-body source is given by Planck's equation:

$$E_\lambda = \frac{c_1}{\lambda^5 [\exp(c_2/\lambda T) - 1]} \quad (13.15)$$

where E_λ is the spectral emissive power, and c_1 and c_2 are the first and second radiation constants with values $3.74 \times 10^{-16} \text{ W m}^2$ and $1.4388 \times 10^{-2} \text{ m K}$, respectively.

For any wavelength λ , let the efficiency of the grating be denoted by ε_λ , and the transmittance of the order sorting filter by T_λ . Then the stray light due to the m th order, expressed as a fraction of the energy in the first order, will be given by

$$r_m = \frac{\varepsilon_{\lambda/m} E_{\lambda/m} T_{\lambda/m}}{m \varepsilon_\lambda E_\lambda T_\lambda} = \frac{\varepsilon_{\lambda/m} (\lambda/m) E_{\lambda/m} T_{\lambda/m}}{\varepsilon_\lambda \lambda E_\lambda T_\lambda}$$

where we have multiplied the numerator and denominator by λ . The permissible magnitude of r_m depends on the number of orders which are involved in producing significant interference. Let this number be N and let the total amount of permissible stray light be given by S , which is expressed as a fraction of the total first-order energy. Then we can require that

$$r_m = S/N$$

and the maximum transmission which can be permitted at wavelength λ/m is given by

$$T_{\lambda/m} = T_\lambda \left(\frac{S}{N} \right) \left(\frac{\lambda E_\lambda}{(\lambda/m) E_{\lambda/m}} \right) \frac{\varepsilon_\lambda}{\varepsilon_{\lambda/m}}. \quad (13.16)$$

Now $\varepsilon_\lambda / \varepsilon_{\lambda/m}$ will be greater than unity except on the blaze wavelength. Without affecting the accuracy too greatly, we can make the assumption that each order m is effective only over the range $2\lambda_0/(2m+1)$ to $2\lambda_0/(2m-1)$ and that $\varepsilon_\lambda / \varepsilon_{\lambda/m}$ is unity over this range. Elsewhere we can assume that the m th order does not produce interference and omit it.

To complete the calculation, we need the value of $\lambda E_\lambda / (\lambda/m) E_{\lambda/m}$. The function λE_λ is plotted in figure 13.12. To make it possible to apply this figure

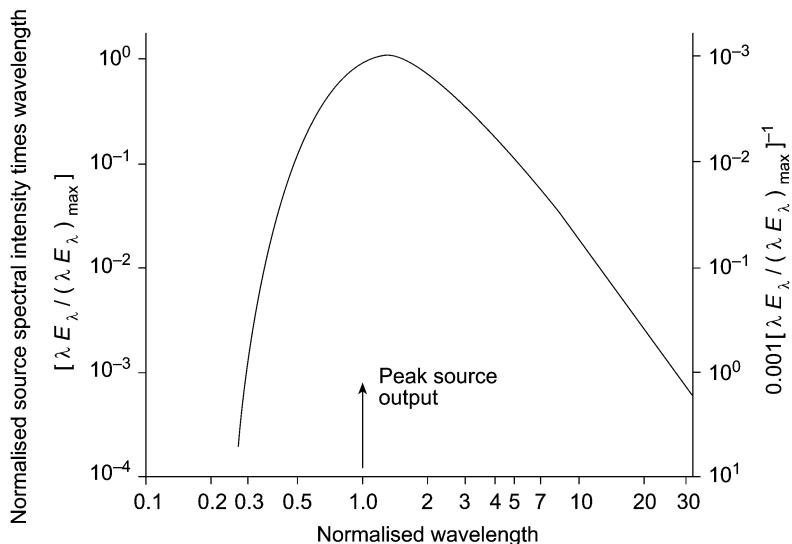


Figure 13.12. Curve showing the variation of λE_λ with wavelength for a black-body source.

generally, the variables have been normalised in the manner shown and the scales are logarithmic so that any particular set of conditions can be reproduced simply by sliding the scales along the axes.

The first step in drawing up the specification for a practical set of filters will be to decide on the required number of filters. Even one single diffraction grating has a useful wavelength range of 3:1, which is greater than the range which can be covered by just one filter.

Let the limits of the wavelength region over which the grating or set of gratings are to be used be λ_F and λ_S , where $\lambda_F > \lambda_S$. If we start with the longest wavelength, then the final filter in the series must block wavelengths $\lambda_F/2$ and shorter. An ideal longwave pass filter would have a rectangular edge shape and it would be possible to use it over the whole of the range $\lambda_F/2$ to λ_F . Real filters have sloping edges and must be allowed some tolerance in edge position otherwise manufacture becomes impossible. This means that the specification must show the start of the transmission region of the final filter as $(1 + \alpha)\lambda_F/2$. Assuming that all the filters in the set are of more or less similar construction, then the same expression will also apply to the next filter in the set, which will have a transmission region specified to start at a wavelength given by $[(1 + \alpha)^2/2^2]\lambda_F$ and to finish at $[(1 + \alpha)/2]\lambda_F$. The regions for the other filters are found in exactly the same way. If there are n filters in the set, then the first filter must have the specified start wavelength at $[(1 + \alpha)^n/2^n]\lambda_F$. We can equate this start wavelength to λ_S

Table 13.1.

Filter number	Pass region (μm)	Longwave edge of rejection zone (μm)
5	19–30	15
4	12–19	9.5
3	7.6–12	6
2	4.8–7.6	3.8
1	3–4.8	2.4

and solve for α :

$$\alpha = 2(\lambda_S/\lambda_F)^{1/n} - 1.$$

This expression can be evaluated in a practical case for several possible values of n and the set of filters giving the optimum arrangement of filters and the best degree of tightness of tolerance selected.

The advantage of using this type of specification is that any particular filter from any set of filters made to the specification is interchangeable with the corresponding filter in any other set. If this interchangeability is not required, it is possible to slacken the tolerances slightly, but this makes the problem of making up each individual set rather more of a puzzle.

To illustrate the method, let us consider the specification for a set of filters for use with a pair of gratings for the region 3–30 μm . The first grating can be the one already considered with blaze at 5 μm , while the second will be a similar one with blaze at 15 μm . The region 3–3.3 μm will not be covered with quite as great efficiency as the rest of the region, but the source will be rather more efficient here, which in fact counterbalances the fall off in grating efficiency to some extent.

First we decide on the number of filters. By inspection we arrive at the conclusion that the minimum number of filters is four, but that this number leads to a specification which is rather tight, and it is better to use five filters. If we assume that the tolerances should be shared equally amongst them, then the limits of the pass regions and the edges of the rejection zones are as shown in table 13.1.

We then decide on the acceptable level of stray light in this case as, say, 1% of the true first-order signal. We must also decide on the acceptable minimum transmittance of the filters in the pass region, say 50%. In practice the level will almost certainly be rather greater than this, but the use of a low figure in setting up the specification gives a pessimistic figure for the specified levels in the rejection region.

Next we compute the regions over which the various orders are effective in producing stray light. The results are shown in table 13.2. Both the actual

Table 13.2.

Order	Range (μm)	Corresponding range in the first order (μm)
15 μm grating		
1st	30–10	30–10
2nd	10–6	20–12
3rd	6–4.29	18–12.85
4th	4.29–3.33	17.15–13.33
5th	3.33–2.72	16.70–13.65
6th	2.72–2.31	16.35–13.85
7th	2.31–2.00	16.15–14.00
8th	2.00–1.76	16.00–14.10
9th	1.76–1.58	15.90–14.20
10th	10th and higher order beyond germanium edge	
5 μm grating		
1st	10–3.33	10–3.33
2nd	3.33–2.00	6.67–4.00
3rd	2.00–1.43	6.00–4.28
4th	4th and higher orders beyond germanium edge	

wavelength of the interfering energy and the corresponding wavelengths in the first order with which it interferes are given. We can choose to use germanium as substrate material for the filters and therefore safely neglect all wavelengths shorter than 1.6 μm , because they will be effectively suppressed by the intrinsic absorption of the germanium.

The first filter we consider is filter number 4, which includes the blaze wavelength of the longer-wave grating in its transmission region. At the blaze wavelength, the highest significant order is the ninth and N therefore is 8, i.e.

$$T_{\lambda_0} S/N = 0.5 \times 0.01/8 = 0.000\,625.$$

We therefore set the scale on the right-hand side of figure 13.12 to correspond to 0.000 625 at 15 μm and read off the allowable transmissions at the higher-order wavelengths from the curve. This is shown in figure 13.13.

To simplify the task of setting up the specification, we assume that the transmission levels which are thus established apply to the complete range for each appropriate order, i.e. for the m th order, the transmission found in this way applies to the range $2\lambda_0/(2m + 1)$ to $2\lambda_0/(2m - 1)$, a slightly pessimistic result. The only exception which we make to this is that portion of the rejection zone immediately beside the edge of the transmission zone. Here it is important that the specification should not be tighter than is strictly necessary. The end of the transmission region is 19 μm . From the range of the higher order interference

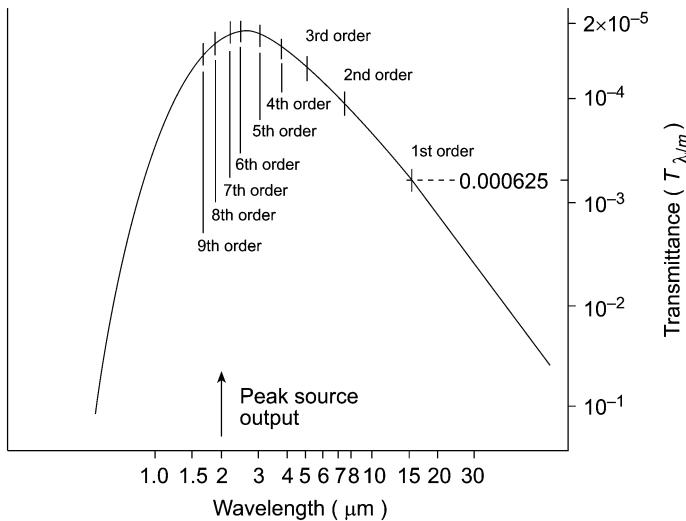


Figure 13.13. How the maximum transmittance levels are established for filter 4.

we see that only one order, the second, is effective at that wavelength. $T_\lambda S/N$ is therefore $0.5 \times 0.01/1.0 = 0.005$. Setting this value on the right-hand scale of figure 13.13 against the point on the curve corresponding to $19 \mu\text{m}$, we read off 0.0009 against $9.5 \mu\text{m}$, which is therefore the maximum allowable transmittance at that wavelength. At $18 \mu\text{m}$, the second and third orders are involved and the value of $T_\lambda S/N$ becomes 0.0025. Setting this against the point on the curve corresponding to $18 \mu\text{m}$, we read off 0.0004 against $9 \mu\text{m}$, which is the maximum allowable transmission at that point. At $17.15 \mu\text{m}$ there are three orders involved so that the transmission at $8.6 \mu\text{m}$ should be not greater than 0.0003. This procedure is repeated at each wavelength where a further order becomes significant until the full number of orders is reached. Points corresponding to these are plotted on a diagram and a horizontal line through each is linked with a vertical line through the adjacent point on the shortwave side. The specification for the filter is then completed by adding a minimum transmittance level of 0.50 from $12\text{--}19 \mu\text{m}$. Figure 13.14 shows the complete arrangement.

Next we consider the longest-wave filter, number 5. Here the conditions are not nearly so severe, because the filter is being used for a region that does not include the first-order blaze wavelength and there is therefore only slight higher-order interference. According to table 13.2 the second-order interference is falling off sharply beyond $20 \mu\text{m}$ and the third order is not effective anywhere within the pass region. The critical region is therefore $9.5\text{--}10 \mu\text{m}$.

$T_\lambda S/N$ is once again 0.005 and setting this value against the point corresponding to $20 \mu\text{m}$ in figure 13.12, we find the permissible transmission in the rejection region at $10 \mu\text{m}$ as 0.0009. Outside the $9.5\text{--}10 \mu\text{m}$ range the simple

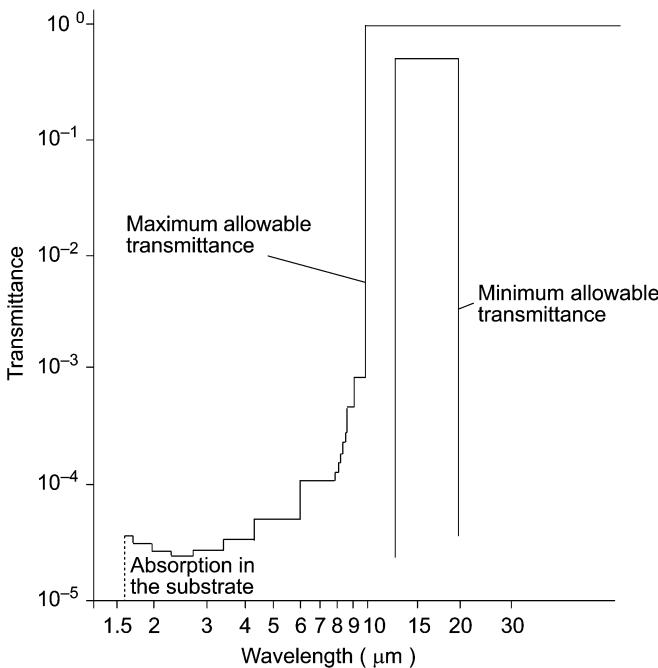


Figure 13.14. Specification of filter 4.

theory which predicts no interference at all is once again not sufficiently accurate. A convenient pessimistic assumption is that the transmission at the very edge of the rejection zone, i.e. at $15 \mu\text{m}$, should be around 0.01 and then a straight line can be drawn from this point to that at $10 \mu\text{m}$. On the shortwave side of $9.5 \mu\text{m}$ we can retain the transmittance as 0.0009. The resulting transmission specification for the filter is given in figure 13.15.

Filter number 3 covers the changeover from one grating to the next. Beyond $10 \mu\text{m}$, the grating is blazed at $15 \mu\text{m}$. The significant range for second-order interference is $12\text{--}20 \mu\text{m}$ so that, except just at $12 \mu\text{m}$, second-order interference will be low. At $12 \mu\text{m}$, $T_\lambda S/N$ is 0.005, and from figure 13.13 the permissible transmission at $6 \mu\text{m}$ is just over 0.001. We can specify this level of transmission as far as $5 \mu\text{m}$, which corresponds to $10 \mu\text{m}$ in the first order, the grating changeover wavelength. On the short wavelength side of $10 \mu\text{m}$ the $5 \mu\text{m}$ grating is used. Table 13.2 predicts that there will be no interference from the edge of the pass band at $7.6 \mu\text{m}$ right to $10 \mu\text{m}$. However to be safe we assume that there will be second-order interference at $7.6 \mu\text{m}$, and setting a value of 0.005 against $7.6 \mu\text{m}$ in figure 13.13, we establish a value for the transmittance at $3.8 \mu\text{m}$, the second-order wavelength. This is shown in figure 13.15 and we further assume that it applies to the whole region between the germanium edge and $5 \mu\text{m}$.

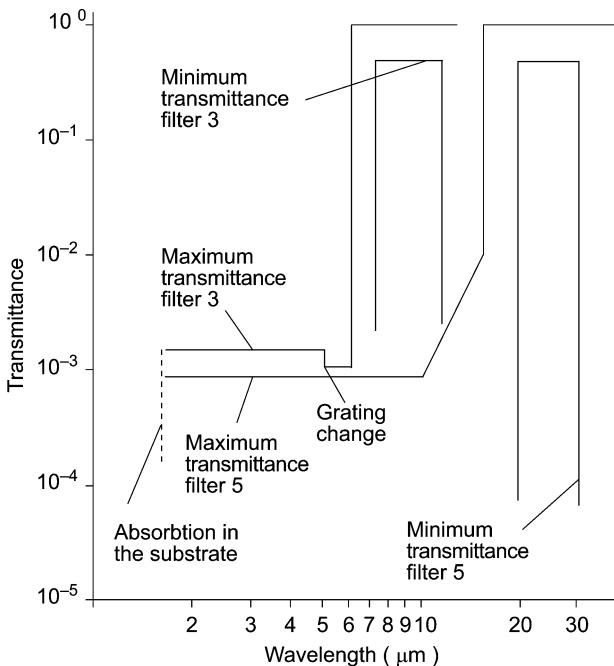


Figure 13.15. Specification of filters 3 and 5.

The specification for filter number 2 (figure 13.16) is set up in exactly the same way as for filter number 4 since it includes the blaze wavelength. However, the requirements are not nearly so severe, because both the peak of the source and the absorption edge of the germanium substrates are much closer to the pass band of the filter.

Filter 1 is similar to the others (figure 13.17). The short band from $1.6 - 2 \mu\text{m}$, where the simple theory predicts no higher order interference (second order missing and third order corresponding to first-order wavelengths beyond $4.8 \mu\text{m}$, the edge of the pass band), is filled in by a horizontal line at the same level as the allowable transmission at $2 \mu\text{m}$.

As far as the optical performance of the filters is concerned, there is only one further point to be specified, the bandwidth of the measuring spectrometer used for inspecting the filters. The requirement here is that the bandwidth should be not greater, nor appreciably less, than the bandwidth of the final instrument in which the filters are to be used⁴. Any spikes of transmittance not resolved by this arrangement will not be resolved by the instrument itself. There is clearly no point in carrying out too strict a test, which would not only be an unnecessary

⁴ i.e. the fractional bandwidth of the measuring instrument should be equal to the fractional bandwidth of the final instrument in the transmission region of the particular filter under test.

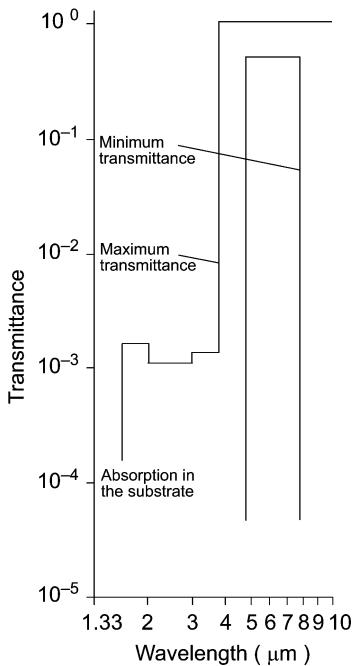


Figure 13.16. Specification of filter 2.

waste of time and expense, but could also lead to a filter being rejected when in fact it is perfectly satisfactory for the job.

Once the specification has been established, the design of the filters is just a straightforward application of the principles discussed in chapter 6. A study of the results suggests some general rules. The first is that the filters which include the first-order blaze wavelength in their pass regions are the most critical in their specifications, and to ease, as far as possible, their edge steepness the blaze wavelength should be arranged to be nearer the shortwave limit of the pass region than the longwave limit. The second point is that since the filters which do not include the first-order blaze have very much reduced rejection requirements, it is useful to make sure that the longest-wave filter, which will be the most difficult to fabricate, has a pass region clear of the blaze wavelength even if in some applications it means an extra filter.

13.5 Glare suppression filters and coatings

Glare is a term that is extensive in its coverage. What we mean by the term in this context is specular reflection of illumination from a bright source that enters the eyes and masks a, usually weaker, desired visual image.

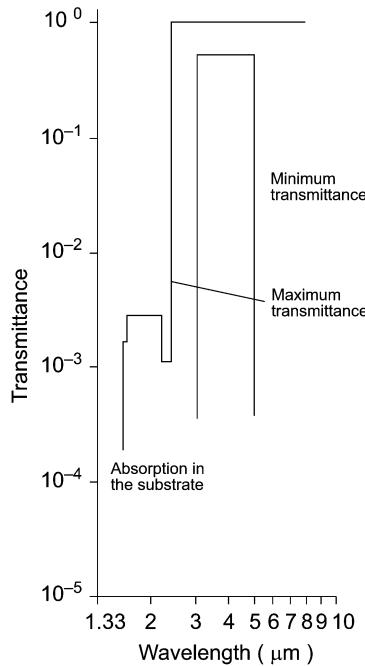


Figure 13.17. Specification of filter 1.

Polarising sunglasses represent an early example of glare reduction. Sunlight reflected by water or silica sand is a common source of glare. When the sun is at an angle that makes the glare a problem the reflection is usually in the vertical plane and at or near the Brewster angle so that the reflected light is principally s-polarised. A person who is upright will receive this glare light as primarily linearly horizontally polarised and it can therefore be virtually eliminated by a suitably oriented polariser.

This solution depends on reflection in the vicinity of the Brewster angle and is not available for the now common glare caused by unsuitable lighting where visual display units are concerned. In this case the signal light from an emitting phosphor at the rear surface of a glass plate is masked by specularly reflected ambient light from the two surfaces of the plate. The orientation of the plane of incidence can vary enormously and the glare can be reflected at angles that are near normal. A solution that has been much used in electronic instruments consists of a circular polariser inserted before the display. Specular reflection at near normal incidence reverses the handedness of the circularly polarised glare light that has already passed through the polariser on its inward journey, and makes it impossible for it to pass through it again on the outward journey. This works well when the specular reflectance of the outer surface of the polariser is

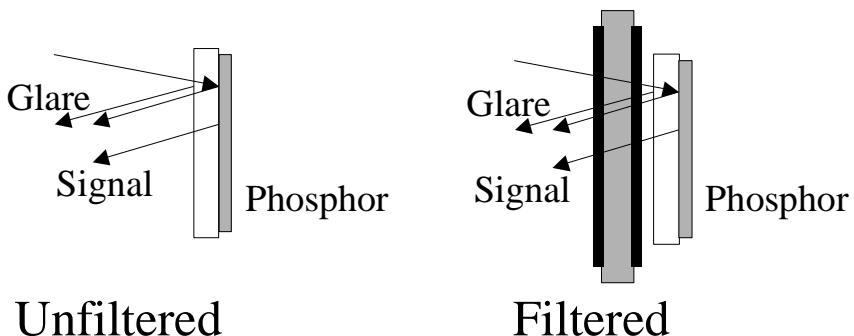


Figure 13.18. The principle of an external antiglare filter. Glare light passes through the filter twice while signal light passes through only once.

appreciably less than that of the underlying display. In other cases the reflectance must be reduced by application of an antireflection coating. Since the circular polariser protects against glare from its own rear surface the antireflection coating is required over the front surface only.

Later it was found that a quite satisfactory reduction in glare could be achieved by replacing the circular polariser by a simple neutral density filter such as a sheet of absorbing glass or plastic. Specular reflectance from the filter is eliminated by antireflection coatings front and back. Glare light then passes through the filter twice while signal light passes through only once. This nominally reduces the glare-to-signal ratio by a factor equal to the transmittance of the filter. However, the brightness of the display can be raised to compensate and so a typical glare reduction is equal to the square of the transmittance. A transmittance of 50%, then, reduces the glare by a factor of four, a quite acceptable figure.

The glare reduction filter of this type is a separate component that is fitted at a late stage to the display unit as an accessory. A very recent tendency is to make the glare reduction component an integral part of the display unit. In its simplest form this is a coating that prevents absorption and acts also as an antireflection coating. The simplest way of achieving this is to replace the normal completely transparent high-index materials by high-index absorbing materials. The most common arrangement takes the four-layer high-efficiency antireflection coating and replaces the usual zirconia or titania with indium tin oxide (ITO). A good antireflection coating that is completely transparent reduces the glare by 50%. Normally it is arranged to have a certain amount of absorption that acts to reduce the glare still further. Figure 13.20 shows a calculated characteristic that uses ITO data from Gibbons *et al* [19]. The overall transmittance of the coating is around 90% and so the glare is further reduced by a factor of 0.8. The glass in the display faceplate is frequently absorbing also and this contributes also to a reduction. The ITO in the coating is a conducting material and acts to reduce electromagnetic

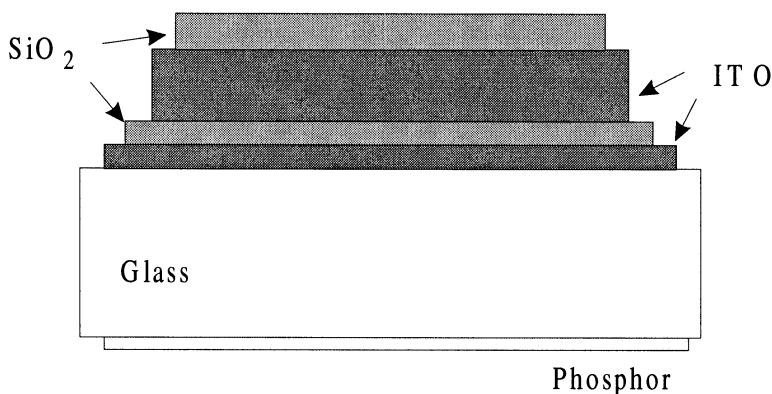


Figure 13.19. Glare-reduction filter applied to the face of a display. The high-index material is made both absorbing and conducting.

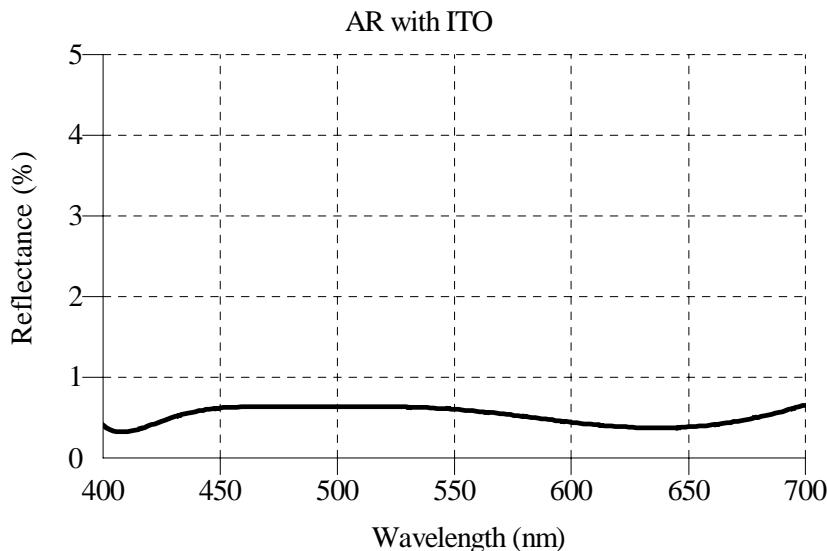


Figure 13.20. Response of a four-layer antireflection coating using silica and ITO. The ITO constants are taken from Gibbons *et al* [19].

emission and static electric fields, but not low-frequency magnetic fields.

To enhance the absorption still further and increase the glare reduction materials that are still more absorbing may be used. Transition metal nitrides, such as titanium nitride, are one possibility [20]. Wolfe [21] has used layers of silver and nickel to increase the absorption and at the same time assure the electrical conductance. Silver was incorporated in the form of a subsystem

consisting of around 8 nm of silver surrounded by 1.2–2.0 nm of NiCrN_x that was in turn surrounded by some 20–30 nm of SiN_x or SiZrN_x . An outer layer of SiO_2 then completed the coating. Alternatively a layer of nickel, perhaps 6–9 nm thick surrounded by protecting layers of SiN_x to protect it from oxidation was found satisfactory. Coatings involving these materials could be made to have transmittances in the range of 30% to 80%.

An ingenious family of two-layer coatings for glare reduction has recently been proposed. Early development was carried out by a group at the Asahi Glass Company Ltd in Japan [22]. A further description is given by Ishikawa and Lippey [23]. Absorbing two-layer coatings are also discussed in detail by Zheng and colleagues [24].

At the shortest wavelength the coating can be considered to consist essentially of a typical V-coat with a thin high-index layer next to the substrate and a rather thicker low-index layer outermost. For simplicity the substrate in this description is transparent but this is not a necessary condition. Now let the wavelength move to a longer value. The physical thicknesses of the layers will remain constant but in the absence of dispersion both optical thicknesses will become smaller fractions of the wavelength and so the admittance loci will shrink. Now imagine that as the wavelength changes the reflectance of the coating remains at zero. The outermost low-index layer can be considered to be a normal dielectric material, like silica, and so it will exhibit negligible dispersion. The end point is firmly fixed at unity on the real axis, the admittance of the incident medium, and so, since the locus is shorter, the starting point moves around the existing circle. Similarly, if the index of the high-index inner layer remains constant and the starting point is firmly fixed at the admittance of the substrate on the real axis, the end point will move around the high admittance circle and a gap will open up in the locus so that it is no longer valid. Now let the optical constants of the inner layer, the high-index layer, be completely adjustable. By adjusting both the index of refraction and the extinction coefficient, the end point of the locus can be swept over a quite large area of the admittance diagram. The gap in the admittance locus can be closed so that it becomes valid and the reflectance remains at zero. By arranging for appropriate smooth variations in both n and k the reflectance can be retained at zero over the entire visible region.

The properties of tungsten-doped titanium nitride are very close to ideal. Measured values taken from Ishikawa and Lippey (estimating from their graph) are given in table 13.3. The thicknesses of the tungsten-doped titanium nitride film and the silica film were 10 nm and 80 nm respectively.

We use a cubic spline interpolation to smooth the constants given in the table and then calculate the performance assuming a normal dispersive index for the glass substrate and arrive at the performance given in figure 13.21. This is impressive.

The calculated transmittance of the coating is shown in figure 13.22. It is surprisingly neutral and will contribute to a satisfactory reduction in glare. Although no figures are given, the authors mention that the coating also reduces

Table 13.3.

Wavelength	Refractive index	Extinction coefficient
405.00	2.5	0.7
510.00	1.8	1.3
632.80	1.2	1.7

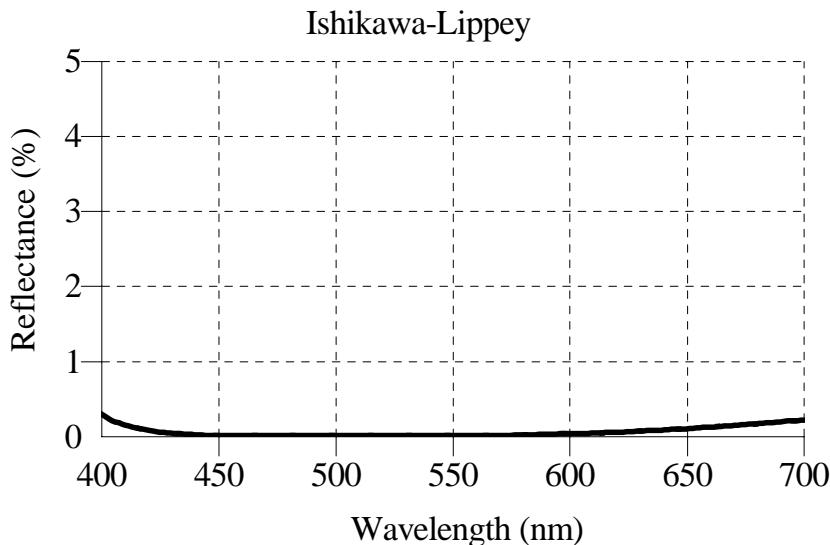


Figure 13.21. The performance calculated for design: Air | SiO₂: 80 nm | TiN_xW_y: 10 nm | Glass. (Calculation parameters from Ishikawa and Lippey [23].)

emissions from the display unit.

The admittance diagram in figure 13.23 shows clearly the way in which the dispersion of the optical constants of the absorbing layer holds the termination of the locus in the vicinity of the incident medium admittance and keeps the reflectance low.

13.6 Some coatings involving metal layers

13.6.1 Electrode films for Schottky-barrier photodiodes

A simple diode photodetector consists of a metal layer deposited over a semiconductor forming a Schottky barrier. High quantum efficiency can be achieved. The incident light passes through the metal layer into the depletion

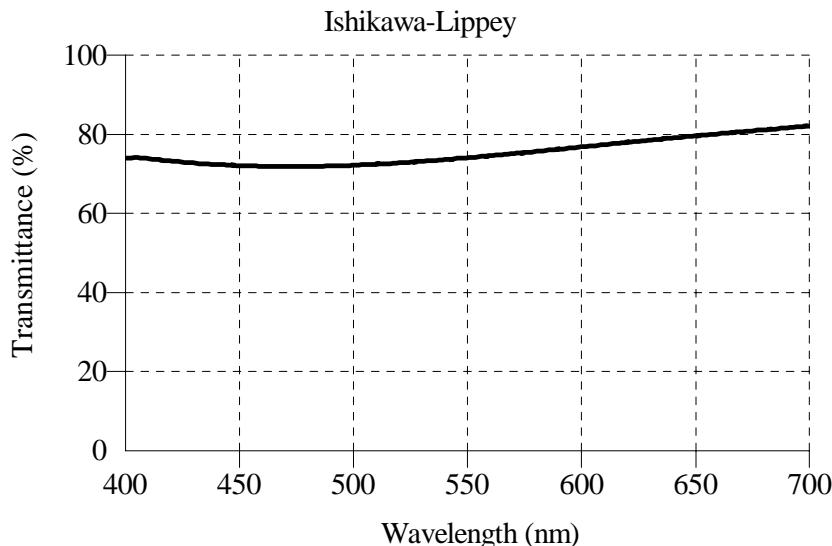


Figure 13.22. The calculated transmittance of the coating of figure 13.21.

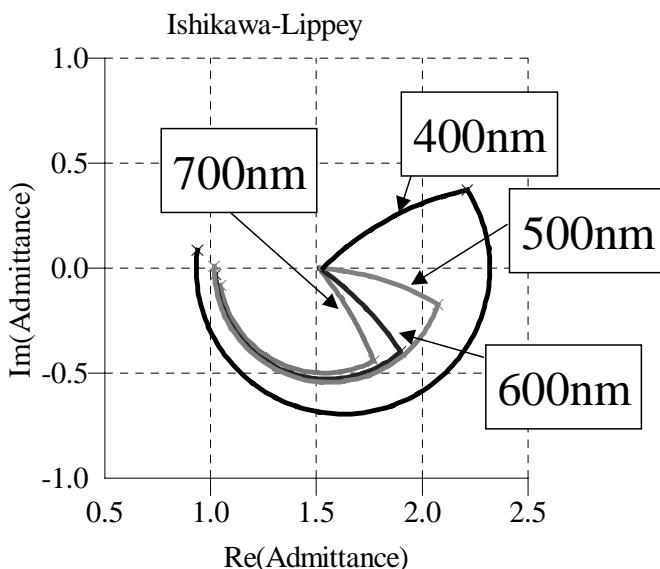


Figure 13.23. The admittance locus of the antireflection coating of figure 13.21 showing how the dispersion of the optical constants of the layer next to the substrate compensates for the shortening of the locus as the wavelength increases.

layer of the diode where it creates electron–hole pairs. The metal contact layer must transmit the incident light and since it has intrinsically high reflectance, it must be coated to reduce its reflection loss. We give here a very simple approach to the design of a combination of electrode and antireflection coating. A number of workers [25–27] have made contributions in this area with probably the most complete account of an analytical approach being that of Schneider [27].

The substrate for the thin films is the semiconducting part of the diode and it is fixed in its optical admittance. The metal layer goes directly over the semiconductor (in some arrangements there is a very thin insulating layer that has negligible optical interference effect) and so the potential transmittance is fixed entirely by the thickness of the metal. All that can be done to maximise actual transmittance is simply to reduce the reflectance to zero.

We take as an example a gold electrode layer deposited on silicon. We assume a wavelength of 700 nm and optical constants of $0.131 - i3.842$ for gold and $3.92 - i0.05$ for silicon [28]. The optical constants of silver and copper are quite similar to those of gold at this wavelength and the results apply almost equally well to these two alternative metals. The admittance locus of a single gold film on silicon is shown in figure 13.24. An antireflection coating must bridge the gap between the appropriate point on the metal locus to the point $(1, 0)$ corresponding to the admittance of air. We can assume that the maximum and minimum values of dielectric layer admittance available for antireflection coating are 2.35 and 1.35, respectively. Using these values, we can add to the admittance diagram two circles that pass through the point $(1, 0)$ and correspond to admittance loci of dielectric materials of characteristic admittances 2.35 and 1.35, respectively. These loci define the limits of a region in the complex plane. Provided a metal locus ends within this region, then it will be possible to find a dielectric overcoat of admittance between 1.35 and 2.35 that, when the thickness is correctly chosen, will reduce the reflectance to zero. It is clear from the diagram that the thicker the metal film, the higher must be the admittance of the antireflection coating. Once the metal locus extends beyond this region, a single dielectric layer can no longer be used and a multilayer coating (or a single absorbing layer, although it would reduce transmittance and so would not be very useful in this particular application) becomes necessary. We have already considered multilayer coatings in the section on induced transmission filters. Here we limit ourselves to a single layer and take the highest available index of 2.35.

The remaining task in the design is then to find the thicknesses of metal and dielectric corresponding to the trajectories between the substrate and the point of intersection, and between the point of intersection and the point $(1, 0)$ in figure 13.24. The points marked along the metal locus correspond to intervals of $0.005\lambda_0$ in geometrical thickness, that is to thickness intervals of 3.5 nm. Visual estimation suggests a value of $0.013\lambda_0$ for the thickness to the point of intersection. A more accurate calculation gives $0.0133\lambda_0$, that is a thickness of 9.3 nm. The dielectric layer has an optical thickness of somewhere between an eighth- and a quarter-wave, and accurate calculation yields $0.186\lambda_0$.

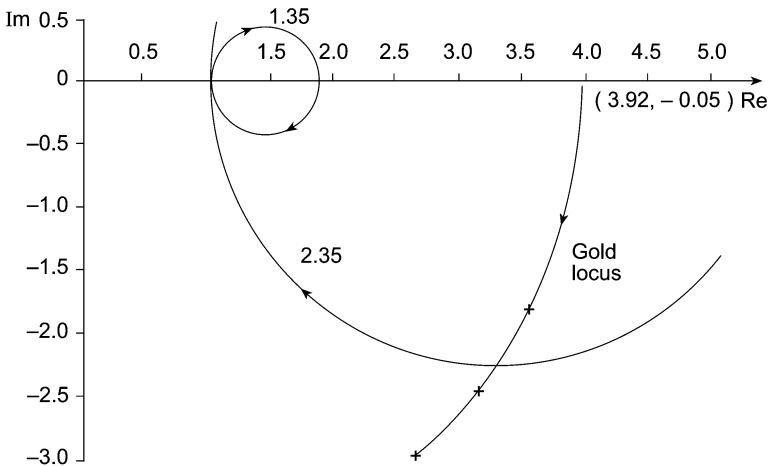


Figure 13.24. Admittance diagram showing some of the factors in the antireflection of a metal electrode layer in a photodiode. The optical constants of gold are assumed to be $(0.131 - i3.842)$ at a wavelength λ_0 of 700 nm. The gold is deposited on silicon with optical constants $(3.92 - i0.05)$. The crosses on the gold locus mark thickness increments of $0.005\lambda_0$ i.e. 3.5 nm. Also shown are loci corresponding to dielectric layers of indices 1.35 and 2.35 that terminate at the point 1.00.

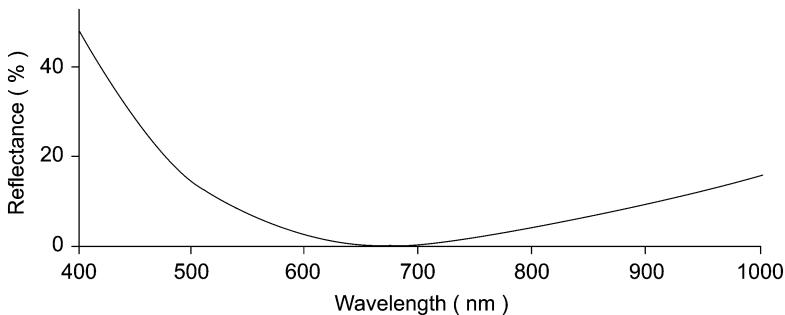


Figure 13.25. The calculated transmittance, including dispersion, of the gold electrode film and antireflection coating designed in the text.

The calculated performance of this coating is shown in figure 13.25. Of course, the thickness of the metal film is rather small and it is unlikely that the values of optical constants measured on thicker films would apply without correction, but the form of the curve and the basic principles of the coating are as discussed here.

13.6.2 Spectrally selective coatings for photothermal solar energy conversion

Coatings for application in the field of solar energy represent a complete subject in their own right. They have been discussed in detail by Hahn and Seraphin [29]. Here we consider simply a limited range of coatings based on antireflection coatings over metal layers that have much in common with the electrode film of the previous section.

Solar absorbers that operate at elevated temperatures can lose heat by radiation unless steps are taken to reduce their emittance in the infrared. Yet to operate efficiently they must have high solar absorptance in the visible and near infrared. Optimum results are obtained from an absorbing coating that exhibits a sharp transition from absorbing to reflecting at a wavelength in the near infrared that varies with the operating temperature of the absorber. One way of constructing such a coating is to start with a thick metal film or a metal substrate and apply an antireflection coating that is efficient over the visible but which becomes ineffective in the infrared, so that at longer wavelengths the reflectance is high and the thermal emittance, as a result, low. Fortunately, we are interested simply in a reduction of reflectance. Transmittance is unimportant. The energy that is not absorbed in the coating is absorbed in the substrate. Thus the antireflection coating can include absorbing layers.

A useful approach to the design is the use of a semiconducting layer over a metal. The semiconductor becomes transparent in the infrared beyond the intrinsic edge and so in that region the reflectance of the underlying metal predominates. In the visible and near infrared the absorption in the semiconductor is sufficient to suppress the metallic reflectance, and to complete the design it is sufficient to add an antireflection coating to reduce the reflectance of the front face of the semiconductor. Since the metal is to dominate the infrared performance either the semiconductor layer must be relatively thin in the infrared or the metal must have sufficiently high k/n to be only slightly affected by the high index of the semiconductor in its transparent region. From the point of view of optical constants, silver is therefore the most favourable metal but it suffers from a lack of stability at elevated temperatures that cause it to agglomerate so that its optical constants are shifted and its reflectance reduced. Seraphin and his colleagues (see their article [29] for a readily available summary and more detailed references) have developed coatings in which the silver is stabilised by layers of chromium oxide (Cr_2O_3) which act as diffusion inhibitors. The silicon films are produced by chemical vapour deposition in which the silicon–hydrogen bonds in silane gas flowing over the substrate are broken by elevated substrate temperature and, as a result, silicon deposits. Adding oxygen or nitrogen to the gas stream gives an antireflection coating of silicon oxide or nitride that can be graded in composition by continuous variation of gas-stream composition. Such coatings can withstand temperatures in excess of 600 °C without degradation.

The design of such coatings is straightforward. First of all, the thickness

of silicon must be such that the visible absorption is sufficiently high to mask the underlying silver but not so thick that interference effects reduce reflectance and increase emittance in the infrared. In the visible region, the light that enters the silicon layer and is reflected from the silver at the rear surface should be sufficiently attenuated that only a very small proportion ever re-emerges. We can assume that the attenuation of this light depends on a law of the form $\exp(-4\pi kd/\lambda)$ and for the entire round trip from front surface to rear of film and back again to the front surface we should have a value roughly in the range 0.01–0.05. Let us choose a design wavelength of 500 nm in the first instance at which silicon in thin-film form has optical constants of 4.3 – i0.74 [28]. Then for $\exp(-4\pi kd/\lambda)$ to be 0.05, the value of d must be 160 nm. Since this is for the entire round trip, the film thickness should be half this value, or 80 nm. An antireflection coating must then be added to reduce the visible reflectance of the front surface of the silicon layer. Since we have reduced the interference effects to a low level, the front surface will be similar to bulk silicon with optical constants characteristic of the material. Seraphin and his colleagues used a graded-index film of silicon nitride and silicon dioxide, but for simplicity we assume here a homogeneous film of roughly 2.0 admittance and a quarter-wave thick at 500 nm. We can take zirconium dioxide with its characteristic admittance of 2.07 as an example. The performance of the complete coating is shown in figure 13.26. The extra dip at 600 nm is a result of the thickness of the silicon. The silicon admittance locus spirals around, converging on the optical constants. At 600 nm, the spiral is somewhat shorter but the end point is passing through a region where the zirconium oxide layer can act as a reasonably efficient antireflection coating once again and so the dip appears. The silver begins to assert itself at around 700 nm in this design. We can shift the reflectance trough to a longer wavelength, say 750 nm, by carrying out a completely similar procedure but this time using 4.17 – i0.37 for the optical constants. Now a double-pass reduction of 0.05 leads to a round-trip thickness of 480 nm, representing a film thickness of 240 nm. The performance is also shown in figure 13.26. In both traces the optical constants of silicon and silver were derived from [28].

An alternative arrangement makes use of metal layers as part of an antireflection coating for silver. The great problem in designing an antireflection coating for a high-efficiency metal using entirely dielectric layers is that the admittance where the locus of the first dielectric layer, that is the layer next to the metal, first cuts the real axis is far from the point (1, 0) where we want to terminate the coating, and with each pair of subsequent quarter-waves we can modify that admittance by only $(n_H/n_L)^2$. Many quarter-waves are needed, as we have seen with the induced transmission filters. A metal layer, on the other hand, follows a different trajectory from a dielectric layer, cutting across dielectric loci, and can be used to bridge the gap between the large radius circle of the dielectric next to the metal and a dielectric locus that terminates at (1, 0).

The metal locus itself can be arranged to pass through (1, 0) but the extra dielectric layer is capable of giving a slightly broader characteristic and also some

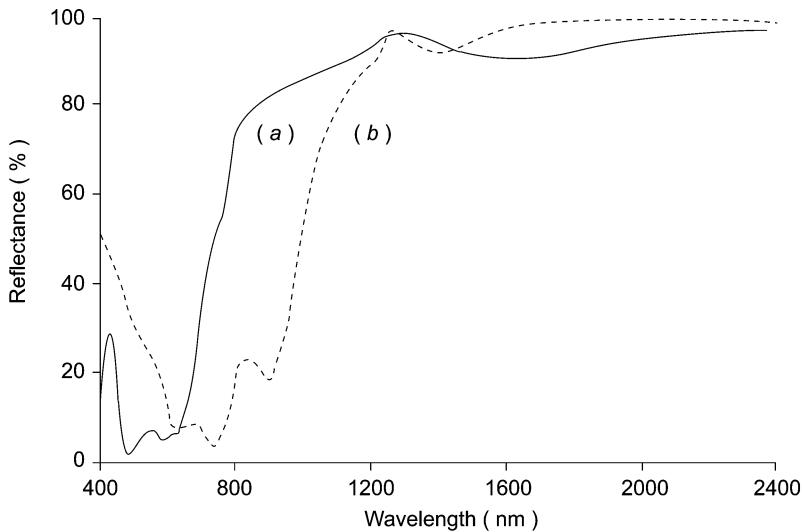


Figure 13.26. The calculated performance including dispersion of solar absorber coatings consisting of antireflected silicon over silver. Designs

(a)	$\left \begin{array}{c} \text{ZrO}_2 \\ 0.25\lambda_0 \end{array} \right $	$\left \begin{array}{c} \text{Si} \\ 80 \text{ nm} \end{array} \right $	$\left \begin{array}{c} \text{Ag} \\ \lambda_0 = 500 \text{ nm} \end{array} \right.$
(b)	$\left \begin{array}{c} \text{ZrO}_2 \\ 0.25\lambda_0 \end{array} \right $	$\left \begin{array}{c} \text{Si} \\ 240 \text{ nm} \end{array} \right $	$\left \begin{array}{c} \text{Ag} \\ \lambda_0 = 750 \text{ nm} \end{array} \right.$

Further details are given in the text.

protection to the metal layer. Silver could be used as the matching metal but its high k/n ratio leads to rather narrow spike-like characteristics even with the terminating dielectric layer, and a metal with rather greater losses is better. We use chromium here as an illustration with aluminium oxide as dielectric. These materials have figured in published coatings (see Hahn and Seraphin [29] for further details). We choose a wavelength of 500 nm for the design and the optical constants we assume for our materials are silver: 0.05 – i2.87; aluminium oxide: 1.67; and chromium: 2.86 – i4.11. Again the optical constants of the metals were obtained from Hass and Hadley [28] with interpolation if necessary. An admittance diagram of a coating of design:

Air	$\left \begin{array}{c} \text{Al}_2\text{O}_3 \\ 0.184\lambda_0 \end{array} \right $	$\left \begin{array}{c} \text{Cr} \\ 7.5 \text{ nm} \end{array} \right $	$\left \begin{array}{c} \text{Al}_2\text{O}_3 \\ 0.184\lambda_0 \end{array} \right $	Ag ($\lambda_0 = 500 \text{ nm}$)
-----	---	---	---	-------------------------------------

is shown in figure 13.27. The chromium locus bridges the gap between the two dielectric layers. Because of its rather lower k/n ratio than silver its trajectory is flatter and the entire characteristic less sensitive to wavelength changes. The

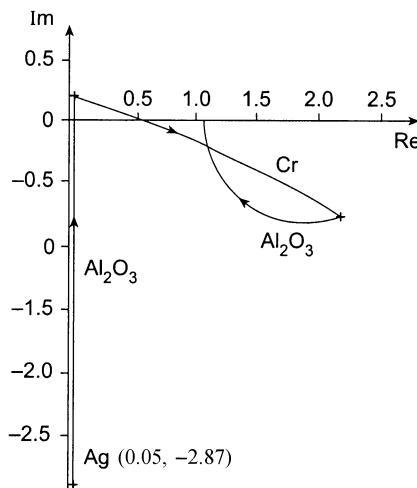


Figure 13.27. Admittance diagram at λ_0 of an absorber coating of design:

Air	$\left \begin{array}{c} \text{Al}_2\text{O}_3 \\ 0.184\lambda_0 \end{array} \right $	$\left \begin{array}{c} \text{Cr} \\ 7.5 \text{ nm} \end{array} \right $	$\left \begin{array}{c} \text{Al}_2\text{O}_3 \\ 0.184\lambda_0 \end{array} \right $	Ag
-----	---	---	---	----

$\lambda_0 = 500 \text{ nm}$. See the text for an explanation.

arrangement helps to keep the final end point of the coating in the vicinity of (1, 0) as the loci increase or decrease in length with changing wavelength or g .

No attempt was made to refine this design although clearly, because of the wide range of possible thickness combinations that would lead to zero reflectance at the design wavelength, there must be scope for performance improvement by refinement. The characteristic of the coating is shown in figure 13.28. The reflectance minimum can be shifted to longer wavelengths by repeating the design process with appropriate values of the optical constants. This gives the desired zero but then at shorter wavelengths, where the dielectric loci are departing further and further from ideal and the chromium layer is unable to bridge the gap between them, a peak of high reflectance is obtained. At still shorter wavelengths, there is a second-order minimum where the dielectric layers make a complete revolution and are once again in the vicinity of the correct position. For the ideal values we have used in these calculations the central peak of high reflectance is very high indeed. Practical coatings also show this double minimum (see Hahn and Seraphin [29]), but the central maximum is very much less prominent, the most likely explanation being that the layers in practice have much greater losses than we have assumed. In particular, the thin chromium layers are unlikely to have ideal optical constants. High losses would make the loci spiral in towards the centre of the diagram and reduce the wavelength sensitivity.

The major problems associated with such coatings are not their design but

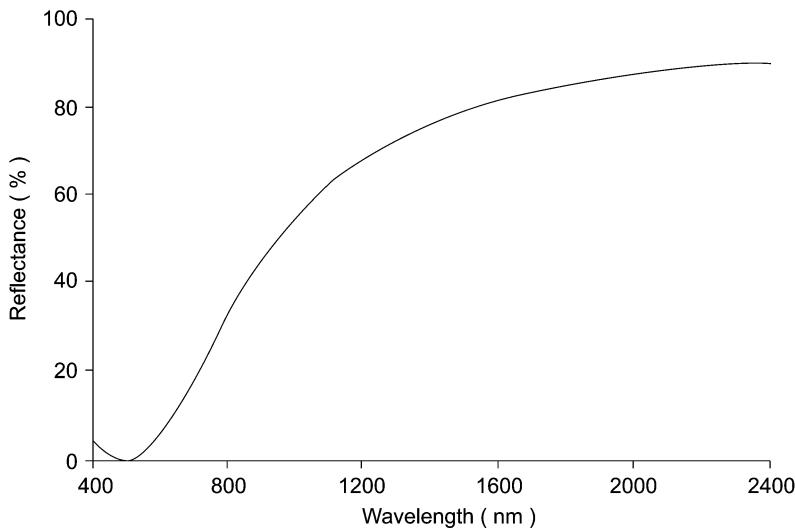


Figure 13.28. The calculated performance, including dispersion, of the absorber coating of figure 13.27.

the necessary high-temperature stability. Spectrally selective solar absorbers are only economically viable when they are used to produce high temperatures and, indeed, it is only at high temperatures that they offer an advantage over the more conventional spectrally flat black absorbing surfaces that can be produced very much more cheaply. They are used under vacuum to eliminate gas conduction heat losses and so the major degradation mechanism is diffusion within the coatings. Silver is particularly prone to agglomerate at high temperatures and much development effort has resulted in the incorporation of thin diffusion barriers such as chromium oxide that inhibit diffusion and agglomeration of the components without affecting the optical properties. The achievements in terms of lifetime at high temperatures are impressive. Further details will be found in Hahn and Seraphin [29].

13.6.3 Heat reflecting metal–dielectric coatings

There are several applications where a cheap and simple heat-reflecting filter would be valuable. For example, a normal, spectrally flat solar absorber can be combined with such a filter so that the combination acts as a spectrally selective absorber. It is possible to construct a very simple band-pass filter that has the desired characteristics from a single metal layer surrounded by two dielectric matching layers [30–33]. The filter is similar in some respects to the induced transmission filter, although the maximum potential transmittance that is theoretically possible cannot usually be achieved. One design technique uses the

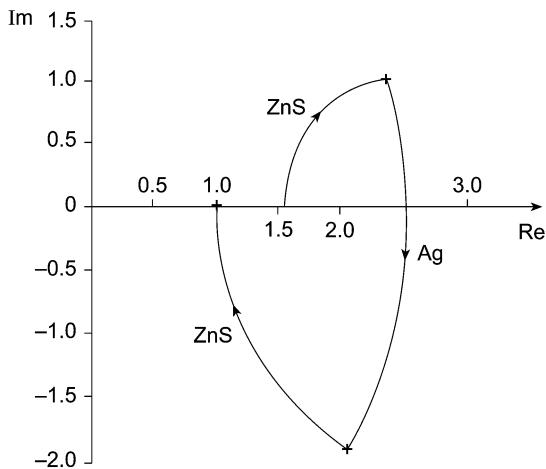


Figure 13.29. Admittance diagram of a metal–dielectric heat-reflecting filter. The diagram shows the locus at a wavelength of 600 nm of a ZnS | Ag | ZnS combination deposited on glass.

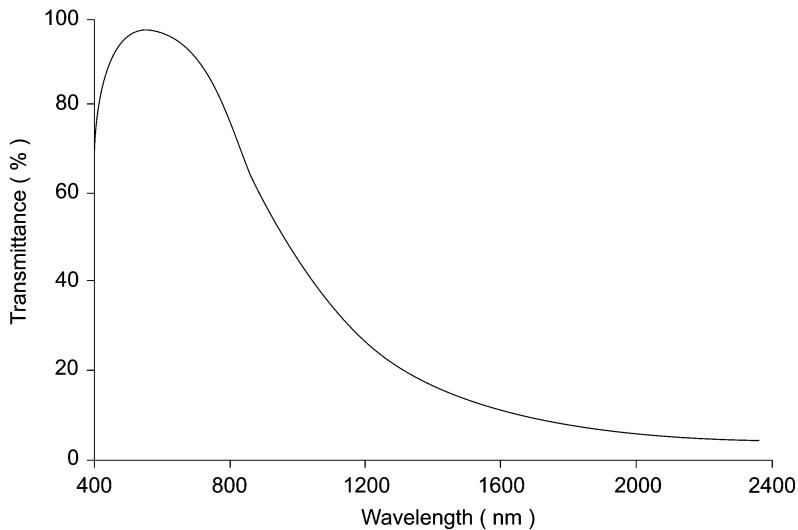


Figure 13.30. Transmittance, calculated with dispersion included, of the heat-reflecting coating of figure 13.29. Details of the design are given in the text.

admittance diagram and we can illustrate it with an example in which we consider a glass substrate and an incident medium of air or vacuum. Silver, with optical constants of $0.06 - i3.75$ at 600 nm, can serve as the metal and we assume a

dielectric layer material of index 2.35. Zinc sulphide, which has such an index, has been used in this application, but the most durable and stable coatings are ones incorporating a refractory oxide. Figure 13.29 shows an admittance diagram in which one dielectric locus begins at the substrate and a second terminates at (1, 0) corresponding to the incident medium. If the complete coating is to have zero reflectance then the remaining layers must bridge the gap between these two loci. Once again, it is easy to see that a metal layer can do this and also that the particular optical constants of the metal are unimportant; they will simply alter somewhat the points of intersection with the two loci. The loci shown correspond approximately to the thickest silver film that will still give zero reflectance. To increase the silver thickness without sacrificing the zero reflectance requires that the indices of the two dielectric layers be increased. A small increase in thickness of metal without a gross alteration in the design could be achieved by the insertion of a low-index quarter-wave layer next to the substrate to move the starting point of the next high-index dielectric layer, the upper one in the admittance diagram, further along the real axis towards the origin. The new locus would be outside the existing one demanding a thicker metal matching layer. In the absence of such a low-index layer, the final three-layer design is:

Air	ZnS	Ag	ZnS	Glass	
1.0	2.35	0.06 – i3.75	2.35	1.52	$\lambda_0 = 600 \text{ nm}$
	$0.146\lambda_0$	15 nm	$0.141\lambda_0$		

with performance shown in figure 13.30. The steep fall towards the infrared is partly due to the drop in efficiency of the matching, but an inspection of the admittance diagram quickly reveals that the reduction in length of each locus accompanying an increase in wavelength should not by itself change the reflectance grossly. The dispersion of the silver, however, keeps the value of $(2\pi kd/\lambda)$ high and, hence, the locus long, and is primarily responsible for the increase in reflectance in the infrared. The coating could be based on virtually any metal with high infrared reflectance and high-index dielectric material. Gold and bismuth oxide have been successfully used [33].

References

- [1] Matteucci J and Baumeister P W 1980 Integration of thin-film coatings into optical systems *Proc. Soc. Photo-Opt. Instrumentation Eng.* **237** 478–85
- [2] Jacquinot P 1954 The luminosity of spectrometers with prisms, gratings or Fabry–Perot etalons *J. Opt. Soc. Am.* **44** 761–5
- [3] Baum W A 1962 The detection and measurement of faint astronomical sources *Stars and Stellar Systems* ed W A Hiltner (Chicago: University of Chicago)
- [4] Bowen I S 1964 Telescopes *Astron. J.* **69** 816–25
- [5] Courtes G 1964 Interferometric studies of emission nebulosities *Astron. J.* **69** 325–33
- [6] Ring J 1956 The Fabry–Perot interferometer in astronomy *Astronomical Optics and Related Subjects* ed Z Kopal (Amsterdam: North Holland) pp 381–8

- [7] Meaburn J 1967 A search for nebulosity in the high galactic latitude radion spurs *Z. Astrophys.* **65** 93–104
- [8] Meaburn J 1976 *The Detection and Spectrometry of Faint Light* (Boston: D Reidel)
- [9] Kaplan L D 1959 Inference of atmospheric structure from remote radiation measurements *J. Opt. Soc. Am.* **49** 1004–7
- [10] Smith S D 1961 Design of interference filters for the observation of infra-red emission from atmospheric carbon dioxide by an earth satellite *Quart. J. R. Meteorological Soc.* **87** 431–4
- [11] Smith S D and Pidgeon C R 1965 Application of multiple beam interferometric methods to the study of CO₂ emission at 15 μm *Mémoires Soc. R. Sci. Liège* **9** 336–49
- [12] Houghton J T 1961 The meteorological significance of remote measurements of infra-red emission from atmospheric carbon dioxide *Quart. J. R. Meteorological Soc.* **87** 102–4
- [13] Houghton J T and Shaw J H 1965 The deduction of stratospheric temperature from satellite observations of emission by the 15 micron CO₂ band *Mémoires Soc. R. Sci. Liège* **9** 350–6
- [14] Houghton J T 1963/4 Stratospheric temperature measurements from satellites *J. Br. Interplanetary Soc.* **19** 381–5
- [15] Ellis P J, Peckham G, Smith S D, Houghton J T, Morgan C G, Rogers C D and Williamson E J 1970 First results from the selective chopper radiometer on Nimbus 4 *Nature* **228** 139–43
- [16] Houghton J T and Smith S D 1970 Remote sounding of atmospheric temperature from satellites. I. Introduction (For part II see Abel *et al* *Proc. R. Soc. A* **320** 35–55) *Proc. R. Soc. A* **320** 23–33
- [17] Abel P G, Ellis P J, Houghton J T, Peckham G, Rodgers C D, Smith S D and Williamson E J 1970 Remote sounding of atmospheric temperature from satellites. II. The selective chopper radiometer for Nimbus D *Proc. R. Soc. A* **320** 23–55
- [18] Alpert N L 1962 Infra-red filter grating spectrophotometers—design and properties *Appl. Opt.* **1** 437–42
- [19] Gibbons K P, Carniglia C K, Laird R E, Newcomb R, Wolfe J D and Westra S W T 1997 ITO coatings for display applications *40th Annual Technical Conference (New Orleans)* (Society of Vacuum Coaters) pp 314–18
- [20] Bjornard E J Viratec Thin Films 1992 *Electrically-Conductive, Light Attenuating Antireflection Coating* USA Patent 5 091 244
- [21] Wolfe J 1995 Anti-static, anti-reflection coatings using various metal layers *38th Annual Technical Conference (Chicago)* (Society of Vacuum Coaters) pp 272–5
- [22] Oyama T and Katayama Y Asahi Glass Company Ltd, Tokyo, Japan 1997 *Light Absorptive Antireflector* USA Patent 5 691 044
- [23] Ishikawa H and Lipsey B 1996 Two layer broad band antireflection coating *10th International Conference on Vacuum Web Coating (Fort Lauderdale, FL)* (Englewood, NJ: Bakish Materials Corporation) pp 221–33
- [24] Zheng Y, Kikuchi K, Yamasaki M, Sonio K and Uehara K 1997 Two-layer wideband antireflection coating with an absorbing layer *Appl. Opt.* **36** 6335–9
- [25] Hovel H J 1976 Transparency of thin metal films on semiconductor substrates *J. Appl. Phys.* **47** 4968–70
- [26] Yeh Y C M, Ernest F P and Stirn R J 1976 Practical antireflection coating for metal-semiconductor solar cells *J. Appl. Phys.* **47** 4107–12

- [27] Schneider M V 1966 Schottky barrier photodiodes with antireflection coating *Bell Syst. Tech. J.* **45** 1611–38
- [28] Hass G and Hadley L 1972 Optical constants of metals *American Institute of Physics Handbook* ed D E Gray (New York: McGraw Hill) pp 6.124–56
- [29] Hahn R E and Seraphin B O 1978 Spectrally selective surfaces for photothermal solar energy conversion. Reprint from *Phys. Thin Films* **10** 1–69
- [30] Fan J C C and Bachner F J 1976 Transparent heat mirrors for solar energy applications *Appl. Opt.* **15** 1012–17
- [31] Fan J C C, Bachner F J, Foley G H and Zavracky P M 1974 Transparent heat-mirror films of TiO₂/Ag/TiO₂ for solar energy collection and radiation insulation *Appl. Phys. Lett.* **25** 693–5
- [32] Bhargava B, Bhattacharya R and Shah V V 1977 A broad band (visible) heat reflecting mirror *Thin Solid Films* **40** L9–11
- [33] Holland L and Siddall G 1958 Heat-reflecting windows using gold and bismuth oxide films *Br. J. Appl. Phys.* **9** 359–61

Chapter 14

Other topics

14.1 Rugate filters

The term *rugate* is derived from biology where the meaning is essentially that of corrugated. It was introduced to describe a structure exhibiting a regular cyclic variation of refractive index resembling a sine or cosine wave. Such structures have the property of reflecting a narrow spectral region and transmitting all others. They exhibit properties similar to a quarter-wave stack but without the higher-order reflection bands. Thus they are notch filters and particularly useful in removing bright spectral lines from weaker continua. Many of their applications involve laser sources and they are especially relevant in the field of laser protection.

It can easily be shown that all the beams emerging from the front surface of a multilayer constructed from a series of quarter-wave layers of alternate high and low index are exactly in phase. This leads to high reflectance but it is limited in width in terms of wavelength or frequency because the constructive interference condition applies only to the wavelength for which the layers are exact quarter-waves. Outside the zone of high reflectance it is the transmittance that is high. The quarter-wave stack, therefore, acts as a notch filter. The lower the ratio of the high-to-low refractive index at the interfaces, the lower will be the amplitude reflection coefficients and the greater the number of beams required to achieve a given reflectance. The rate at which the interference condition decays with change in wavelength determines the width of the high reflectance zone. Smaller index contrast implies more beams, faster decay of the constructive interference and hence, narrower reflectance zones. A narrow zone of high reflectance in turn implies a large number of layers of low-index contrast. All this is considered in greater detail in chapters 4 and 5.

A limitation of systems made up of discrete dielectric layers is that a change in wavelength does not change the amplitude of the beams, except for slight changes due to dispersion. The same beams with the same amplitudes exist over a wide spectral region. It is impossible to distinguish between the interference

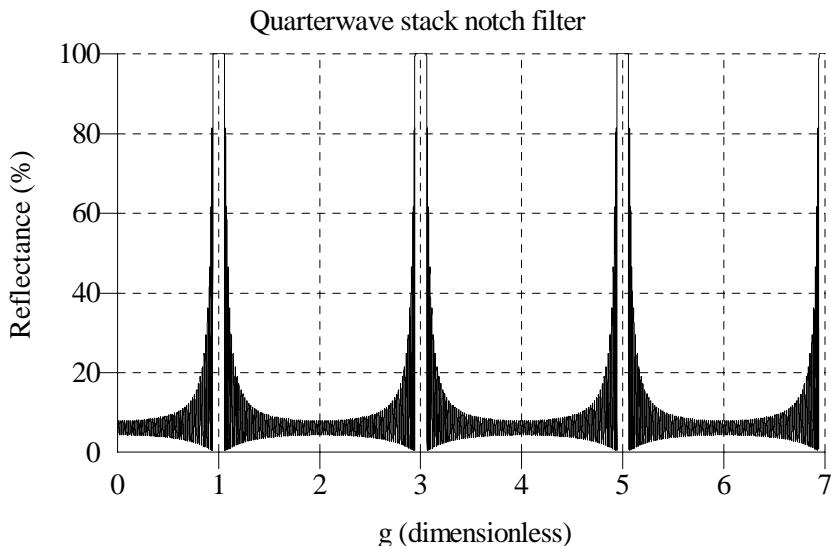


Figure 14.1. A typical characteristic of a quarter-wave stack used as a notch filter showing the higher orders at g of 3, 5 and 7. The fringes in the pass regions are so tightly packed they cannot be distinguished.

effect between two beams with phase difference φ and two beams of exactly the same amplitude and phase difference $\varphi \pm 2m\pi$ where m is an integer. In the case of the quarter-wave stack, the interference condition that exists at wavelength λ_0 also exists at wavelengths $\lambda_0/3, \lambda_0/5, \lambda_0/7$ and so on, leading to the higher-order reflectance zones that limit its usefulness as a notch filter. A typical characteristic curve plotted in terms of g , that is λ_0/λ , is shown in figure 14.1.

The higher orders may not present any problem in certain applications and for these the discrete layer design will be quite satisfactory. For those others where the peaks are a problem, we do need to suppress them. They have their origins in the interference between beams reflected at all the interfaces. In other words their origin is distributed throughout the multilayer. We need, therefore, a distributed solution. We need to retain the beams at the fundamental peak at $g = 1.0$, but we must remove them at all other integral values of g . An antireflection coating that does not affect the performance at $g = 1$, but that operates at values of g greater than unity, is required for each interface. An inhomogeneous layer is such an antireflection coating.

We shall return shortly to the derivation of performance of such systems. For the moment let us accept the two possible profiles for inhomogeneous layers shown in figure 14.2. If we assume that the layers have an optical thickness of one quarter wavelength at $g = 1.0$ then the performances, in terms of reflectance against g , are those shown in figure 14.3.

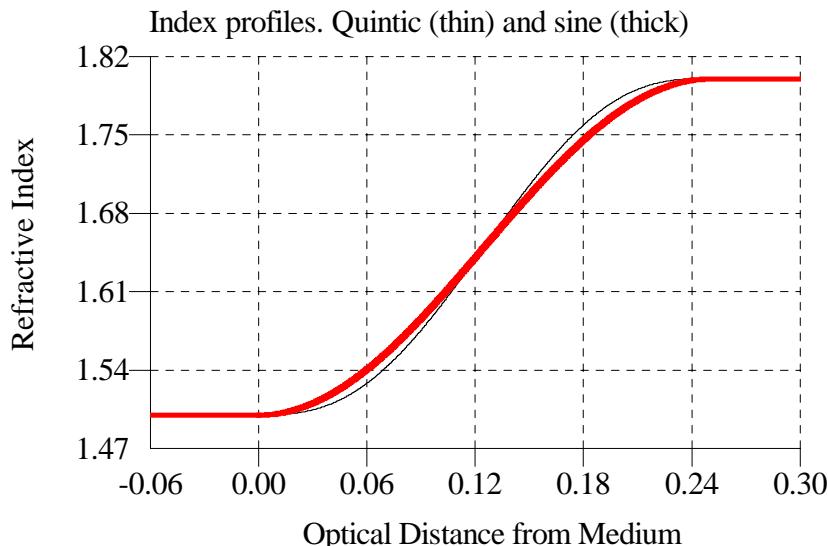


Figure 14.2. Inhomogeneous layer profiles rising from 1.50 to 1.80. The layers are one-quarter of a wavelength in optical thickness and the profile of refractive index follows a sine law (shallower curve—thick line) or a fifth-order polynomial (steeper curve—thin line) with zero first and second derivatives with respect to thickness at the end points.

This antireflection coating must now be inserted at each interface in the discrete layer coating. Figure 14.4 shows the resulting profile of optical admittance. The coating now has a sinusoidal variation of index throughout and is known as a rugate structure because of the smooth cyclic variation.

The new variation of reflectance is shown in figure 14.5. Note the small residual peak at $g = 2.0$. This is due to the failure of the sinusoidal variation of refractive index to act completely like the absentee half-wave layers of the discrete design. The slight residual reflectance change accumulates in a coating with a large number of layers and gives the slight perturbation from the regular fringe pattern that appears elsewhere. Southwell [1] has pointed out that an inhomogeneous layer based on an exponential sine does act as an absentee layer at even values of g , even though its profile is almost indistinguishable from that of a sine function.

The inhomogeneous antireflection coating is a very robust one from the point of view of errors. There is an insensitivity to the actual profile of the index. As long as the thickness at a given wavelength is greater than roughly a half-wave then the reflectance at that wavelength should be very low. Thus even quite large errors in the profile of a rugate filter are not normally serious unless they are systematic and lead to a change in the pitch of the cycle. Such errors tend to broaden the fundamental peak. Quite severe errors are required before the higher-

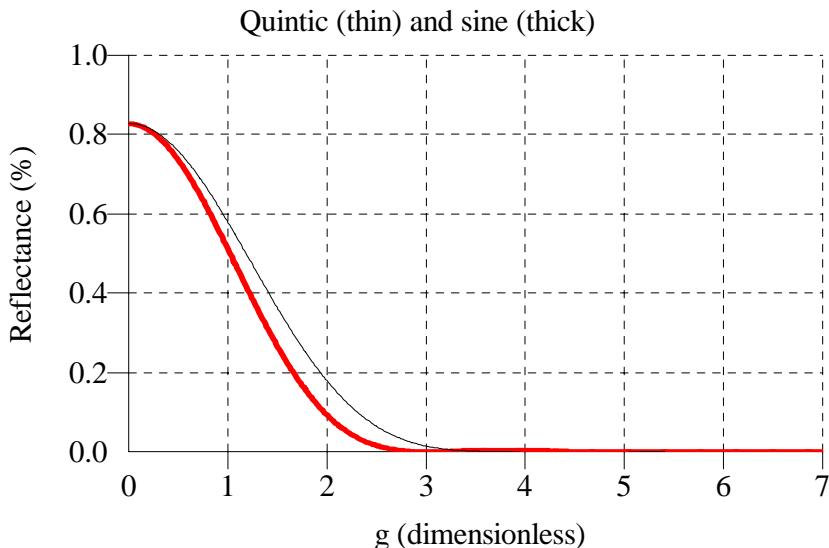


Figure 14.3. Reflectance against g for the inhomogeneous layers shown in figure 14.2. The sine law variation is less steep than the fifth-order polynomial so the curve of reflectance (left-hand curve—thick line) drops faster but the fifth-order polynomial (right-hand curve—thin line) gives lower reflectance at greater values of g .

order peaks begin to return. This has useful implications for the manufacturing of such filters.

The control of the deposition of rugate filters is a rather more involved task than for a simple discrete-layer quarter-wave stack. In discrete-layer deposition, it is optical thickness that has always been the object of the closed loop control system. Refractive index has been considered to be characteristic of the particular material being deposited and so the control of that aspect of the layers has been open loop. The deposition methods have concentrated on the control of source temperature, rate of deposition and so on. The rugate filter represents a greater challenge because there is no natural material that yields the desired profile of refractive index. It must be engineered. Compositional changes are necessary and, in the true rugate filter, these changes should be smooth. This tends to imply some form of active index control.

The absence of the need for direct index control, however, makes discrete layers very attractive. Although they are not strictly true rugates, nevertheless it is possible to create discrete-layer structures that have, up to a point, similar properties. To replace a rugate structure by a discrete-layer structure, we can imagine slicing a rugate period into a large number of thin layers of equal optical thickness. Each thin slice has an inhomogeneous index profile but we can convert it into a homogeneous index that has simply the central value. This gives a

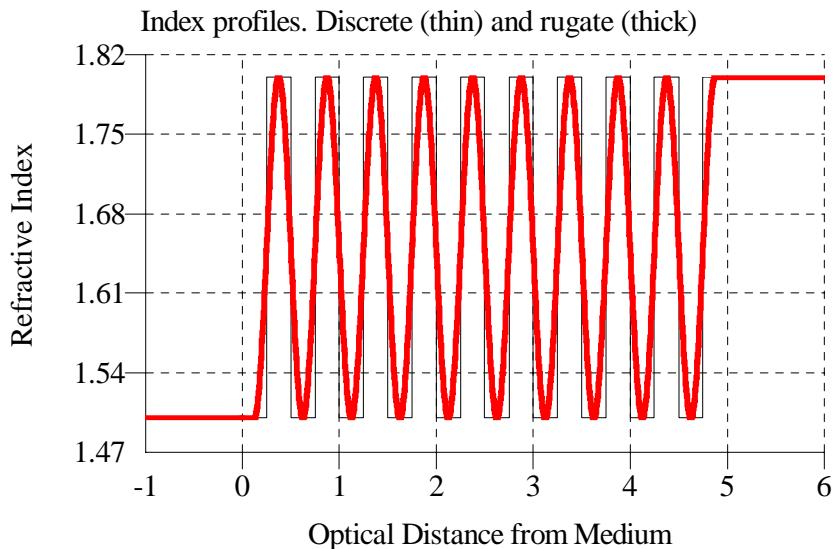


Figure 14.4. The result of replacing each discrete interface (square plot—thin line) by one graded to have a sine profile (rounded plot—thick line). This gives the rugate structure.

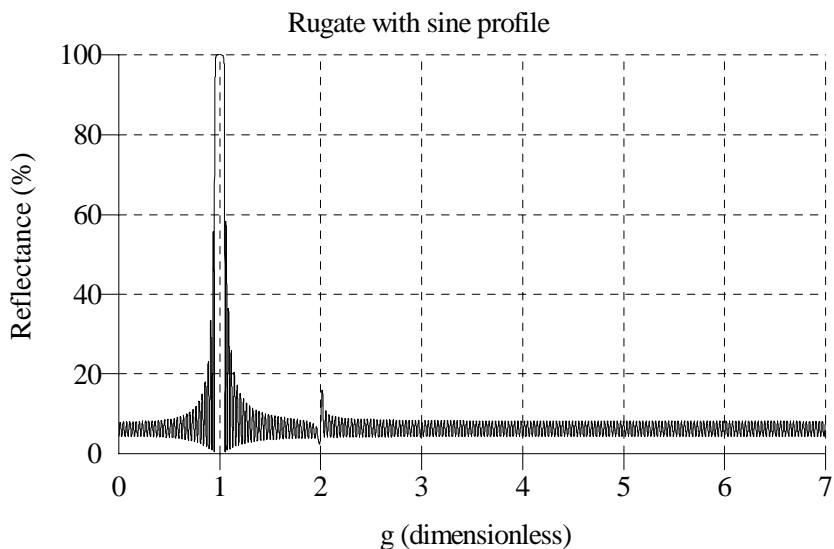


Figure 14.5. The reflectance curve of the rugate filter. The variation of index is shown in figure 14.4 except that the filter actually calculated had the equivalent of 64 discrete layers.

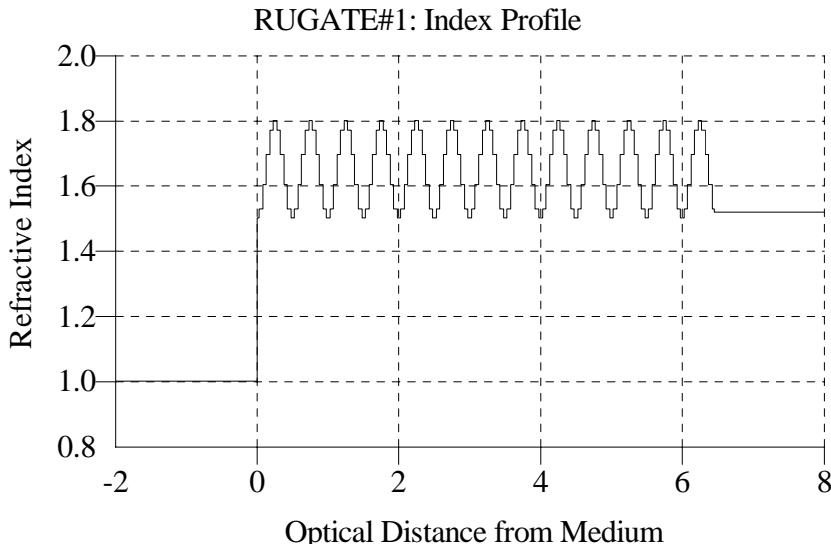


Figure 14.6. The profile of a rugate filter with a cycle consisting of ten discrete layers rather than a continuously varying profile.

staircase profile of index. In fact, and we return to this point later in this section, the calculation of the properties of rugate filters with arbitrary profile is normally carried out in this way with the thicknesses chosen to be so thin that further subdivision makes no changes to the results. Here we use rather thicker slices.

Figure 14.6 shows the profile of a rugate filter that has been converted in this way. The steps are arranged so that in each rugate cycle there are ten of them. This means that at the reflectance peak where the rugate cycle is one half-wave thick, the individual discrete layers are just one-twentieth of a wave thick. As long as the individual layers are thin compared with a quarter-wave, then the discrete version works well. However, as the wavelength reduces, the phase thickness of the individual layers increases and eventually becomes much thicker compared with a wavelength. However, the behaviour of the system does not just simply deteriorate but is quite regular and understandable. At a value of g of zero, the layers are effectively of zero phase thickness and so the reflectance of the system is that of the uncoated substrate. At $g = 1.0$, the rugate cycle is now a half-wave and the reflectance is high. As g increases, the cycle, at first, retains its antireflecting properties and the higher-order peaks are suppressed. Now let us jump to the case where g is large enough for the layers to be of half-wave thickness. Here we have absentee layers and the reflectance is that of the uncoated substrate. At this value of g we still have exactly the same beams taking part in the interference as at all other values of g . The phase shifts between them, however, are exactly the same as at $g = 0$ except that, in every case, there is an additional

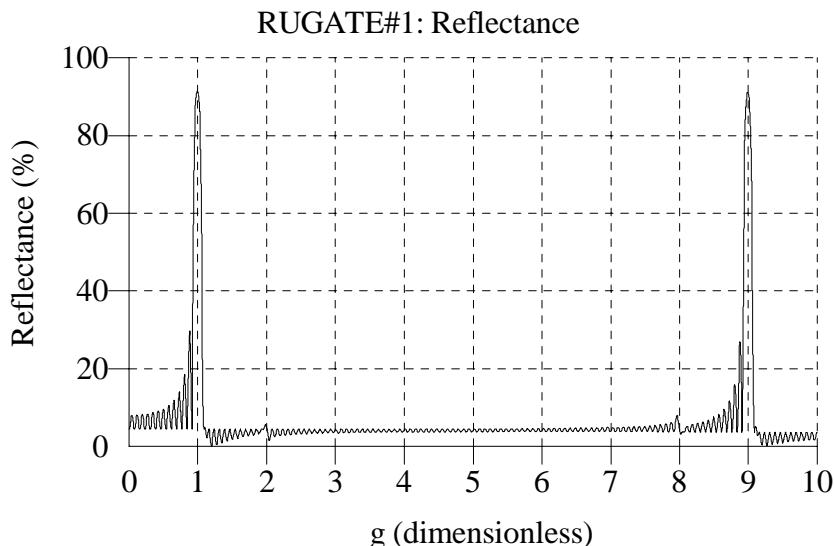


Figure 14.7. The performance of the rugate of figure 14.6 as a function of g showing the harmonic peak at $g = 9.0$. Note the subtle differences in the low reflectance performance from $g = 0$ to $g = 2$ and from $g = 8$ to $g = 10$. This is due [1] to the use in figure 14.6 of a half cycle that is the mirror image of the alternate half cycle only if the outer layers are half the thickness of the others.

wavelength, that is 360° , which is indistinguishable from zero. Furthermore, as we now reduce g from this value, we find exactly the same interference pattern as a function of the reduction in g that we find as a function of the increase in g from zero in the normal way. Thus, if we have ten equal steps or discrete layers making up the rugate cycle with a fundamental peak at $g = 1$, then there will be a similar peak at $g = 9$. A cycle made up of four layers will have a further peak at $g = 3$ and so on. Figure 14.7 illustrates this for the rugate of figure 14.6. Figure 14.8 shows similar performance for a rugate with a four-layer cycle. In this case the harmonics begin at $g = 3$ and so the sole peak that is eliminated is at $g = 2$. This may not appear to be any different from a two-layer cycle but, in fact, the extra layers help to suppress the half-wave-hole peak that appears at $g = 2$ when the coating based on the two-layer cycle is tilted.

Southwell [1] has pointed out that the slight lack of symmetry in the result in figure 14.7 is a consequence of the use of a set of sublayers of identical thickness such that there are no two adjacent sublayers with the same index. This effectively makes the rugate period symmetrical only if the two outermost layers are considered to be half the thickness of the others. A rearrangement where the outermost sublayers have the same index and the full sublayer thickness, implying a merging of the innermost layer pair and the ending layer of each cycle with the

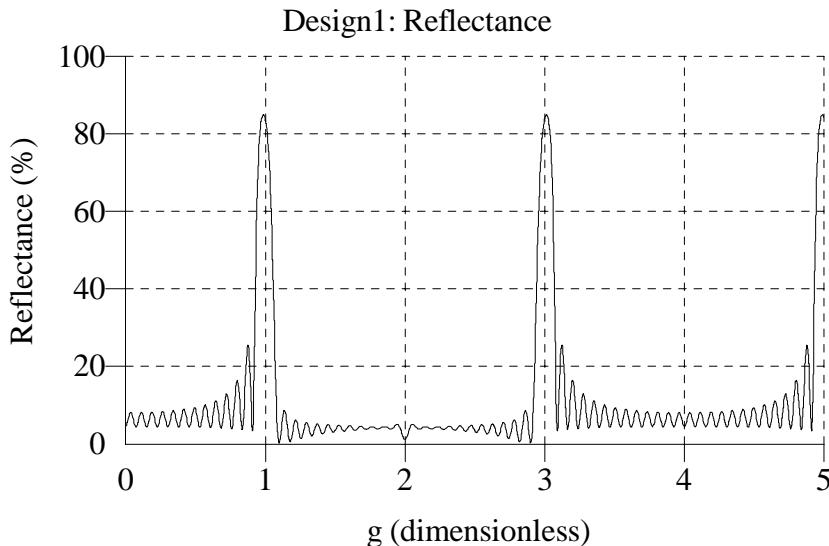


Figure 14.8. The performance of a rugate similar to that of figure 14.6 except that the cycle is made up of four discrete layers of equal thickness. The harmonic peak appears now at $g = 3.0$.

starting of the next, gives a perfectly symmetrical performance.

An alternative technique for the replacement of the continuous variation with a series of discrete layers uses two materials with fixed indices of refraction. One of the indices must be equal to or less than the lowest in the rugate structure and the other equal to or greater than the highest. The method uses the properties of the characteristic matrices of the films. There are two variants. The first uses the result that the matrix of any symmetrical arrangement of layers, absorbing and inhomogeneous layers included, can be replaced by the characteristic matrix of a single equivalent homogeneous layer [2, 3]. This equivalence is dealt with more fully in chapter 3 and is a purely analytical relationship and certainly not physical, but it is valid wherever the properties involve only the characteristic matrices. This relationship can be reversed so that the homogeneous film matrix can be replaced by the matrix of a symmetrical combination of layers. Since the eventual result involves identical matrices, properties such as reflectance and transmittance at one particular angle of incidence and wavelength are unchanged when the equivalent sequences are interchanged. One of the most useful aspects of this relationship is the replacement of a layer of intermediate index by a symmetrical combination of layers of given high and low index. At one angle of incidence and one wavelength this equivalence holds completely for any property that can be calculated using the characteristic matrices. For the equivalence to be retained exactly with changes in wavelength demands a particular dispersion

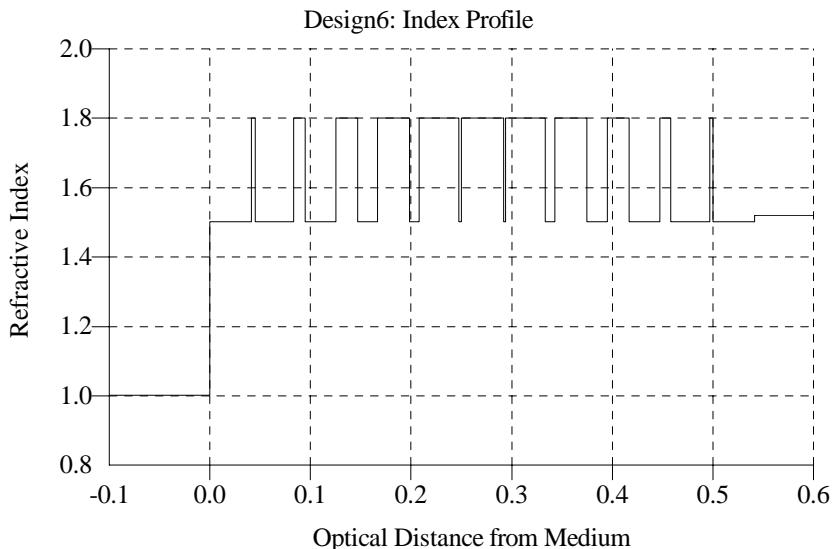


Figure 14.9. A 22-layer representation of a single half-wave rugate cycle. The layers are either of high (1.8) index or low (1.5) and their thicknesses are varied so that the overall effect is similar to the smooth variation of the classical rugate.

of the indices of the replacement layers. This implies that when real layers are involved with their natural dispersion the equivalence becomes gradually poorer as the wavelength changes, especially as the wavelength decreases. The equivalence strictly does not extend to changes in angle of incidence although the deterioration is not usually very rapid. The second variant uses an approximate method based on pairs of layers. When both members of a layer pair are thin compared with a wavelength then the characteristic matrix of the combination of the two layers is equivalent to that of a single layer of intermediate index [4]. Again this relationship is not valid for changes in angle of incidence and it becomes poorer as the wavelength decreases. Both variants can take the staircase approximation to the rugate cycle and convert it into an equivalent series of alternate high- and low-index layers of differing thicknesses.

We illustrate the method by using the second variant, the two-layer approximation. Figure 14.9 shows a single cycle. (There is an extra layer at the end that is strictly the first layer of a following cycle.) The performance of a rugate filter based on 14 of these cycles in series is shown in figure 14.10.

The important point about these calculations is that a discrete-layer approximation to a rugate filter can give performance that is nevertheless acceptable. The range of transparency of the materials is rarely greater than the clear ranges shown in figures 14.7 and 14.9. Many of the techniques for coating production lend themselves much better to the construction of discrete-

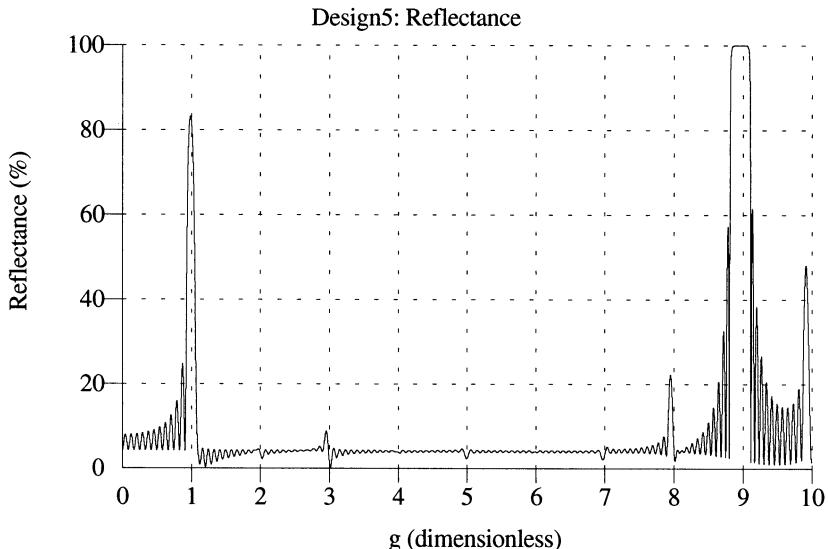


Figure 14.10. The performance of the rugate of figure 14.9. The performance has characteristics similar to those of the stepped version from which it was derived.

layer systems than to the creation of smoothly varying index profiles.

We now consider the theoretical problems in more detail. Figure 14.11 shows a representation of an inhomogeneous layer that is linking two media. The optical admittance, y , is plotted against the optical thickness, z . Accurate calculation of such layers involves the slicing of them into sufficiently thin homogeneous sublayers and then using the normal calculation techniques. The slices should be rather thinner than a quarter-wave at the shortest wavelength in the calculation. To test the adequacy of the approximation, the layers can be made still thinner and the calculation repeated. A completely unchanged performance is an indication that the approximation is satisfactory. For the design of such structures it is usual to employ an approximate technique based on what is essentially an application of the vector method. If the performance is to be calculated at the plane denoted by $z = 0$ then the vector that is derived from the step at the plane z will be given by

$$\rho \exp(-i2\delta) = \frac{\Delta y}{2y} \exp(-i2\kappa z)$$

where κ , the wavenumber, is given by $2\pi/\lambda$, λ being the free space wavelength. If we represent twice the optical thickness, z , by x then we can write the sum of all the various vectors as

$$\sum \frac{\Delta y}{2y} \exp(-i\kappa x). \quad (14.1)$$

In the simple vector method this sum is simply equal to the amplitude reflection coefficient. However, when many such vectors are involved with a quite thick inhomogeneous structure, a correction may be made that represents a better approximation. The conversion of the sum of (14.1) to an integral then yields

$$\int_{-\infty}^{\infty} \frac{dy}{dx} \left(\frac{1}{2y} \right) [\exp(-ikx)] dx = Q(\kappa) \exp[i\varphi(\kappa)] \quad (14.2)$$

connecting a function of performance with a function of the distribution of characteristic admittance through a Fourier integral expression. This may be inverted so that the distribution of y may be calculated from the distribution of performance. Q is a function of performance, $\kappa = 2\pi/\lambda$, and x is twice the optical path. $\varphi(\kappa)$ is a phase factor that must be an odd function to ensure that $n(\kappa)$ is real. Although multiple beam effects are neglected, a judicious choice of Q can reduce the errors that arise from this approximation. Note that equation (14.2) is frequently written with a positive argument for the exponential. This is simply a consequence of the particular sign convention that is used.

Functions that have been proposed and used for Q include (the first represents the simple amplitude reflection coefficient):

$$\begin{aligned} Q &= \sqrt{R} \\ Q &= \sqrt{\frac{R}{T}} \\ Q &= \sqrt{\frac{1}{2} \left(\frac{1}{T} - T \right)} \\ Q &= \sqrt{\frac{1}{\sqrt{T}} - \sqrt{T}}. \end{aligned} \quad (14.3)$$

The great advantage of this approach is the analytical connection in either direction of a function of design with a function of performance. If we know the performance we can find a design and vice versa. Disadvantages are that the technique is approximate and considerable skill and experience are required in the choice of the appropriate Q function and phase factor φ . Although the resulting design is a continuously varying admittance profile, it can be converted into a discrete-layer design, the thicknesses being chosen thin enough not to affect performance at the shortest wavelength of interest.

Finally we note that the term rugate is sometimes used for any layer system in which there is a deliberate attempt to induce an inhomogeneity, whether or not it is of a cyclic kind, in spite of the rather more restricted meaning of the original term.

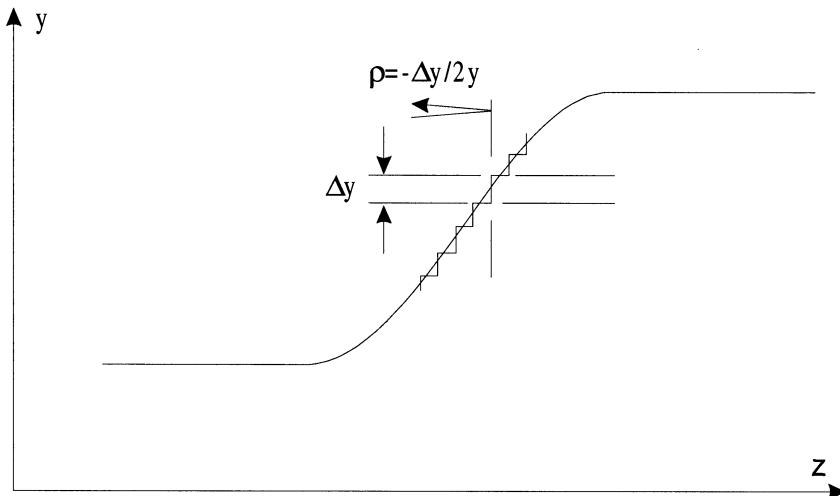


Figure 14.11. To derive an expression for the performance of a dielectric inhomogeneous layer we first divide the layer into a series of separate steps. These steps are chosen close enough so that closer spacing still yields an unchanged result. Each step has an amplitude reflection coefficient of $-\Delta y/(2y)$.

14.2 Ultrafast coatings

Traditionally, coating designers have been able to rely on the steady-state nature of the effects they seek to produce. There are now laser systems, known as ultrafast, capable of generating pulses of light that are short enough for transient response to become significant. A normal high reflector consisting of a quarter-wave stack might be some 25 quarter-waves in thickness. At a wavelength of $1 \mu\text{m}$ this implies a trip length for light travelling from the front to the rear of the coating and back again of $12.5 \mu\text{m}$ or a trip time of around 42 fs (one femtosecond is $1/1000$ picoseconds). Pulses that are around 50 fs in length are now common and the shortest current pulses are some 5 fs in length. It is clear that the transient response of coatings must now be considered important in such applications, but the effects, in fact, can be significant even with pulses some two or three orders of magnitude longer. The idea that coating properties should have an influence on short pulses and that they might be engineered to have prescribed effects is not new. It is, however, only recently that the field has expanded and the technology advanced to the stage where the application is becoming of major importance.

A short pulse can be thought of as an envelope over a carrier. The carrier contains the phase information associated with the pulse and it travels at what is known as the *phase velocity*. The energy is obviously associated with the envelope that travels at what is known as the *group velocity*. In the presence of dispersion, the group velocity and the phase velocity are different, normal dispersion making

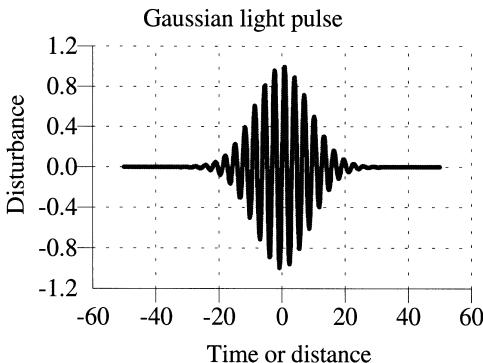


Figure 14.12. A short Gaussian-shaped pulse consisting of an envelope over a carrier of constant frequency. The carrier phase may move faster than the pulse when it will appear to run through the envelope as it travels.

the phase velocity greater. Thus the carrier appears to run through the pulse envelope. A short pulse with Gaussian envelope is shown in figure 14.12.

The pulse may also be visualised in a different way, as a collection of monochromatic component waves with a continuous distribution of frequencies over a given band. The coherent combination of these monochromatic waves yields the envelope and carrier of the alternative model. Both of the models are entirely equivalent and, if we wish, we can pass from one to the other by way of a Fourier transform.

Pulse envelopes frequently have a Gaussian shape [5, 6]. For simplicity we can look at the temporal variation at the origin of our coordinates, $z = 0$, and then, if the peak of the pulse corresponds to $t = 0$,

$$F(t) = \mathcal{A}e^{-\frac{t^2}{2\mu^2}} \quad (14.4)$$

where μ has the dimension of time. The Fourier transform gives the frequency distribution and it is also a Gaussian function,

$$G(\omega) = \mathcal{B}e^{-\frac{\mu^2(\omega-\omega_0)^2}{2}}. \quad (14.5)$$

If the time between the half-maximum points is τ and the width of the pulse (angular) frequency distribution also at half-maximum is $\Delta\omega$ then

$$\tau \Delta\omega = 4\ln_e 2.$$

Note that both these quantities are functions of μ . For example,

$$\tau = (2\sqrt{\ln_e 2})\mu.$$

The centre of the pulse is the point where all of the component waves are exactly of identical phase. If all the component waves travel at the speed of light *in vacuo* then the phase coincidence will also travel at that speed and the centre of the pulse will move with it. Similarly if all waves slow down equally then the pulse will slow down to the same extent but will otherwise be unchanged.

The relative phase of the carrier within the pulse is set by the value of the phase where all the component waves coincide. If the phase of the waves is zero then the carrier will have a peak exactly at the peak of the pulse. We can find the position of the pulse peak at any time by a simple procedure.

The pulse can be considered to be made up of monochromatic component waves. As these propagate the phase relationships between them will change, but if the pulse shape is unaltered as it propagates then at any particular time there must be a distance along the path where the phase is identical for all the component waves, and this must correspond to the pulse centre. We use the normal thin-film convention of $(\omega\tau - \kappa z)$ in the phase factor where $\kappa = 2\pi n/\lambda$ with λ the free space wavelength. We write the component wave phase at distance z and time t as $\varphi - \varphi_0 + \Delta\varphi$. Then for coincidence of all component phases, $\Delta\varphi$ must be zero.

This condition is

$$\begin{aligned} (\omega_0 + \Delta\omega)t - (\kappa_0 + \Delta\kappa)z &= \varphi_0 + \Delta\varphi \\ \omega_0 t - \kappa_0 z &= \varphi_0 \\ \Delta\varphi &= 0 = \Delta\omega t - \Delta\kappa z \\ z &= \frac{\Delta\omega}{\Delta\kappa}t = v_g t. \end{aligned} \tag{14.6}$$

The quantity $\Delta\omega/\Delta\kappa$ is known as the group velocity, v_g , and clearly it must remain constant if the position z is to be the same for all the component waves and the shape of the pulse is to remain unchanged.

An alternative visualisation involves a simple diagram. We plot the z -direction horizontally and ω vertically. We sketch the bundle of component waves making up the pulse as a set of lines through the appropriate values of ω and parallel to the z axis. We mark contours of constant φ on the lines. Provided there is one contour that runs normally across the lines then the pulse peak will be positioned there and the pulse shape will be unchanged.

In a nondispersive medium the phase at the peak will be zero because all the component waves will be travelling at identical velocity even though it may be less than the velocity in free space. In a dispersive medium, the component waves travel at different velocities according to the particular value of refractive index. Provided the variation in velocities still permits a phase coincidence somewhere, then the pulse will appear there and will be unchanged in shape, although the phase of the carrier wave will be altered. It is clear from equation (14.6) that the critical condition is for the group velocity to remain constant across the frequency spectrum of the pulse.

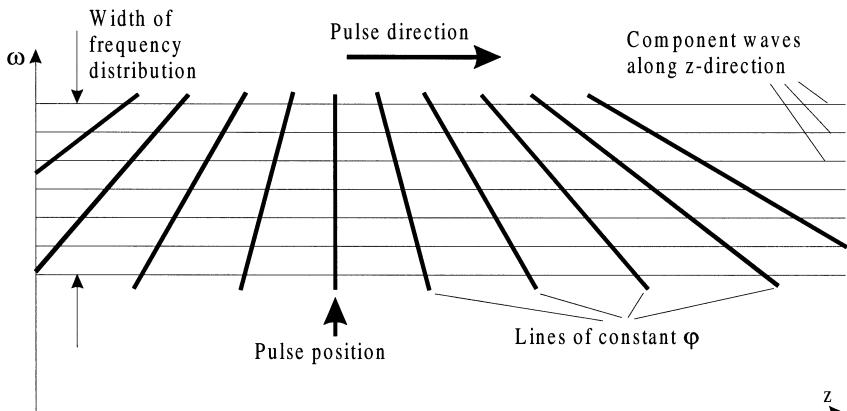


Figure 14.13. Sketch showing the component waves of the pulse as horizontal lines along the direction of propagation and with their relative phases marked as contour lines across them. The pulse peak coincides with the position where the phase of all the components is exactly equal.

In a dispersive medium, the refractive index changes with frequency. We can calculate the group velocity in terms of this change.

$$\kappa = \frac{2\pi n(\omega)}{\lambda} = \frac{\omega n(\omega)}{c}$$

$$\begin{aligned} \frac{dk}{d\omega} &= \frac{n(\omega)}{c} + \left(\frac{\omega}{c} \right) \frac{dn(\omega)}{d\omega} \\ v_g &= \frac{c}{n(\omega) + \omega \frac{dn(\omega)}{d\omega}}. \end{aligned} \quad (14.7)$$

In a medium with normal dispersion, this is not constant.

There is thus no guarantee that the group velocity should be constant with changing frequency. If the second derivative of ω with respect to κ is nonzero then there can be no phase coincidence and the pulse will be perturbed. Again we can consider the operation in two different equivalent ways. If we limit ourselves to the second derivative then we can write the expression for the phase of an arbitrary component wave as:

$$(\omega_0 + \Delta\omega) t - \left(\kappa_0 + \Delta\omega \frac{dk}{d\omega} \Big|_0 + \frac{1}{2} (\Delta\omega)^2 \frac{d^2\kappa}{d\omega^2} \Big|_0 \right) z = \varphi + \Delta\varphi \quad (14.8)$$

and we can immediately identify a problem. The third term in the coefficient of z is even in $\Delta\omega$ and so cannot be compensated by the other terms. We must

therefore split the frequency distribution of the pulse into two parts, one with positive $\Delta\omega$ and the other with negative $\Delta\omega$, and look at each separately. In each case we ensure that the value of $\Delta\varphi$ is zero. This gives two equations instead of the usual one. We keep the value of z the same in each and introduce a different time t representing the interval in time between the pulse centres that correspond to each part of the split distribution. If the spectral width of the split distribution were halved then each component pulse would have twice the basic pulse width. As a crude correction for this effect, therefore, we treat the $\Delta\omega$ in the following expressions as the width of the frequency distribution of the basic initial pulse.

$$\begin{aligned}\Delta\omega t_1 - \Delta\omega \frac{d\kappa}{d\omega} \Big|_0 z - \frac{1}{2}(\Delta\omega)^2 \frac{d^2\kappa}{d\omega^2} \Big|_0 z &= 0 \\ -\Delta\omega t_2 + \Delta\omega \frac{d\kappa}{d\omega} \Big|_0 z - \frac{1}{2}(\Delta\omega)^2 \frac{d^2\kappa}{d\omega^2} \Big|_0 z &= 0.\end{aligned}$$

Then, since

$$\begin{aligned}\frac{d}{d\omega} \left(\frac{d\kappa}{d\omega} \right) &= \left(-\frac{1}{v_g^2} \right) \frac{d}{d\omega} (v_g) \\ \Delta t = (t_1 - t_2) &= -\Delta\omega \left(\frac{d^2\kappa}{d\omega^2} \right) z = \Delta\omega \left(\frac{dv_g}{d\omega} \right) \left(\frac{1}{v_g^2} \right) z\end{aligned}\quad (14.9)$$

and the result, (14.9), is similar to that of a much more strict derivation using Gaussian pulses. The pulse is broadened and the carrier frequency of each part of the pulse is different. A pulse with a varying carrier frequency along its length is said to be chirped.

Alternatively we can use the diagram to see the way in which the phase coincidences are affected by the variation of group velocity. Figure 14.14 shows the modified arrangement of the various component waves and their contours of equal phase. The phase broadening itself causes a widening of the pulses corresponding to each band of frequencies and so there is a still greater broadening as the pulse propagates.

The effect, because it is due to a change in the group velocity across the frequency range of the pulse, is usually known as group velocity dispersion. Similar effects occur in waveguides and optical fibres. Group velocity dispersion, often abbreviated to GVD, is measured in units of $(\text{time})^2 (\text{length})^{-1}$ and is given by

$$\text{group velocity dispersion} = \frac{d^2\kappa}{d\omega^2} \Big|_0. \quad (14.10)$$

If the original pulse is of Gaussian shape as in (14.4) then if we write:

$$\tau_g^2 = \frac{d^2\kappa}{d\omega^2} \Big|_0 z$$

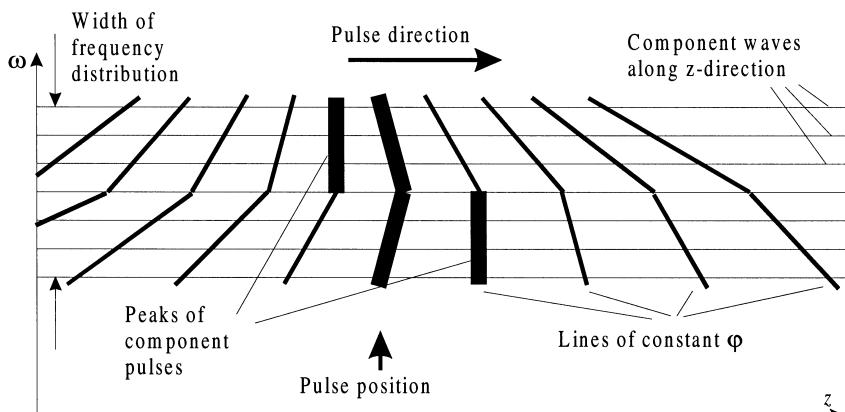


Figure 14.14. The pulse frequency distribution is now split into two parts, each of which represents a component pulse with its own centre position. Since the group velocity is different for the two component pulses they separate such that one lags behind the other and the combined pulse is broadened.

it can be shown [5] that the new pulse width is given by

$$\tau_{\text{new}} = \tau \left[1 + \frac{\tau_g^4}{\mu^4} \right]^{\frac{1}{2}}. \quad (14.11)$$

All of these effects are linear and so they can be undone by a similar but opposite effect. Further, the order in which the effects occur is unimportant. A dispersive broadening may be cancelled by an opposite dispersion.

A pulse, consisting of an envelope over a carrier, may be subjected to a modification, by passing through a crystal modulator for example, in which the phase of the carrier is gradually varied throughout the length of the pulse. If this variation is a linear function of time then the effect is just as though the frequency of the carrier had been changed. There is little other effect. However, if the phase is changed as a quadratic function of time then it is as though the frequency of the carrier were shifted gradually throughout the length of the pulse [6]. The pulse has sliding frequency and is therefore *chirped*.

$$\cos(\omega t + at^2) = \cos[(\omega + at)t] \quad (14.12)$$

has frequency $(\omega + at)$. This chirped pulse appears indistinguishable from a short pulse that has been dispersion broadened, except that the apparent dispersion can be opposite in sign to normal dispersion. The pulse can then be subjected to the action of a dispersive medium where there is significant group velocity dispersion. Provided this dispersion is of the correct magnitude and sense then it will undo the

artificially induced effect in the pulse leaving it considerably narrowed. Various components have been used for this purpose but the flexibility of optical coatings makes them particularly attractive in this application [7–9].

Optical coatings affect both the amplitude and the phase of incident light. They can therefore, in principle, make the kinds of adjustments to incident light that we have been considering. They have an advantage over dispersive systems in that the correction is made immediately. We first must consider the nature of the effect that thin-film coatings have on the pulse.

Amplitude reduction over part of the range of frequencies leads to pulse broadening because the narrower the frequency spectrum the broader is the pulse. We therefore limit ourselves to consideration of those systems that have flat performance in terms of either transmittance or reflectance and that make adjustments to the phase. The sign convention is important. We use the normal thin-film convention.

The coordinate system has its origin at the surface where the reflection is said to be taking place and the phase shift is measured at that surface. The electric field retains its incident positive direction. An incident wave, say, $\mathcal{E} \cos(\omega t - \kappa z + \varphi_{\text{inc}})$, say, suffers a phase change φ_{ref} at the surface $z = 0$. The electric field at that surface for the reflected beam therefore becomes $\mathcal{E} \cos(\omega t - \kappa z + \varphi_{\text{inc}} + \varphi_{\text{ref}})$. This then forms a reflected beam that has expression $\mathcal{E} \cos(\omega t + \kappa z + \varphi_{\text{inc}} + \varphi_{\text{ref}})$. The returned beam is now propagating along the negative direction of the z axis. We can avoid the sign change in z if we introduce the idea of the total path travelled by the wave that we denote by x , which always increases as the wave propagates and is along the positive direction of the z axis before reflection and along the negative direction after reflection. (Note the temptation when using the alternative phase factor convention of $(\kappa z - \omega t)$ to reverse the direction of the wave by incorrectly writing $(\kappa z + \omega t)$, reversing the direction of time rather than, correctly, $(-\kappa z - \omega t)$, reversing the propagation direction.)

The expression for the wave now becomes

$$\mathcal{E} \cos(\omega t - \kappa x + \varphi_{\text{inc}} + \varphi_{\text{ref}}) \quad (14.13)$$

where x is always positive for increasing propagation length.

Now let us examine the effects of the various phase angles on the pulse and its components. We take equations (14.7) and we rewrite the left-hand side to include a change of phase on reflection. Then

$$\begin{aligned} & \omega_0 t - \kappa x + \Delta \omega t - \Delta \omega \frac{d\kappa}{d\omega} \Big|_{\omega_0} x - \frac{1}{2} (\Delta \omega)^2 \frac{d^2 \kappa}{d\omega^2} \Big|_{\omega_0} x + \varphi_0 + \Delta \omega \frac{d\varphi}{d\omega} \Big|_{\omega_0} \\ & + \frac{1}{2} (\Delta \omega)^2 \frac{d^2 \varphi}{d\omega^2} \Big|_{\omega_0} \\ & = (\omega_0 t - \kappa x) + \Delta \omega \left\{ t - \left(\frac{d\kappa}{d\omega} \Big|_{\omega_0} x - \frac{d\varphi}{d\omega} \Big|_{\omega_0} \right) \right\} \end{aligned}$$

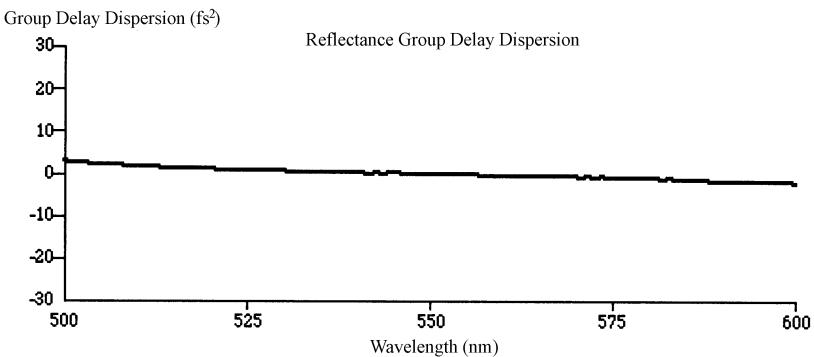


Figure 14.15. The calculated group delay dispersion for a 19-layer classical quarter-wave stack of zinc sulphide and cryolite, the zinc sulphide outermost. Reference wavelength is 550 nm. The effect is clearly quite small and this is normal for quarter-wave stacks in general.

$$-\frac{1}{2} (\Delta\omega)^2 \left(\left. \frac{d^2\kappa}{d\omega^2} \right|_{\omega_0} x - \left. \frac{d^2\varphi}{d\omega^2} \right|_{\omega_0} \right). \quad (14.14)$$

$-(d\varphi/d\omega)$ has units of time and we can identify it as equivalent in its effect to the group delay due to dispersion and it is therefore known as the group delay, sometimes abbreviated to GD. The next term, $-(d^2\varphi/d\omega^2)$ has an effect equivalent to the group velocity dispersion. Since the negative first derivative is known as group delay this second derivative is known as group delay dispersion, abbreviated to GDD, and has units of (time)². Although we have said little about it here, the third derivative is sometimes called the third-order dispersion, with units of (time)³, and abbreviated to TOD. Third-order dispersion is usually small but, if it is significant, it can adversely affect the shape of the pulse. The group delay dispersion is particularly important because it can be adjusted in sign and therefore can be used to offset the effects of group velocity dispersion and also to operate on chirped pulses.

For most simple reflectors, φ increases with wavelength. This is the case with the classical quarter-wave stacks. φ increases slowly with λ , the rate of change being a minimum at the central wavelength, and the greater the index contrast in the layers, the slower the change. An outer low-index layer actually reduces still further the rate of change. The calculated group delay dispersion for a zinc sulphide and cryolite quarter-wave stack is shown in figure 14.15. The outermost layer in this case is zinc sulphide. Cryolite outermost leads to a slight gain but gives an antinode of electric field at the outer surface and may, therefore, be undesirable. It is obvious that the calculated group delay dispersion for quarter-wave stacks will normally be very small and so it is a particularly safe type of reflector to use with short pulses.

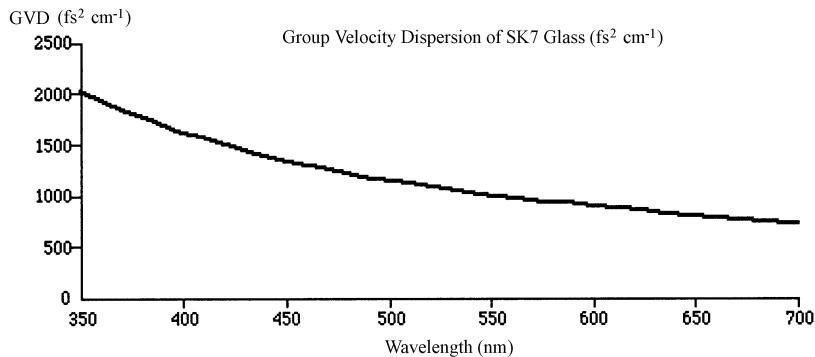


Figure 14.16. The group velocity dispersion in $\text{fs}^2 \text{ cm}^{-1}$ for SK7 glass calculated from the manufacturer's data.

Transparent optical materials with normal dispersion show a refractive index n that reduces as wavelength increases. The rate of reduction, however, falls with increasing wavelength through most of the transparent region and so the second derivative of n with λ is positive. Since

$$\kappa = \frac{2\pi n}{\lambda}$$

the group velocity dispersion is

$$\frac{d^2\kappa}{d\omega^2} = \left(\frac{\lambda^3}{2\pi c^2} \right) \left(\frac{d^2n}{d\lambda^2} \right). \quad (14.15)$$

For typical optical materials the group velocity dispersion can be of the order of $1000 \text{ fs}^2 \text{ cm}^{-1}$. Figure 14.16 shows the group velocity dispersion calculated from the manufacturer's data for SK7 glass [10].

The net group delay dispersion is given by

$$\left(\frac{d^2\kappa}{d\omega^2} \Big|_0 L - \frac{d^2\varphi}{d\omega^2} \Big|_0 \right). \quad (14.16)$$

Straightforward quarter-wave stacks show small group delay dispersion implying that although useful in reflecting short pulses, it is not likely to be useful in compensating for the group velocity dispersion of a reasonable thickness of optical material. Some way of increasing the magnitude of the negative values of group delay dispersion of an optical coating is required. The addition of a weak cavity to the front of the quarter-wave stack has been shown to be one fairly successful way of achieving this result provided the wavelength region is limited, that is the pulse is reasonably long. Such an arrangement is usually known as a Gires–Tournois interferometer after the originators [11, 12]. The weak cavity

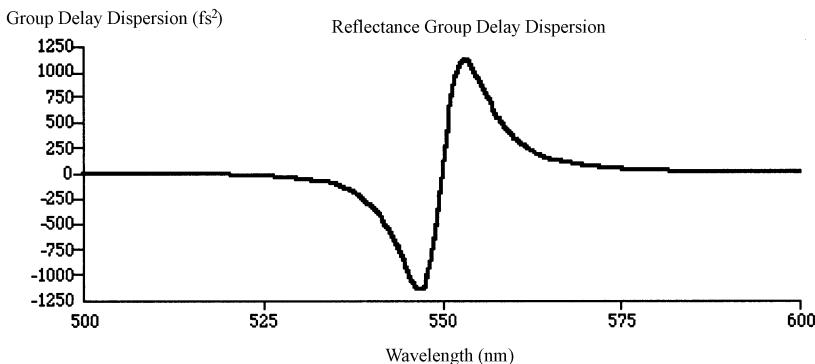


Figure 14.17. The group delay dispersion calculated for the coating in expression (14.17).

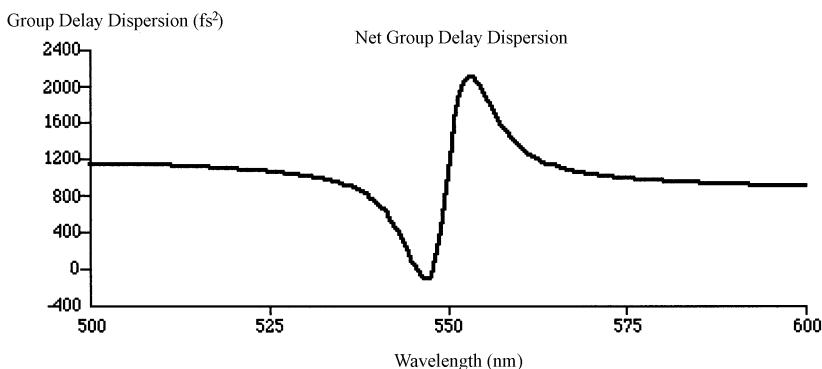


Figure 14.18. The resultant group delay dispersion for the system of SK7 and coating. Over a short spectral region the group delay dispersion has been reduced to the vicinity of zero.

does not reduce the reflectance too much but the effect is a very rapid change of phase on reflection that leads to the desired effect.

We can assume a 1-cm thick slice of SK7 glass and attempt the compensation of the resulting group delay dispersion by the use of the interferometer. Figure 14.17 shows the group delay dispersion of a Gires–Tournois interferometer of design

$$\text{Air} | H \text{L} 6\text{H} (L\text{H})^9 | \text{Glass} \quad (14.17)$$

using zinc sulphide and cryolite as materials. Over a limited region the group delay dispersion is capable of compensating for the effect of the 1 cm of SK7. Figure 14.18 shows the composite group delay dispersion and it is near zero at wavelengths just shorter than 550 nm, the central wavelength of the interferometer.

Table 14.1. Design of chirped reflector. (Courtesy of Thin Film Center Inc.)

λ_0 700 nm					
Layer	Material	Optical thickness	Layer	Material	Optical thickness
Medium	Air	Massive			
1	TiO ₂	0.048	13	TiO ₂	0.282
2	SiO ₂	0.239	14	SiO ₂	0.285
3	TiO ₂	0.336	15	TiO ₂	0.275
4	SiO ₂	0.208	16	SiO ₂	0.291
5	TiO ₂	0.231	17	TiO ₂	0.306
6	SiO ₂	0.197	18	SiO ₂	0.324
7	TiO ₂	0.225	19	TiO ₂	0.362
8	SiO ₂	0.292	20	SiO ₂	0.320
9	TiO ₂	0.292	21	TiO ₂	0.355
10	SiO ₂	0.287	22	SiO ₂	0.323
11	TiO ₂	0.279	23	TiO ₂	0.273
12	SiO ₂	0.288	Substrate	Glass	Massive

This version of the interferometer is quite weak in its effect. It is possible to increase the group delay dispersion by much more than an order of magnitude by appropriate design so that the effect of much greater thicknesses of material can be accommodated. The limitation of the interferometer is its rather small spectral range of correction so that its principal application must be to longer pulses.

The principle of coatings of this type is that light may penetrate into them to a rapidly varying extent and therefore show rapid phase dispersion, which in turn is translated into the high group delay dispersion that is required for the system. Broadband reflectors with extended zones also exhibit this effect, and incidentally may have a considerable broadening effect when used as simple reflectors. They are, however, useful for operating on chirped pulses [8, 13] and because they often have a structure that exhibits a gradual tapering of layer thickness through the structure they are often known as chirped mirrors. Table 14.1, figures 14.19 and 14.20 show the details of the design and calculated performance of such a coating with a group delay dispersion of -30° over the region 750–900 nm. This is an example of a design arrived at purely by synthesis with no starting information other than the materials silica and titania that were to be used. Szipöcs and Köházi-Kis [14] give a detailed account of a more systematic approach to the design of such chirped mirrors.

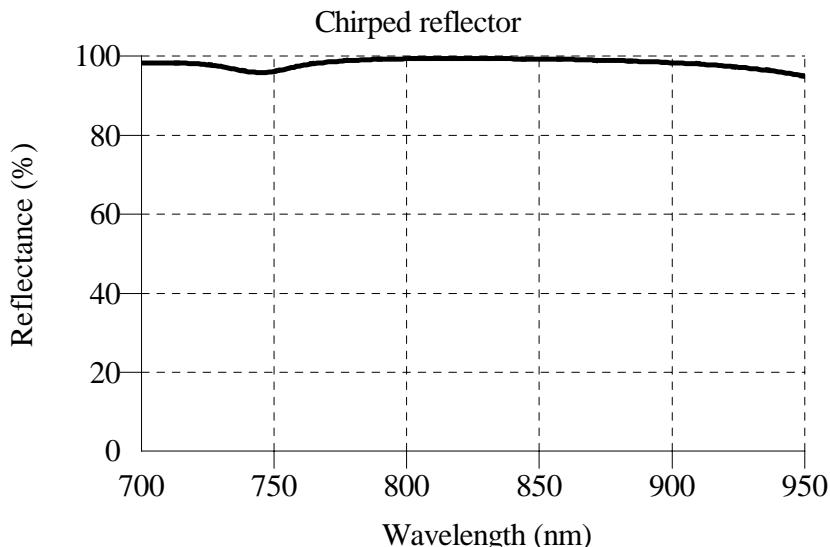


Figure 14.19. Calculated reflectance of the coating of table 14.1.

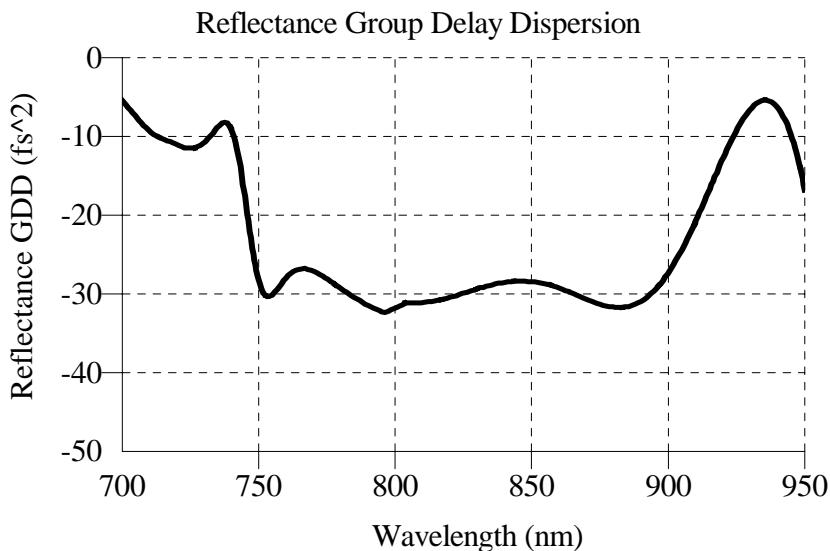


Figure 14.20. Calculated group delay dispersion of the coating of table 14.1

14.3 Automatic methods

Given a possible solution to a thin-film design problem, can we devise an objective method to change the parameters so that it becomes a better design? Can

we continue the process to make the design as good as possible? And, of course, can we finally devise a way of achieving all this using an automatic computer? The answer to all these questions is a conditional affirmative.

An automatic process that makes adjustments to an already existing design without making major changes is known as *refinement*. An automatic process that involves an element of design construction is usually known as *synthesis*. The term synthesis may denote anything from a mild complication of an almost acceptable design to a process that builds an acceptable design from nothing more than a list of materials and a performance specification. The term *optimisation* simply means improving performance and includes both refinement and synthesis. These are not by any means universal definitions and there is no universal agreement on the meanings of the terms.

Before we can make a coating better, we must define what we mean by *better*, and our definition must be one that can be applied to automatic methods. At the current stage of development of the subject the concept is invariably expressed in terms of changes in a single number, the *figure of merit*. The usual arrangement is for a smaller figure of merit to be better than a larger one and a figure of merit to be zero if the coating has exactly the desired performance. However, automatic processes can work as well with a figure of merit that increases as the merit improves. The figure of merit is derived from a comparison of the actual calculated performance of a design and a specification of a desired performance. The derivation involves the application of a set of rules and it is important that the rules should yield a completely unambiguous figure of merit.

Performance may include any attributes of the coating that can be quantified, but it is frequently taken as the reflectance, or transmittance, or some such normal expression of performance, at specified points over a prescribed wavelength range. Each individual expression of performance is known as a *target*. Usually the form of the rules for calculating the figure of merit will be similar to the following expression:

$$F = \frac{\sum_j [W_j |T_j - P_j|^q]}{\sum_j W_j} \quad (14.18)$$

where F is the figure of merit, T_j is the j th target, P_j is the corresponding calculated value of performance and W_j is a weight that indicates the relative importance of the particular target, or its tolerance, and may include an allowance for the scale of the particular performance attribute represented in the target. It is usual to normalise the expression so that the refinement or synthesis process has always approximately the same working range and this is indicated in equation (14.18) by dividing by the sum of the weights. The quantity q , the power to which the performance gap is raised, may be completely free for the user to choose or may, in some procedures, be completely defined. Experience shows that a value of q of 2 works well in many cases. Increasing the value of q makes the process more responsive to larger performance gaps at the expense of smaller.

The figure of merit depends on the particular set of design parameters and we can consider it as a function of the design parameters as variables. In this case we call it the *function of merit*. For efficient and reliable optimisation the function of merit should be a continuous, single-valued function of the parameters. Abrupt changes in the function of merit as parameters vary inhibit efficient refinement and should be avoided. Hard constraints on the process can have the same effect as abrupt changes and so it is often more efficient to soften the constraints by expressing their effect in terms of penalty functions attached to the function of merit rather than rigid boundaries.

If we have the same number of targets in the definition of the merit function as we have parameters in the design, then in principle, provided the targets are attainable and not mutually exclusive, the problem should be completely soluble, although it may require impossible optical constants or thicknesses. In most cases, however, we will have rather fewer parameters, or those that we have will be incapable of achieving completely the desired performance, and then the objective of the optimisation process becomes to make the figure of merit as small as possible. We can visualise the function of merit as represented by a surface in multidimensional space, one dimension for each adjustable parameter and one for the figure of merit. Making the figure of merit as small as possible, then, is translated into finding a minimum of the merit function, and thence into finding the lowest possible minimum, or, as it is known, the *global minimum*. If there are constraints on the parameters, such as permissible ranges, then the lowest possible minimum within the constraints is known as the *constrained global minimum*. Since there always are constraints (we cannot permit infinite thicknesses for instance) the minimum that concerns us will be the constrained global minimum. Unfortunately, although it is relatively easy to find a minimum of the merit function, it is not nearly as easy to find, or even to be sure that one has found, the constrained global minimum. Unless the function of merit is analytically friendly, the only way to be absolutely sure is to carry out an exhaustive search of the given parameter region. We can illustrate the problems involved in this by assuming a 20-layer design with 20 possible values of thickness for each layer, where refractive indices are already prescribed. Assume that one complete figure of merit can be generated in 1 ns. Then an exhaustive search of all possible designs will occupy a time of 20^{20} ns, that is around 2×10^9 years. This problem is considerably constrained, but already it gives some idea of what is involved in an exhaustive search. All optimisation techniques, therefore, carry out a more limited procedure that arrives at a local minimum that may be as good a minimum as is economically possible. The adjective *global* is sometimes applied to processes that essentially search in constrained parameter space for more than one merit function minimum so that they have an improved chance of finding the constrained global minimum.

We may have major gaps in our ideas of a starting design. Perhaps we do not have any idea of the indices for the layers beyond the range of possibilities that are available, or we may not know the number of layers beyond perhaps a

prescribed maximum. In that case we have the synthesis problem. If we have a reasonably good design which simply needs minor adjustment then we have refinement. Synthesis clearly has rather greater dimensions than refinement. To begin we will concentrate on refinement and assume that we have a starting design of a certain number of layers that the process will alter only in some limited way such as in terms of layer thicknesses or refractive indices, or possibly both.

In optical thin-film design we do have many techniques capable of establishing good designs that can be already almost satisfactory. In other words, they are already in the region of an acceptable minimum of the merit function and all that is required is to reach the actual minimum as quickly as possible. This is the objective of many of the optimisation techniques that are used in optical coating work. Such is the complicated nature of the function of merit that all do not necessarily find the same minimum from the same starting design. Then there are techniques designed especially so that they do not necessarily choose a neighbouring minimum. Instead they range over a region of the parameter space, in a gradually more and more constrained manner. This permits them the opportunity of discovery of any other merit function minimum that might offer improved performance over that nearest to the point of departure.

There are many ways of classifying the various optimisation techniques. They can be divided into those that use a single design that is gradually altered in prescribed ways until a minimum is reached, and those that use a family of designs, rejecting members of the family and replacing them by other designs, and reaching the minimum in this way. They may also be classified as those that attempt continuously to move towards a minimum of the merit function and those that may take some time before they finally choose the particular merit function minimum, and, therefore, have greater chance of finding a more satisfactory minimum.

Only an analytical technique can involve continuous alteration of parameters. In computer optimisation the parameters are altered in finite steps that are usually adjusted in size as the process continues. It consists, essentially, of probing the merit function surface. The results of previous probing are used to guide the choice of future ones. The optimisation is normally divided into repeated units called iterations. Each iteration will usually involve a single or multiple adjustment of the design or designs according to a set prescription and a reassessment of a new figure of merit. The process is continued until either a satisfactory outcome is attained or fresh iterations are unable to achieve any further improvement. The nature of the adjustment of the design and the way in which it is predicted is what principally distinguishes the various techniques [15].

It is tempting to find the best slope of the merit function as a function of the adjustable design parameters and simply to move down this slope as quickly as possible by changing the design parameters depending on the steepness of the slope. However, it is easy for the technique to become violently unstable with one overcorrection following another if precautions are not taken. The *steepest descent* method picks the maximum slope and follows it but the parameter

changes are usually restrained according to the derivative of the slope. If this is high, indicating that the slope appears to be changing rapidly, then the parameter changes are kept small. In the method of *damped least squares* the steepest slope down which the optimisation will travel is chosen as the slope that minimises the sum of the squares of the differences between the desired changes in the merit function parameters and the changes predicted from the local slope. The rate of travel along that direction is restrained by the introduction of a damping parameter and this avoids the slope change instabilities. Then there are several *univariate search techniques* in which only one parameter is altered at each iteration. The most common is probably the *golden section* technique. Here a minimum of the merit function is achieved for each parameter in turn. The parameters may be chosen in the order of some prescribed scheme or at random. The search for the minimum in each case involves the process of bracketing, where three values of the parameter are maintained, with the figure of merit of the central one less than either of the two outer values. This means that a minimum exists between the two outer parameters. By always dividing the appropriate region in the ratio of $1:(3 - \sqrt{5})/2$, that is $1:0.382$, the golden section, the most efficient search can be performed. Linear search techniques are like the univariate search techniques but they may freely choose the directions along which they search in parameter space. The most effective techniques change the directions from time to time based on previous progress. They are usually called *direction set methods*. The most efficient try to find a set of conjugate directions, that is a set of directions that are decoupled from each other with respect to the minimisation process—minimising along a second direction after a first should not alter the minimum of the first direction. Just one pass through the directions is then sufficient to reach the minimum. This works perfectly for simple quadratic functions. Unfortunately the thin-film functions are very complicated and they have to be searched over quite large regions so they rarely reach the final minimum in just one pass but the search can be made more efficient if a continuous attempt is made to achieve conjugate directions.

Flip-flop optimisation [4] is a relatively new term. It is a digital technique, in a sense. A design is set up consisting of a large number of very thin layers of equal geometrical or optical thickness. These thin layers may have either of only two possible indices, or admittances, usually a high value and a low value. A merit function is set up and the figure of merit calculated. Now the layers of the design, from one end to the other, are scanned. At each iteration step, the figure of merit of the coating is assessed, with the index of the appropriate layer set to both of the permitted values in turn. The better arrangement, in the sense of a lower figure of merit, is chosen, and the index of the layer set to that value. The process then passes to the adjacent layer, and so on. Several complete passes of the design may be employed, and the order in which the layers are examined may be changed. Usually the design stabilises at a minimum of the merit function after only a few passes. The designs often consist of quite long blocks of one or the other index, corresponding to normal discrete layers, separated by blocks

that clearly correspond to discrete layers of intermediate index, and occasionally a structure that represents a thicker inhomogeneous layer is obtained. The process appears very stable. It is relatively easy to take a normal discrete layer design and turn it into a suitable starting design for this process, although it appears to work quite well with all layers initially set to one or the other of the two indices.

A process that does not immediately necessarily choose the minimum towards which it shall move, is *simulated annealing* [15]. This uses a Boltzmann probability distribution:

$$\text{Prob}(E) = \exp(-E/kT) \quad (14.19)$$

where E is replaced by a merit function and kT by an annealing parameter T . Then if the existing figure of merit is E_1 and a suggested new design has E_2 , the probability that the new design is accepted in place of the old is

$$p = \text{probability} = \exp[-(E_2 - E_1)/T] \quad (14.20)$$

except that for $E_2 < E_1$ the probability is unity. The process involves calculating a new figure of merit based on a random choice of parameters within an assigned domain. If the merit function is less than the old the new design replaces the old. If the merit function is greater than the old it will be accepted with probability p based on the drawing of a random number. An *annealing schedule* is required that decides on the way in which T is allowed to fall until no further improvement is achieved.

One of the better techniques, that uses a family of designs rather than one single one, is the *simplex* technique, sometimes called *nonlinear simplex* to distinguish it from a similarly named technique in linear programming. The family of designs is known as the simplex, and numbers one more than the number of design parameters involved. At each iteration the worst design, that is the design with the greatest figure of merit, is rejected in favour of a new better design. The alternative new designs are generated in three possible ways. First the worst design is reflected in the centre of gravity of the simplex and the figure of merit calculated. If this yields a better design then a further equal move is made in the same direction and, again, the corresponding figure of merit calculated. The better of these two designs replaces the existing worst design. If the first move fails to yield a better performance then the worst design is moved halfway towards the centre of gravity, which will then normally be an improvement. In the rare cases where none of the alternatives yields a better design, a completely new simplex is generated by moving all the designs half way towards the existing best design [15].

The *statistical testing* method of Tang and Zheng [16] also involves a family of designs. Like simulated annealing it does not move immediately down a particular slope but takes rather longer and so has a better chance of finding a more acceptable minimum. A starting region of parameter space is chosen and then this region gradually shrinks around, it is hoped, a good, and perhaps even a

global, minimum. Designs are chosen at random within the starting domain until a prescribed number have been found with merit function less than a starting target. The region then shrinks until it contains only those designs, and a new target that is now the mean of the merit functions is chosen. The process is repeated until a final minimum is reached.

There is a great deal of debate about which technique is better than another and it is clear that there are differences in performance for different starting designs and coating types. A few comparative studies have been performed [17, 18] but they have not unambiguously identified any technique always superior to all others. The secret of success in refinement is a good starting design that offers scope for improvement. In that context, there is little difference between the various methods.

Synthesis is similar to refinement but involves some construction of the design beyond the adjustment of the existing layers. The number of possible designs is infinite and so the synthesis problem can be solved only by introducing some constraints. Imagine that we have a very efficient refinement technique that is capable of dealing with starting designs that are rather far from ideal. Let us now set up targets and merit function in the normal way. Next we create a starting design that uses a very small number of layers, perhaps only one. We refine this design until it is optimum. Then we add layers according to some prescribed rules. Perhaps the figure of merit will now be rather larger than before, but we refine again and eventually achieve an optimum figure of merit that is lower. Again we add layers according to our prescription and refine as before. We continue this process until we reach a stage where no improvement is taking place and at that stage we accept the best design. This is a viable synthesis technique and represents fairly well the few techniques that are sometimes used in practice. The way in which layers are added is the major difference between them. Dobrowolski [19] was the major pioneer in this field. He recognised that the addition of one single layer was often ineffective and addition of more layers was indicated. Some spectacular results have been obtained by the *needle variation* method [20]. This searches the design for the best place to add a thin slice of material. The definition of best is the maximum negative derivative of the merit function with respect to the added layer thickness. The addition of this thin slice, known as the needle, effectively adds two layers because it cuts the existing layer in two. Some commercial techniques, not otherwise published, add varying numbers of layers depending on the stage of the synthesis and on the constraints. All depend on a powerful and efficient refinement technique. The statistical refinement techniques tend to be less suitable because already they use considerable computer time and it is more usual to use either the gradient, damped least squares or linear search techniques in synthesis.

It may sometimes be said in support of a particular technique that it opens up new possibilities in design and arrives at performance levels that cannot be achieved in any other way. However, any design, however achieved, lies in the constrained parameter space. We may think of it as already existing. All that

the various techniques can do is to search the constrained parameter space to find a suitable merit function minimum. They cannot find a minimum that does not exist. Although it may seem that synthesis is an ideal technique, the difficulties in finding the constrained global, or even a very good, minimum, which are compounded by the rapid increase in complexity as layers are added, mean that the final design may not be as good as one arrived at by a process of establishing a very good starting design and then carrying out a minimum of refinement [21]. In some techniques quite thin layers that are difficult to manufacture may form part of the final design that must then be processed to remove them. The needle method, for example, introduces such thin layers as a necessary part of the process and they may remain at termination. Synthesis is therefore best used when the designer is hard pressed with little idea of how to proceed and it works most effectively when the total number of layers is not large.

Refinement and synthesis work best when the targets call for high transmittance. High reflectance presents certain problems. The performance of an optical coating is essentially a set of interference fringes. Refinement targets should be set so that they are closer together than the fringe spacing otherwise the performance in between the targets may be seriously in error. The problem is sometimes called aliasing. For sine or cosine fringe profiles avoidance of aliasing implies roughly that if the film is m quarter-waves thick then the spacing for wavelength target points should be λ/m . We often tend to work in constant increments of wavelength rather than wavenumber and so the target for a film m quarter-waves thick at λ should have $m + 1$ points to cover the octave λ to 2λ . A film that is 25 wavelengths thick then should have a target function with 100 wavelength points per octave. This modest requirement is adequate for coatings with low reflectance but, unfortunately, completely inadequate for coatings where reflectance must be high [22]. The reason is that fringe profiles are not always approximately sine or cosine functions. In an antireflection coating, the reflectance is small and multiple beam interference is weak. The fringes are then virtually sinusoidal and so the simple calculation applies. In high reflectance coatings the fringes are invariably the result of multiple-beam interference and therefore are very narrow. This increases enormously the required number of targets necessary to ensure that a fringe cannot creep in between them. Additionally, there is a definite tendency for narrow fringes of lower reflectance to appear in coatings where high reflectance is required. We can readily understand the reason. Figure 14.21 shows the reflectance curves of two similar coatings. One is a quarter-wave stack with high reflectance. The other is derived from it by increasing the thickness of one of the central quarter-waves to one half-wave. Although this converts the coating into a single-cavity narrowband filter, the width of the high-reflectance zone is considerably increased. The price is a very narrow central fringe. A density curve, figure 14.22, of the same filter, shows that there is really no fundamental gain but most merit functions are based on reflectance or transmittance, and would assign a lower figure of merit to the broader curve. Small changes in the thickness of the nominal half-wave layer can then adjust the

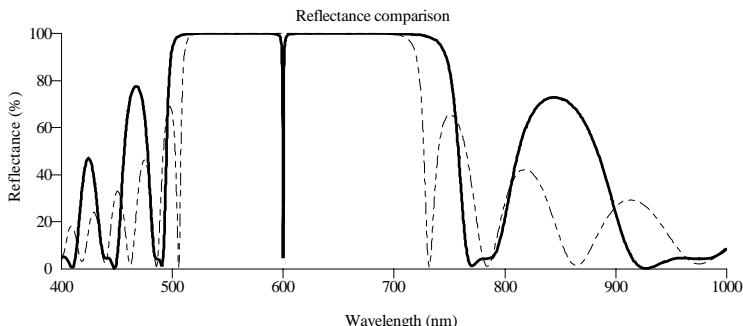


Figure 14.21. The insertion of a narrow fringe into the centre of a high-reflectance coating can actually cause an apparent increase in the width of the high-reflectance zone. The basic quarter-wave stack high reflector is the dashed line.

lateral position of the fringe with virtually no other changes. Thus the appearance of such features, sitting in between the target points in broadband reflectors, is not surprising. They are persistent and exceedingly difficult to eliminate, particularly by automatic means. Adding extra target points at the fringe is not very successful because a simple adjustment of the cavity layer thickness can move the fringe to where the target points are wider. It is therefore a very simple process for the refinement to alter slightly the thickness of one layer and move the sharp fringe exactly midway between two target points, with resulting substantial decrease of the figure of merit. This is a much easier operation for the process than the removal of a fringe, and sharp deep fringes are, therefore, persistent features that naturally position themselves between the target points, because a small change in the thickness of virtually any layer, but especially the cavity layer, will simply translate the fringe with almost no change in shape.

The fringe peaks are at their narrowest when the coating takes the form of a single cavity in the centre of the coating surrounded by maximum reflectors. Let us assume a total thickness for the coating of x full waves and arrange it as a series of quarter-waves of alternate high and low index and with a central half-wave cavity layer. The halfwidth of such an assembly is given approximately by

$$\frac{\Delta\lambda}{\lambda} = \frac{4y_L^{2x-1} y_{\text{sub}}}{\pi y_H^{2x}} \quad (14.21)$$

where we neglect any dispersion of phase shift. The spacing of the wavelength points should be perhaps half this value:

$$\frac{\Delta\lambda}{\lambda} = \left(\frac{2}{\pi}\right) \left(\frac{y_L^{2x}}{y_H^{2x}}\right) \quad (14.22)$$

where we have assumed the substrate admittance equal to y_L . We can take the

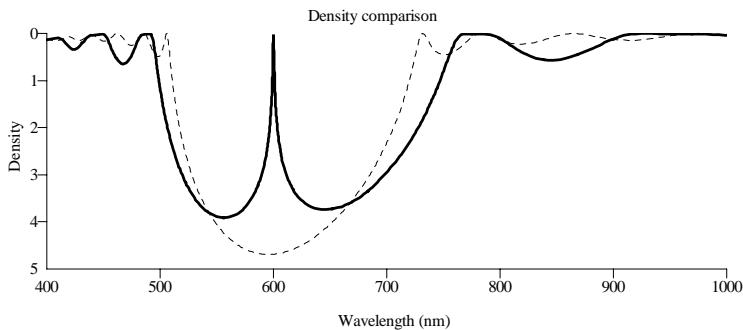


Figure 14.22. A look at the density variation shows that the performance is not better but most merit functions are based on transmittance or reflectance not density and would prefer the broader zone in figure 14.21.

wavelength interval as λ to 2λ , say, and the ratio of admittances as $\sqrt{2}$, so that the total number of points in the specification becomes:

$$N = \pi 2^{x-1} \approx 2^x. \quad (14.23)$$

Every time another full wave is added the number of points in the specification for the merit function should double.

It can be argued that the calculations are too pessimistic but it is certainly clear that there is an inexorable increase in computing requirements with coating thickness. The increased burden of calculation becomes rapidly severe if not impossible. Many of the newer processes are capable of very large numbers of layers and, especially in the case of polymeric films, coatings with thousands of layers are achievable.

Automatic methods have revolutionised the design of coatings. They have not eliminated the older techniques but have rather changed their role. The drudgery of hand calculation has been completely removed. However, as the complexity of optical coatings increases, the completely automatic methods approach a barrier to further progress in the form of suitable measures of merit and further developments in design techniques are required. The advent of the computer has certainly not reduced the need for the skill, experience and innovation that has characterised the field until now.

References

- [1] Southwell W H 1998 *Rugate Filter Structures* Private communication (Rockwell Science Center)
- [2] Epstein L I 1952 The design of optical filters *J. Opt. Soc. Am.* **42** 806–10
- [3] Epstein L I 1955 Improvements in heat reflecting filter *J. Opt. Soc. Am.* **45** 360–2

- [4] Southwell W H 1985 Coating design using very thin high- and low-index layers *Appl. Opt.* **24** 457–60
- [5] Saleh B E A and Teich M C 1991 *Fundamentals of Photonics* 1st edn (New York: Wiley)
- [6] Yariv A and Yeh P 1984 *Optical Waves in Crystals* 1st edn (New York: Wiley)
- [7] Ferencz K and Szipocs R 1993 Recent developments of laser optical coatings in Hungary *Opt. Eng.* **32** 2525–38
- [8] Szipöcs R, Ferencz K, Spielmann C and Krausz F 1994 Chirped multilayer coatings for broadband dispersion control in femtosecond lasers *Opt. Lett.* **19** 201–3
- [9] Stingl A, Spielmann C, Krausz F and Szipöcs R 1994 Generation of 11-fs pulses from a Ti:sapphire laser without the use of prisms *Opt. Lett.* **19** 204–6
- [10] Schott 1992 *Schott Optical Glass* (Duryea: Schott Glass Technologies)
- [11] Gires F and Tournois P 1964 Interféromètre utilisable pour la compression d'impulsions lumineuses modulées en fréquence *C. R. Acad. Sci.* **258** 6112–15
- [12] Kuhl J and Heppner J 1986 Compression of femtosecond optical pulses with dielectric multilayer interferometers *IEEE Trans. Quantum Electron.* **QE-22** 182–5
- [13] Szipöcs R and Krausz F 1998 *Dispersive Dielectric Mirror* USA Patent 5 734 503
- [14] Szipöcs R and Köházi-Kis A 1997 Theory and design of chirped dielectric laser mirrors *Appl. Phys. B* **65** 115–35
- [15] Press W H, Flannery B P, Teukolsky S A and Vetterling W T 1986 *Numerical Recipes. The Art of Scientific Computing* 1st edn (Cambridge: Cambridge University Press)
- [16] Tang J F and Zheng Q 1982 Automatic design of optical thin-film systems—merit function and numerical optimization method *J. Opt. Soc. Am.* **72** 1522–8
- [17] Aguilera J A, Aguilera J, Baumeister P, Bloom A, Coursen D, Dobrowolski J A, Goldstein F T, Gustafson D E and Kemp R A 1988 Antireflection coatings for germanium IR optics: a comparison of numerical design methods *Appl. Opt.* **27** 2832–40
- [18] Dobrowolski J A and Kemp R A 1990 Refinement of optical multilayer systems with different optimization procedures *Appl. Opt.* **29** 2876–93
- [19] Dobrowolski J A 1965 Completely automatic synthesis of optical thin film systems *Appl. Opt.* **4** 937–46
- [20] Furman S A and Tikhonravov A V 1992 *Basics of Optics of Multilayer Systems* 1st edn (Gif-sur-Yvette: Editions Frontières)
- [21] Thelen A 1998 Computer aided design *Optical Interference Coatings* (Washington, DC: Optical Society of America) pp 268–70
- [22] Macleod H A 1996 Recent trends in optical thin films *Rev. Laser Eng.* **24** 3–10

Chapter 15

Characteristics of thin-film dielectric materials

This list gives some details of the more common thin-film dielectric materials. It is not a definitive list but is intended to show the wide range of available materials. The metals exhibit enormous dispersion and so an abbreviated table of values is of little use. For extended tables of the optical constants of metals consult [1–4]. Surveys of many thin-film materials are given by Ritter [5, 6] and by Palik [2–4]. For a fuller account of the fluorides of the rare earths consult Lingg [7].

In most cases the materials in the table can be deposited by many different processes. Where thermal evaporation is possible it is the main process listed. Many of the materials, with the principal exception of the fluorides, can be sputtered in their dielectric form by either radio frequency sputtering or neutral ion-beam sputtering. A few materials, the nitrides especially, are not capable of evaporation or reactive evaporation and require an energetic process such as ion-assisted deposition.

The optical properties of thin films are very dependent on deposition conditions and other factors. The values quoted should be interpreted simply as values that were reported at some time, and not as necessarily intrinsic and repeatable properties of the materials.

Table 15.1.

Materials	Deposition technique	Refractive index	Region of transparency	Remarks	References
Aluminum oxide (Al_2O_3)	E-beam	1.62 at 0.6 μm 1.59 at 1.6 μm 1.62 at 0.6 μm 1.59 at 1.6 μm	$T_s = 300^\circ\text{C}$ $T_s = 40^\circ\text{C}$	Can also be produced by anodic oxidation of Al in ammonium tartrate solution [8]	[9]
Aluminum oxynitride (AlO_xN_y)	E-beam evaporation of Al with nitrogen ion assist and oxygen background	1.71–1.93 at 350 nm 1.65–1.83 at 550 nm	<300 nm–6.5 μm	Index varies continuously as function of composition	[10, 11]
Antimony trioxide (Sb_2O_3)	Molybdenum boat	2.20 at 366 nm 2.04 at 546 nm	300 nm–1 μm	Important to avoid overheating otherwise decomposes	[12]
Antimony sulphide (Sb_2S_3)		3.0 at 589 nm	500 nm–10 μm	Brief note [13, p 189]	[14, 15]
Beryllium oxide (BeO)	Tantalum boat. Reactive evaporation of Be metal in activated oxygen	1.82 at 193 nm 1.72 at 550 nm	190 nm infrared (IR)	Highly toxic	[16]
Bismuth oxide (Bi_2O_3)	E-beam [17]. Also reactive sputtering of bismuth in oxygen [18]	2.7 at 600 nm 2.2 at 9 μm 2.45 at 550 nm	<550 nm–12 μm (E-beam) (Sputter)	Good infrared material but less abrasion resistant than other oxides [17]	[17, 18]
Bismuth trifluoride (BiF_3)	Graphite Knudsen cell	1.74 at 1 μm 1.65 at 10 μm	260 nm–20 μm		[19]
Cadmium sulphide (CdS)	Quartz crucible with spiral filament in contact with charge	2.6 at 600 nm 2.27 at 7 μm	600 nm–7 μm	Avoid overheating. Filament temperature must be $\leq 1025^\circ\text{C}$	[14, 20]
Cadmium telluride (CdTe)	Molybdenum boat	3.05 in near IR			[21] (brief)

Table 15.1. (Continued)

Materials	Deposition technique	Refractive index	Region of transparency	Remarks	References
Calcium fluoride (CaF ₂)	Molybdenum or tantalum boat. E-beam [17]	1.23–1.26 at 546 nm 1.40 at 600 nm 1.32 at 9 μm	(porous) (E-beam)	150 nm–12 μm	[14, 17, 21, 22]
Ceric oxide (CeO ₂)	Tungsten boat	2.2 at 550 nm 2.18 at 500 nm 2.42 at 250 nm 2.2 in near IR	$T_s = 50^\circ\text{C}$ $T_s = 350^\circ\text{C}$	400 nm–16 μm	[23–26]
Cerous fluoride (CeF ₃)	Tungsten boat. E-beam	1.63 at 550 nm 1.59 at 2 μm 1.57 at 9 μm	(E-beam)	300 nm–12 μm	Hot substrate. Craze on cold substrate [23]. High tensile stress
Chiolite (5NaF3AlF ₃)	Howitzer or tantalum boat			Similar to cryolite	[21]
Chromium oxide (Cr ₂ O ₃)	E-beam	2.242 at 700 nm 2.1 at 8 μm		<600 nm–8 μm	[17]
Cryolite (Na ₃ AlF ₆)	Howitzer or tantalum boat	1.35 at 550 nm		<200 nm–14 μm	Slightly hygroscopic Soft, easily damaged
Gadolinium fluoride (GdF ₃)	E-beam	1.55 at 400 nm		140 nm–>12 μm	[7]
Germanium (Ge)	E-beam or graphite boat	4.25 in IR (usually slightly higher than bulk value)		1.7–100 μm	Absorption band centred at approx. 25 μm
Hafnium dioxide (HfO ₂)	E-beam	2.088 at 350 nm 2.00 at 500 nm 1.88 at 8 μm		220 nm–12 μm	[17, 24, 30, 31]
Hafnium fluoride (HfF ₄)	E-beam	1.57 at 600 nm 1.46 at 10 μm		<600 nm–12 μm	[17]

Table 15.1. (Continued)

Materials	Deposition technique	Refractive index	Region of transparency	Remarks	References
Lanthanum fluoride (LaF ₃)	Tungsten boat. E-beam [17]	1.59 at 500 nm 1.57 at 2 μ m 1.52 at 9 μ m	200 nm–12 μ m (E-beam)	Heated substrate	[17, 23, 24, 27, 32, 33]
Lanthanum oxide (La ₂ O ₃)	Tungsten boat	1.95 at 550 nm 1.86 at 2 μ m	350 nm–>2 μ m	Hot substrate (~300 °C)	[23, 24, 27]
Lead chloride (PbCl ₂)	Platinum or molybdenum boat	2.3 at 550 nm 2.0 at 10 μ m	300 nm–>14 μ m		[21, 34]
Lead fluoride (PbF ₂)	Platinum boat. E-beam [17]	1.75 at 500 nm 1.70 at 1 μ m 1.3 at 10 μ m	240 nm–>20 μ m		[17, 21, 23, 35, 36]
Lead telluride (PbTe)	Tantalum boat	5.5 in IR	3.4 μ m–>30 μ m	Avoid overheating. Hot substrate (see text)	[37–39]
Lithium fluoride (LiF)	Tantalum boat	1.36–1.37 at 546 nm	110 nm–7 μ m		[14, 40]
Lutetium fluoride (LuF ₃)	E-beam	1.51 at 400 nm	140 nm–12 μ m		[7]
Magnesium fluoride (MgF ₂)	Tantalum boat	1.38 at 550 nm 1.35 at 2 μ m	210 nm–10 μ m	Films on heated substrates much more rugged. High tensile stress	[14, 21, 22, 24, 30, 41–43]
Magnesium oxide (MgO)	E-beam	1.7 at 550 nm 1.74 at 550 nm	$T_s = 50^\circ\text{C}$ $T_s = 300^\circ\text{C}$	210 nm–8 μ m	[44]
Neodymium fluoride (NdF ₃)	Tungsten boat. E-beam [17]	1.60 at 250 nm 1.58 at 2 μ m 1.60 at 9 μ m	220 nm–12 μ m (E-beam)	Hot substrate 300 °C	[17, 23, 24, 27]

Table 15.1. (Continued)

Materials	Deposition technique	Refractive index	Region of transparency	Remarks	References
Neodymium oxide (Nd ₂ O ₃)	Tungsten boat	2.0 at 550 nm 1.95 at 2 μ m	400→2 μ m	Hot substrate 300 °C. Decomposes at high boat temperature	[23, 27]
Praseodymium oxide (Pr ₆ O ₁₁)	Tungsten boat	1.92 at 500 nm 1.83 at 2 μ m	400→2 μ m	Hot substrate 300 °C	[27]
Samarium fluoride (SmF ₃)	E-beam	1.56 at 400 nm	160 nm→12 μ m		[7]
Scandium oxide (Sc ₂ O ₃)	E-beam	1.86 at 550 nm	350 nm–13 μ m		[45]
Silicon (Si)	E-beam with water-cooled hearth. Sputtering	3.5 in IR	1.1–14 μ m		[23]
Silicon monoxide (SiO)	Tantalum boat or howitzer	2.0 at 550 nm 1.7 at 6 μ m	500 nm–8 μ m	Fast evaporation at low pressure	[21] (brief) [9, 14, 23, 30, 46]
Disilicon trioxide (Si ₂ O ₃)	Tantalum boat or howitzer	1.52–1.55 at 550 nm	300 nm–8 μ m		[9, 23, 47–52]
Silicon dioxide (SiO ₂)	E-beam. Mixture in tungsten boat	1.46 at 500 nm 1.445 at 1.6 μ m	<200 nm–8 μ m (in thin films)		[9, 23, 53, 54]
Silicon nitride (Si ₃ N ₄)	Low voltage reactive ion plating	2.06 at 500 nm	320 nm–7 μ m		[55]
Sodium fluoride (NaF)	Tantalum boat	1.34 in visible	<250 nm–14 μ m		[14] (in brief)
Strontrium fluoride (SrF ₂)	E-beam	1.46 at 600 nm 1.3 at 10 μ m	<600 nm→12 μ m		[17]

Table 15.1. (Continued)

Materials	Deposition technique	Refractive index	Region of transparency	Remarks	References
Tantalum pentoxide (Ta ₂ O ₅)	E-beam	2.16 at 550 nm 1.95 at 8 μm	300 nm–10 μm		[17,24]
Tellurium (Te)	Tantalum boat	4.9 at 6 μm	3.4 μm–20 μm		[21,23,56,57]
Titanium dioxide (TiO ₂)	Reactive evaporation of TiO, Ti ₂ O ₃ or Ti ₃ O ₅ in O ₂ , E-beam reactive evaporation	2.2–2.7 at 550 nm depending on structure	350 nm–12 μm	Can also be produced by subsequent oxidation of Ti film	[14,23,47,52,53,58–63]
Thallous chloride (TlCl)	Tantalum boat	2.6 at 12 μm	Visible→20 μm		[21,64]
Thorium oxide (ThO ₂)	E-beam	1.8 at 550 nm 1.75 at 2 μm	250 nm–15 μm	Radioactive	[21,23,65–67]
Thorium fluoride (ThF ₄)	Tantalum boat	1.52 at 400 nm 1.51 at 750 nm	250 nm→15 μm	Radioactive. Note: Thorium oxyfluoride (ThOF ₂) actually forms ThF ₄ when evaporated	[21,23,65–68]
Ytterbium fluoride (YbF ₃)	E-beam	1.52 at 600 nm 1.48 at 10 μm	<600 nm–12 μm		[17]
Yttrium oxide (Y ₂ O ₃)	E-beam	1.82 at 550 nm 1.69 at 9 μm	250 nm–12 μm		[17,24,30,69]
Zinc selenide (ZnSe)	Platinum or tantalum boat	2.58 at 633 nm	600 nm→15 μm		[66]
Zinc sulphide (ZnS)	Tantalum boat or howitzer	2.35 at 550 nm 2.2 at 2.0 μm	380 nm→25 μm		[14,21,23,26,29,39,41,67]

Table 15.1. (Continued)

Materials	Deposition technique	Refractive index	Region of transparency	Remarks	References
Zirconium dioxide (ZrO_2)	E-beam	2.1 at 550 nm 2.05 at 9.0 μm	340 nm–12 μm		[17, 24, 46]
Substance H1 [†] (zirconia/titania)	Tungsten boat or E-beam	2.1 at 550 nm	360 nm–7 μm	Does not melt completely [70]	[70, 71]
Substance H2 [†] (mixed praseodymium and titanium oxides)	E-beam	2.1 at 550 nm	400–7 μm	Some weak absorption bands visible [70]	[70]
Substance H4 [†] (lanthanum and titanium oxide)	E-beam with molybdenum liner	2.1 at 500 nm	$T_s = 300^\circ\text{C}$	360 nm–7 μm	[70]
Substance M1 [†] (mixed praseodymium and aluminium oxides)	E-beam	1.71 at 500 nm	$T_s = 300^\circ\text{C}$	300 nm–9 μm	[70]

[†] Substance H1, Substance H2, Substance H4 and Substance M1 are members of the Patinal® series of optical coating materials manufactured by E Merck, Darmstadt, Germany.

References

- [1] Hass G and Hadley L 1972 Optical constants of metals *American Institute of Physics Handbook* ed D E Gray (New York: McGraw Hill) pp 6.124–56
- [2] Palik E D ed 1985 *Handbook of Optical Constants of Solids* (San Diego: Academic)
- [3] Palik E D 1991 *Handbook of Optical Constants of Solids II* (San Diego: Academic)
- [4] Palik E D 1998 *Handbook of Optical Constants of Solids III* (San Diego: Academic)
- [5] Ritter E 1975 Dielectric film materials for optical applications *Physics of Thin films* ed G Hass, M H Francombe and R W Hoffman (New York: Academic) pp 1–49
- [6] Ritter E 1976 Optical film materials and their applications *Appl. Opt.* **15** 2318–27
- [7] Lingg L J 1990 Lanthanide trifluoride thin films: structure, composition and optical properties *PhD Dissertation* (University of Arizona)
- [8] Hass G 1949 On the preparation of hard oxide films with precisely controlled thickness on evaporated aluminum mirrors *J. Opt. Soc. Am.* **39** 532–40
- [9] Cox J T, Hass G and Ramsay J B 1964 Improved dielectric films for multilayer coatings and mirror protection *J. Phys.* **25** 250–4
- [10] Hwangbo C K, Lingg L J, Lehan J P, Macleod H A and Suits F 1989 Reactive ion-assisted deposition of aluminum oxynitride thin films *Appl. Opt.* **28** 2779–84
- [11] Targove J D, Lingg L J, Lehan J P, Hwangbo C K, Macleod H A, Leavitt J A and McIntyre L C Jr 1987 Preparation of aluminum nitride and oxynitride thin films by ion-assisted deposition *Materials Modification and Growth using Ion Beams Symposium (Anaheim, CA)* (Pittsburgh, PA: Materials Research Society) pp 311–16
- [12] Jenkins F A 1958 Extension du domaine spectral de pouvoir réflecteur élevé des couches multiples diélectriques *J. Phys. Rad.* **19** 301–6
- [13] Heavens O S, Ring J and Smith S D 1957 Interference filters for the infra-red *Spectrochim. Acta* **10** 179–94
- [14] Heavens O S 1960 Optical properties of thin films *Rep. Prog. Phys.* **23** 1–65
- [15] Billings S H and Billings M H Jr 1947 The infra-red refractive index and dispersion of evaporated stibnite thin films *J. Opt. Soc. Am.* **37** 119–21
- [16] Ebert J 1982 Activated reactive evaporation *Proc. Soc. Photo-Opt. Instrumentation Eng.* **325** 29–38
- [17] Kruschwitz J D T and Pawlewicz W T 1997 Optical and durability properties of infrared transmitting thin films *Appl. Opt.* **36** 2157–9
- [18] Holland L and Siddall G 1958 Heat-reflecting windows using gold and bismuth oxide films *Br. J. Appl. Phys.* **9** 359–61
- [19] Moravec T J, Skogman R A and Bernal G E 1979 Optical properties of bismuth trifluoride thin films *Appl. Opt.* **18** 105–10
- [20] Hall J F and Ferguson W F C 1955 Optical properties of cadmium sulphide and zinc sulphide from 0.6 micron to 14 micron *J. Opt. Soc. Am.* **45** 714–18
- [21] Ennos A E 1966 Stresses developed in optical film coatings *Appl. Opt.* **5** 51–61
- [22] Heavens O S and Smith S D 1957 Dielectric thin films *J. Opt. Soc. Am.* **47** 469–72
- [23] Ritter E 1961 Gesichtspunkte bei der Stoffauswahl für dünne Schichten in der Optik *Z. Angew. Math. Phys.* **12** 275–6
- [24] Smith D and Baumeister P W 1979 Refractive index of some oxide and fluoride coating materials *Appl. Opt.* **18** 111–15
- [25] Hass G, Ramsay J B and Thun R 1958 Optical properties and structure of cerium dioxide films *J. Opt. Soc. Am.* **48** 324–7

- [26] Cox J T and Hass G 1958 Antireflection coatings for germanium and silicon in the infrared *J. Opt. Soc. Am.* **48** 677–80
- [27] Hass G, Ramsay J B and Thun R 1959 Optical properties of various evaporated rare earth oxides and fluorides *J. Opt. Soc. Am.* **49** 116–20
- [28] Pelletier E, Roche P and Vidal B 1976 Détermination automatique des constantes optiques et de l'épaisseur de couches minces: application aux couches diélectriques *Nouv. Rev. Opt.* **7** 353–62
- [29] Netterfield R P 1976 Refractive indices of zinc sulphide and cryolite in multilayer stacks *Appl. Opt.* **15** 1969–73
- [30] Borgogno J P, Lazarides B and Pelletier E 1982 Automatic determination of the optical constants of inhomogeneous thin films *Appl. Opt.* **21** 4020–9
- [31] Baumeister P W and Arnon O 1977 Use of hafnium dioxide in multilayer dielectric reflectors for the near uv *Appl. Opt.* **16** 439–44
- [32] Bourg A, Barbaroux N and Bourg M 1965 Propriétés optiques et structure de couches minces de fluorure de lanthane *Opt. Acta* 151–60
- [33] Targove J D, Lehan J P, Lingg L J, Macleod H A, Leavitt J A and McIntyre L C 1987 Ion-assisted deposition of lanthanum fluoride thin films *Appl. Opt.* **26** 3733–7
- [34] Penselin S and Steudel A 1955 Fabry–Perot-Interferometerverspiegelungen aus dielektrischen Vielfachschichten *Z. Phys.* **142** 21–41
- [35] Carl-Zeiss-Stiftung 1965 *Interference filters* UK Patent 994 638
- [36] Lès Z, Lès F and Gabla L 1963 Semitransparent metallic–dielectric mirrors with low absorption coefficient in the ultra-violet region of the spectrum (3200–2400 Å) *Acta Phys. Pol.* **23** 211–14
- [37] Smith S D and Seeley J S 1968 *Multilayer Filters for the Region 0.8 to 100 Microns* (Air Force Cambridge Research Laboratories)
- [38] Yen Y-H, Zhu L-X, Zhang W-D, Zhang F-S and Wang S-Y 1984 Study of PbTe optical coatings *Appl. Opt.* **23** 3597–601
- [39] Ritchie F S 1970 Multilayer filters for the infrared region 10–100 microns *PhD Thesis* (University of Reading)
- [40] Schulz L G 1949 The structure and growth of evaporation LiF and NaCl films on amorphous substrates *J. Chem. Phys.* **17** 1153–62
- [41] Hall J F Jr and Ferguson W F C 1955 Dispersion of zinc sulfide and magnesium fluoride films in the visible spectrum *J. Opt. Soc. Am.* **45** 74–5
- [42] Wood O R II, Craighead H G, Sweeney J E and Maloney P J 1984 Vacuum ultraviolet loss in magnesium fluoride films *Appl. Opt.* **23** 3644–9
- [43] Hall J F 1957 Optical properties of magnesium fluoride films in the ultraviolet *J. Opt. Soc. Am.* **47** 662–5
- [44] Pulker H K 1979 Characterization of optical thin films *Appl. Opt.* **18** 1969–77
- [45] Arndt D P, Azzam R M A, Bennett J M, Borgogno J P, Carniglia C K, Case W E, Dobrowolski J A, Arndt D P, Gibson U J, Hart T T *et al* 1984 Multiple determination of the optical constants of thin-film coating materials *Appl. Opt.* **23** 3571–96
- [46] Hass G and Salzberg C D 1954 Optical properties of silicon monoxide in the wavelength region from 0.24 to 14.0 microns *J. Opt. Soc. Am.* **44** 181–7
- [47] 1957 *Improvements in or Relating to the Manufacture of Thin Light Transmitting Layers* UK Patent 775 002
- [48] Ritter E 1962 Zur Kenntnis des SiO und Si₂O₃—Phase in dünnen Schichten *Opt. Acta* **9** 197–202

- [49] Okamoto E and Hishinuma Y 1965 Properties of evaporated thin films of Si_2O_3 *Trans. 3rd Int. Vac. Congress* **2** 49–56
- [50] Bradford A P, Hass G, McFarland M and Ritter E 1965 Effect of ultraviolet irradiation on the optical properties of silicon oxide films *Appl. Opt.* **4** 971–6
- [51] Bradford A P and Hass G 1963 Increasing the far-ultra-violet reflectance of silicon oxide protected aluminium mirrors by ultraviolet irradiation *J. Opt. Soc. Am.* **53** 1096–100
- [52] Auwärter M 1960 *Process for the Manufacture of Thin Film* USA Patent 2 920 002
- [53] Reichelt W 1965 Fortschritte in der Herstellung von Oxydschichten für optische und elektrische Zwecke *Trans. 3rd Int. Vac. Congress* **2** 25–9
- [54] Libbey-Owens-Ford Glass Company 1947 *Method of Coating with Quartz by Thermal Evaporation* UK Patent 632 442
- [55] Bovard B B, Ramm J, Hora R and Hanselmann F 1989 Silicon nitride thin films by low voltage reactive ion plating: optical properties and composition *Appl. Opt.* **28** 4436–41
- [56] Moss T S 1952 Optical properties of tellurium in the infra-red *Proc. Phys. Soc.* **65** 62–6
- [57] Greenler R G 1955 Interferometry in the infrared *J. Opt. Soc. Am.* **45** 788–91
- [58] Hass G 1952 Preparation, properties and optical applications of thin films of titanium dioxide *Vacuum* **2** 331–45
- [59] Brinsmaid D S, Keenan W J, Koch G J and Parsons W F Eastman Kodak Co 1957 *Method of Producing Titanium Dioxide Coatings* USA Patent 2 784 115
- [60] Balzers Patent und Lizenz Anstalt 1962 *Improvements in and Relating to the Oxidation and/or Transparency of Thin Partly Oxidic Layers* UK Patent 895 879
- [61] Pulker H K, Paesold G and Ritter E 1976 Refractive indices of TiO_2 films produced by reactive evaporation of various titanium-oxide phases *Appl. Opt.* **15** 2986–91
- [62] Heitmann W 1971 Reactive evaporation in ionized gases *Appl. Opt.* **10** 2414–18
- [63] Chiao S-C, Bovard B G and Macleod H A 1998 Repeatability of the composition of titanium oxide films produced by evaporation of Ti_2O_3 *Appl. Opt.* **37** 5284–90
- [64] Perkin-Elmer Corporation 1961 *Infrared Filters* UK Patent 970 071
- [65] Heitmann W and Ritter E 1968 Production and properties of vacuum evaporated films of thorium fluoride *Appl. Opt.* **7** 307–9
- [66] Heitmann W 1966 Extrem hochreflektierende dielektrische Spiegelschichten mit Zinkselenid *Z. Angew. Phys.* **21** 503–8
- [67] Behrndt K H and Doughty D W 1966 Fabrication of multilayer dielectric films *J. Vacuum Sci. Technol.* **3** 264–72
- [68] Ledger A M and Bastien R C 1977 *Intrinsic and Thermal Stress Modeling for Thin-Film Multilayers* (Norwalk, CT: The Perkin Elmer Corporation)
- [69] Lubezky I, Ceren E and Klein Z 1980 Silver mirrors protected with Yttria for the 0.5 to 14 μm region *Appl. Opt.* **19** 1895
- [70] Fritz M, Koenig F, Merck E and Feiman S 1992 New materials for production of optical coatings *35th Annual Technical Conf. Proc.* (Albuquerque, NM: Society of Vacuum Coaters) pp 143–7
- [71] Stetter F, Esselborn R, Harder N, Friz M and Tolles P 1976 New materials for optical thin films *Appl. Opt.* **15** 2315–17

Index

- abrasion resistance, 440–441
absorbers, spectrally selective, 579–583
absorbing media,
 antireflection of, 34–35
 normal incidence, 29
 oblique incidence, 36–39
absorptance, 43–45
absorption, 204–208, 477
absorption coefficient, 18
absorption filters,
 shortwave pass, 246
 thin-film, 210–211
adhesion, 442–444
 aluminium, 443
 direct pull measurement, 442
 scratch test, 442–443
 zinc sulphide, 442
admittance diagram,
 electric field, 60–66
 electric field losses, 62–66
 electric field theory, 60–66
 theory, 55–66
admittances, modified, 349–353, 350
advanced plasma source, 411, 414
all-dielectric Fabry–Perot filter, *see* Fabry–Perot
aluminium, 158, 167, 264–265, 265
aluminium nitride, 453–454
aluminium oxide (Al_2O_3), 163–164, 622
aluminium oxynitride (AlO_xN_y), 453–454, 622
aluminium source, 398, 402
aluminium, reflectance, 159
amplitude reflection coefficient, 22
amplitude transmission coefficient, 22
angle of incidence, effect of, 283–292
antimony sulphide (Sb_2S_3), 622
antimony trioxide (Sb_2O_3), 193, 318, 622
antireflection coatings, 86–159
 antireflection, single layer, 110
 antireflection, single layer, 87–92
 buffer layer, 148–152
 double layer, 111–118
 double layer, 92–101
 double layer, admittance diagram, 118
 double layer, admittance diagram, 119
 double layer, admittance diagram, 95
 double layer, vector diagram, 94
 double layer, vector diagram, 94
Epstein, 137
equivalent admittance, 137
for visible and infrared, 144
four-layer, 128
Frank Rock, 132
glass, 111–156
high-index substrates, 87–108
inhomogeneous layers, 152–155
low-index substrates, 108–156
Mouchart, 143

- multilayer, 102–108
- multilayer, 118–156
- multilayer, vector diagram, 103
- Musset and Thelen technique, 104–108
- quarter-half-quarter coating, 129
- Reichert, 134–136
- Thetford's technique, 118–126
- two zeros, 139–144
- V-coat, 113
- Vermeulen technique, 132
- Vermeulen technique, 137
- W-coat, 120
- W-coat, 127
- Young's technique, 108
- apparent curvature of reflector, 200–203
- applications of coatings, 536–585
- arsenic triselenide, 100
- arsenic trisulphide, 100
- astronomical applications of filters, 545–550
- atmospheric temperature sounding, 550–559
- automatic methods of design, 610–619
- baking, 417
- baking and adhesion, 418
- band-pass filters, 257–345
- barium fluoride substrate, 200
- beam splitters,
 - considerations, 538–540
 - dielectric, 172–176
 - oxide (BeO), 623
 - polarisation, 538–540
- bismuth oxide (Bi_2O_3), 114, 622
- bismuth trifluoride (BiF_3), 622
- blocking of sideband, 293
- boosted reflectors, 164–167
- Boyle, Robert, 1
- Brewster angle, 28–29, 350
 - polarising beam splitter, 362–366
- broad band-pass filters, 257–260
- buffer layer, 148–152
- cadmium sulphide (CdS), 623
- cadmium telluride (CdTe), 623
- caesium iodide, 274
- calcium fluoride (CaF_2), 193, 624
- ceric oxide (CeO_2), 193, 405, 448, 623
- ceric oxide (CeO_2), 89, 127, 193, 623
- cerous fluoride (CeF_3), 96
- characteristic matrix, 39
- characteristic optical admittance, 16
- characteristic shifts due to temperature, 474–477
- chemical vapour deposition, 413–415
- chiolite ($5\text{NaF}\cdot3\text{AlF}_3$), 199, 623
- chirped mirrors, 609
- chirped pulse, 603, 604
- Chromel A, 176–177
- chromium, 159, 170–172
- chromium oxide (Cr_2O_3), 623
- Ciddor, 200–203
- circle diagrams, 80–85
- coating edge, 536–537
- coatings with metal layers, 575–585
- columnar growth, 463, 464
- complex refractive index, 14
- computer refinement, 195, 233, 613–616
- contamination, sensitivity to, 478–485
- copper, reflectance, 159
- critical angle, definition, 350
- cryolite (Na_3AlF_6), 192–193, 193, 196, 197, 203, 264, 274, 279, 318, 364, 405, 447, 623
- cryolite, temperature coefficient of optical thickness, 344
- cube polarisers, 367
- DC planar magnetron sputtering, 405–408

- defects in microstructure, 467–468
 delta, definition, 40
 deposition parameters, influence on
 film properties, 462–463
 DHW filter, 257, 300, 393–300
 didymium fluoride, 96
 dielectric materials beyond critical
 angle, 357–359
 direct monitoring, 515
 direct turning value monitoring,
 515–517
 layer sensitivity, 518
 disilicon trioxide (Si_2O_3), 625
distribution, see also uniformity
 boats, 495
 electron beam source, 495
 howitzer, 495
- E, equivalent optical admittance,
 216–220
 edge filter, 210–255
 design,
 edge steepness, 255
 extending rejection zone, 246–
 248
 extending transmission zone,
 248–253
 practical filters, 244–246
 reducing transmission zone, 253–
 254
 Seeley lumped circuits, 240–244
 Thelen shifted periods, 238–240
 with inhomogeneous matching
 layer, 155
 Young and Crystal, 234–238
 effect of temperature, 474–477
 effective index, in tilting, 284
 electrode films for Schottky-barrier
 photodiodes, 575–578
 electron beam source, 399–403
 distribution, 495
 energetic processes, 405–413
 energy grasp, 540–545
 environmental effects, 530–534
- Epstein, 259
 equivalent optical admittance, 216–
 220
 equivalent phase thickness, 216–220
 error compensation in direct turning
 value monitoring, 516–518
 evaporation, reactive, 448–449
 extended high reflectance zones,
 193–200
 extinction coefficient, 14–15
- Fabry–Perot filter,
 absorption, 275–280
 absorption all-dielectric, 275–280
 absorption metal–dielectric, 265–
 266
 all-dielectric, 266–280
 bandwidth, 268–274
 fused silica spacer, 281
 germanium solid etalon, 282
 germanium spacer, 274
 mica spacer, 280–281
 Mylar spacer, 282
 resolving power, 262–263
 sensitivity to errors, 265–266
 solid etalon, 280–283
 solid etalon, infrared, 282–283
 solid etalon, requirements, 281–
 282
 structure, 267
 typical, 274
 uniformity, 279
 Yttralox spacer, 282
- Fabry–Perot interferometer, 179–
 185
- film performance, influence of mi-
 crostructure, 462–478
- film properties, influence of deposi-
 tion parameters, 462–463
- filters,
 astronomical applications, 545–
 550
 effect of intense illumination,
 344–345

- effect of temperature, 344–345
finesse, 181
flattening characteristic using
halfwave layer, 120, 128,
132
Fraunhofer, Joseph, 2–4
Fresnel rhomb, 384
Fresnel, Augustin Jean, 2
fringe order, m , 181
frustrated total reflectance, *see* FTR
FTR filter, 390
FTR, frustrated total reflectance,
361, 390
- g , definition, 91–92
gadolinium fluoride (GdF_3), 623
gallium arsenide substrate, 89
gamma, equivalent phase thickness,
216–220
GD, 606
GDD, 606
Geffcken, W, 4
germanium (Ge), 96, 193, 274, 451,
623
absorption filter, 210–211
source, 399
substrate, 100, 104, 106, 155,
231, 89, 90, 91, 96
glare suppression filters and coat-
ings, 570–575
gold, 185
reflectance, 159
Greenland and Billington, 362
group delay, 606
group delay dispersion, 606
group velocity, 599, 602
group velocity dispersion, 603, 604
GVD, 603
- hafnium dioxide (HfO_2), 623
hafnium fluoride (HfF_4), 623
half-wave layer, flattening, 120,
128, 132
half-wave retardation, Lostis, 384
- half-wave thicknesses, theory, 52–
53
hard coat on plastic substrates, 415
heat reflector, triple stack, 254
heavy absorption in optical property
measurement, 423
Herpin index, 72–73, 213–232
application to nonquarterwave
stacks, 216–220
hexamethyldisiloxane, 415
high reflectance coatings, 179–208
high reflectance zones, extended,
193–200
high-reflectance zone width, 188–
192
history of optical thin films, 1–4
HMDSO, 415
Hooke, Robert, 1
howitzer source, 399, 403
distribution, 495
- incident cone of light, effect on
filter, 288–292
indirect monitoring, 515
indium antimonide substrate, 89
induced transmission filter, 327–342
bandwidth, 340
design examples, 331–340
uv, measured performance, 343
manufacture, 340–342
matching stack, 345–347
inhomogeneous layers, 152–155,
589–590
intense illumination, effect on fil-
ters, 344–345
introduction, 1
ion-assisted deposition, 410–412
ion-beam sputtering, 408
ionised plasma-assisted deposition,
411–412, 414
irradiance, 17
- Kretschmann and Raether coupling,
361

- lanthanum fluoride (LaF_3), 624
lanthanum oxide (La_2O_3), 624
laser damage, 477–478
lead chloride (PbCl_2), 193, 624
lead fluoride (PbF_2), 318, 624
lead telluride (PbTe), 193, 274, 452–453, 624
absorption filter, 211
temperature coefficient, 345
lithium fluoride (LiF), 624
longwave pass filter,
design, 232
practical performance, 246
losses, 477
losses in reflectors, 204–208
low-voltage ion plating, 408–410
lutetium fluoride (LuF_3), 624
- MacNeille polarising beam splitter, 351, 362–366
magnesium fluoride (MgF_2), 104, 110, 112, 114, 127, 164, 167, 193, 193, 264, 265, 319, 405, 446–447, 624, 96, 97
optical property measurement, 421, 422
magnesium oxide (MgO), 624
magnetron sputtering, 405–408
manufacturing specification, 526
Mary Banning, 362
material properties, summary, 446–456
materials
aluminium, 158, 167, 264–265, 265
aluminium nitride, 453–454
aluminium oxide (Al_2O_3), 164, 622
aluminium oxynitride (AlO_xN_y), 453–454, 622
aluminium source, 398, 402
aluminium, reflectance, 159
antimony sulphide (Sb_2S_3), 622
antimony trioxide (Sb_2O_3), 193, 318, 622
arsenic triselenide, 100
arsenic trisulphide, 100
beryllium oxide (BeO), 622
bismuth oxide (Bi_2O_3), 114, 622
bismuth trifluoride (BiF_3), 622
cadmium sulphide (CdS), 622
cadmium telluride (CdTe), 622
caesium iodide, 274
calcium fluoride (CaF_2), 193, 623
ceric oxide (CeO_2), 193, 405, 448, 623, 89
cerous fluoride (CeF_3), 96, 127, 193, 623
chiolite ($5\text{NaF}\cdot3\text{AlF}_3$), 199, 623
Chromel A, 176–177
chromium, 159, 177–172
chromium oxide (Cr_2O_3), 623
cryolite (Na_3AlF_6), 192–193, 193, 196, 197, 203, 264, 274, 279, 318, 364, 405, 447, 623
cryolite, temperature coefficient of optical thickness, 344
didymium fluoride, 96
disilicon trioxide (Si_2O_3), 625
gadolinium fluoride (GdF_3), 623
germanium (Ge), 97, 193, 274, 451, 623
gold, 185
gold, reflectance, 159
hafnium dioxide (HfO_2), 451, 623
hafnium fluoride (HfF_4), 623
lanthanum fluoride (LaF_3), 318, 624
lanthanum oxide (La_2O_3), 624
lead chloride (PbCl_2), 193, 624
lead fluoride (PbF_2), 318, 624
lead telluride (PbTe), 193, 274, 452–453, 624
lithium fluoride (LiF), 624
lutetium fluoride (LuF_3), 624

- magnesium fluoride (MgF_2), 96, 97, 104, 110, 112, 114, 127, 164, 167, 193, 264, 265, 319, 405, 446–447, 624
magnesium fluoride, optical property measurement, 421, 422
magnesium oxide (MgO), 624
material mixtures, *see* mixtures
neodymium fluoride (NdF_3), 624
neodymium oxide (Nd_2O_3), 625
Nichrome, 159, 176–177
samarium fluoride (SmF_3), 625
sapphire, 164
scandium oxide (Sc_2O_3), 625
silicon (Si), 104, 451–452, 625
silicon dioxide (SiO_2), 198, 415, 450–451, 625
silicon monoxide (SiO), 193, 398, 453, 625, 89–90
silicon nitride (Si_3N_4), 453–454, 625
silicon oxide, 164, 193
silicon oxynitride, 453–454
silicon substrate, 89, 90
silver, 169, 185, 264, 265
sodium fluoride (NaF), 625
stibnite, 199
strontium fluoride (SrF_2), 625
Substance H1, 456, 627
Substance H2, 456, 627
Substance H4, 456, 627
Substance M1, 456, 627
tantalum pentoxide (Ta_2O_5), 626
tellurium (Te), 452, 626
thallous chloride ($TlCl$), 626
thorium fluoride (ThF_4), 193, 453, 626
thorium oxide (ThO_2), 626
titanium dioxide (TiO_2), 193, 405, 449–450, 626
ytterbium fluoride (YbF_3), 626
yttrium oxide (Y_2O_3), 626
zinc selenide ($ZnSe$), 626
zinc sulphide (ZnS), 89, 192–195, 196, 197, 198, 203, 274, 279, 364, 398, 405, 447, 453, 626
zirconium dioxide (ZrO_2), 127, 193, 451, 627
matrix, characteristic, 39
maximum potential transmittance, 331
Maxwell, James Clerk, 2
Maxwell's equations, 12
measured performance of filters, 342–345
measured performance, induced transmission filter for uv, 343
measurement of optical constants, *see* optical property measurement
mechanical property measurement, 436–445
stress, titanium oxide, 439, 441
metal with dielectric overcoat, p-polarisation reflectance dip, 357–358
s-polarisation reflectance dip, 356–357
tilted performance, 355–357
metal–dielectric filters, *see also* induced transmission filters
characteristic, 261
drift, 265
Fabry–Perot filter, 260–266
heat reflecting coatings, 583–585
manufacture, 264–266
typical bandwidth, 264
typical performance, 264
metals at oblique incidence, 353–355
methyltrimethoxysilane, 415
mica as Fabry–Perot spacer, 280–281
microstructure, 462–478
columnar growth, 463, 464
crystalline, 465–467

- defects, 467–468
influence on film behaviour, 462–478
nodules, 469, 471
mid-frequency sputtering, 406–407
mirrors,
 aluminium, 158–169
 neutral, 158–169
mixtures,
 cerium fluoride and zinc sulphide, 454
 cerium oxide and cerium fluoride, 454
 cerium oxide and magnesium fluoride, 454
 germanium and magnesium fluoride, 454
 germanium and selenium, 456
oxides, 455–456
silica, mixed with other oxide, 455
various, 454–456
zinc sulphide and cryolite, 454
zinc sulphide and magnesium fluoride, 455
modified admittances, 349–353, 350
moisture adsorption, 468–474
molybdenum boats, 397
monitoring,
 accuracy and stability, 518–519
 direct, 515
 direct turning value, 515–517
 error compensation in direct turning value, 516–518
layer sensitivity, 518
optical, *see* optical monitoring
quartz crystal, 509–511
quartz crystal error compensation, 519–520
simulation, 513–520
tolerances, *see* tolerances in monitoring
MTMOS, 415
multilayer phase retarders, *see* phase retarders
multiple-cavity filters, 293–306
 effect of tilting, 315
higher performance, 306–319
improved matching, 308–319
Knittl's method, 299
losses, 316–319
metal–dielectric filters, 325–342
ripple, 304–306
Smith's method, 294–300
Thelen's method, 300–306
Musset and Thelen technique, 104–108
Mylar, as Fabry–Perot spacer, 282
n and *k* extraction, *see* optical property measurement
narrowband filters, 260–345
neodymium fluoride (NdF_3), 624
neodymium oxide (Nd_2O_3), 625
neutral density filters, 176–177
neutral mirrors, 158–169
Nevière and Vincent, 357
Newton, Sir Isaac, 2
Nichrome, 159, 176–177
nodules, 467–468, 469, 471
non-polarising coatings, 368–377
 reflectors, high angles of incidence, 374–377
 reflectors, Thelen's technique, 376–377
non-quarterwave monitoring, 506
oblique incidence, 23–30
 oblique incidence metals, 353–355
 oblique incidence optical admittance for, 27–28
optical admittance, 16
 characteristic admittance, 16
 equivalent admittance, 216–220
 oblique incidence, 27–28
optical distance, 15
optical monitoring,

- broadband systems, 508–509
maximètre, 508, 518–519
non-quarterwave, 506
photoelectric, 501
precision in turning value, 504–506
reflectance or transmittance?, 502
Ring and Lissberger, 508
techniques, 500–509
temperature effects, 507
turning value, 504
typical arrangement, 501, 503
visual method, 501
zinc sulphide problems, 507
- optical path, 15
optical property measurement, 418–436
Abelès technique, 428–429
Cauchy expression, 435
ellipsometric technique, 429–432
envelope technique, 427–428
Hacskaylo technique, 429
Hadley method, 425–426
inhomogeneous films, 432–436
Netterfield method, 436
Pelletier method, 426–427
quarterwaves, 420–421
optical tunnel filters, 389–390
 Baumeister, 389
order-sorting filters for grating spectrometers, 559–570
Otto coupling, 361
oxide mixtures, 455–456
- packing density, 463–464
 effect on film index, 463–464
pass band, transmission in, 226–228
peak transmittance, variation over surface, 343–344
PECVD, 415
performance of filters, measured, 342–345
performance specification, 523–529
Pfund, A H, 4
- phase retarders, 482–389
 Apfel's technique, 385–389
 multilayer, 385–389
 quarter and half wave, 382–389
phase shift on reflection, φ , 45, 186
phase shift, on transmission, 45
phase thickness, equivalent, 216–220
phase velocity, 600
phase-dispersion filter, 319–325
physical vapour deposition, 394–413
- plane waves, 14
plasma enhanced chemical vapour deposition, 415
plasma polymerisation, 415–416
plate polariser, 366–368
platinum, 159
polariser cube, 367
polariser plate, 366–368
polarising beam splitter, 351, 362–366
potential absorptance, 204–205
potential transmittance, 45, 50–52, 327–333
 maximum, 331
Poynting vector, 17
p-polarised light, definition, 24
praseodymium oxide (Pr_6O_{11}), 625
production methods, 393–456
production of thin films, 394–418
protection of metal films, 160–169
PVD, physical vapour deposition, 394–413
- quarter and half wave retarders, 382–389
quarter-wave stack, 185–193, 211–213
 Herpin index, 215–220
quarter-wave thicknesses, theory, 52–53
quartz, 193
quartz crystal monitoring, 509–511

- error compensation, 519–520
race track, 405
radio frequency sputtering, 408
Ramsay and Ciddor, 200
Rayleigh criterion, 183
Rayleigh, Lord, 3–4
reactive evaporation, 448–449
reactive low-voltage ion plating, 408–410
refinement and synthesis,
 problems, 617–619
 refinement, 195, 233, 611–616
 synthesis, 611, 616–617
 techniques, 613–617
reflectance, 23, 43–45
reflectance of a thin film, theory, 39–44
reflection, incoherent at two or more surfaces, 67–72
reflectors,
 apparent curvature, 200–203
 losses, 204–208
 multilayer dielectric, 185–208
 non-polarising, high angles of incidence, 376–377
refractive index, 14
 complex, 14
resolution, 183
retarders, *see* phase retarders
rhodium, reflectance, 159
ripple,
 advanced elimination, 233–244
 origin of, 227
 reduction of in pass band, 228–230
Rouard, Pierre, 4
rugate filters, 588–598
 Fourier expression for design, 598
 Q function, 598
samarium fluoride (SmF_3), 625
sapphire, 164
scandium oxide (Sc_2O_3), 625
Schuster diagram, 97, 100, 112
sensitivity to contamination, 478–485
shortwave pass filter,
 absorption filters, 246
 design, 232
sideband blocking, 293
silica, mixed with other oxide, 455
silicon (Si), 104, 452, 625
silicon dioxide (SiO_2), 198, 415, 450–451, 625
silicon monoxide (SiO), 89–90, 193, 398, 453, 625
silicon nitride (Si_3N_4), 454, 625
silicon oxide, 164, 193
silicon oxynitride, 454
silver, 169, 185, 264, 265
 admittance loci, 355
 metal–dielectric filters, 326–342
reflectance, 159
simple boundary, 18–27
simulation of monitoring, 513–520
Smith chart, 77–80
Smith's method, 75–77
sodium fluoride (NaF), 625
sol–gel process, 416
solid etalon filter, 280–283
specification of filters, 523–534
 performance, 523–529
spectrally selective absorbers, 579–583
s-polarised light, definition, 24
sputtering, 405–408
 DC planar magnetron, 405–408
 mid-frequency, 406–407
 radio frequency, 408
 twin magnetron, 407
stibnite, 199
stop band,
 transmission at centre, 225–226
 transmission at edge, 223–225
stress, effects of impurities, 440
Strong, John, 4

- strontium fluoride (SrF_2), 626
Substance H1, 456, 627
Substance H2, 456, 627
Substance H4, 456, 627
Substance M1, 456, 627
substrate cleaning, 497–498
 glow discharge, 498–499
 preparation, 497–499
substrate temperature during deposition, 403–405
substrates, series of, 67–72
surface plasma wave, 361
surface plasmon, 361
surface, effect of second, 67–72
symmetrical multilayers, 213–232
symmetrical periods, 72–73
 in multiple-cavity filters, 300–306
synthesis, *see* refinement and synthesis
- TADI filter, 294
tangential components of field, definition, 26
tantalum boat, 397, 402
tantalum pentoxide (Ta_2O_5), 626
target poisoning, 406
Taylor, Dennis, 4
tellurium (Te), 452, 626
temperature,
 cycling of filters, 344
 effect on filters, 344–345, 474–477
 of substrate during deposition, 403–405
TEOS, 415, 416
test specification, 527–529
 abrasion resistance, 530–532
 adhesion, 533
 environmental resistances, 533–534
jig marks, 529
physical properties, 530–534
pinholes, 528
- Scotch tape test, 533
spatter, 528–529
stains, 529
- tetraethoxysilane, 415
tetraethylorthosilicate, 416
tetramethoxysilane, 415
thallous chloride (TlCl), 626
theory,
 alternative method, 73–75
 basic, 12–85
 summary of important results, 46–50
- thermal evaporation, 395–405
 boats, 397–401
- thickness distribution, *see* uniformity
- thickness monitoring, 499–511
- thin films, production, 394–418
- thin-film absorption filters, 210–211
- thin-film dielectric materials, properties, 621–627
- thin-film materials, 446–456
- third order dispersion, 606
- thorium fluoride (ThF_4), 193, 453, 626
- thorium oxide (ThO_2), 626
- THW filter, 257, 299–300
- tilted antireflection coatings, 377–382
 p-polarisation, 378–379
 s-polarisation, 379–381
 s- and p-polarisation, 381–382
- tilted coatings, 348–391
- tilted non-polarising edge filter, 368–374, 379
- tilting,
 effect on multiple-cavity filters, 315
 effect on single-cavity filters, 283–292
- effective index, 284
- Pidgeon and Smith method of calculation, 284–292

- titanium dioxide (TiO_2), 193, 449–450, 626
measurement of stress, 439, 441
- titanium tetraethoxide, 416
- TMMOS, 415
- TMOS, 415
- TOD, *see* third order dispersion tolerances,
in monitoring, 511–520
Monte Carlo methods, 513–514
permissible in various coatings, 514
survey of early work, 511–513
- toxicity, 445–446
- transmittance, 23, 43–45
potential, *see* potential transmittance
symmetry of, 53–54
- trimethylmethoxysilane, 415
- tungsten boat, 397, 402
- tunnel filters, *see* optical tunnel filters
- turning value monitoring, 504
- twin magnetron sputtering, 407
- ultrafast coatings, 599–609
- ultraviolet, materials for, 451
- uniformity, 488–497
directed surface source, 489
domed work holder, 495, 496
flat plate, 490
- Holland and Steckelmacher's method, 489–495
- planetary jigs, 495
- point source, 489
- rotating substrates, 490–495
- spherical surface, 490
- use of masks, 496–497
- units, 46
- variation of peak wavelength with temperature, 344
- varying angle of incidence, 283–292
- vector method, theory, 66–68
- WADI filter, 293, 299
- Young, Thomas, 2
- Young's technique, 108
- ytterbium fluoride (YbF_3), 626
- Ytralox, as Fabry–Perot spacer, 282
- yttrium oxide (Y_2O_3), 626
- zinc selenide (ZnSe), 626
- zinc sulphide (ZnS), 192–195, 193, 196, 198, 203, 274, 279, 364, 398, 405, 447, 453, 626, 89
temperature coefficient of optical thickness, 344
- zirconium dioxide (ZrO_2), 127, 193, 451, 627