The Physics of Music and Color

Leon Gunther

The Physics of Music and Color

Springer

Leon Gunther
Department of Physics and Astronomy
Tufts University
Medford, Massachusetts
USA
l.gunther@tufts.edu

ISBN 978-1-4614-0556-6 e-ISBN 978-1-4614-0557-3
DOI 10.1007/978-1-4614-0557-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011934793

Dedicated to my mother, Esther (Weiss) Gunther Wand, who nurtured me with a deep appreciation of music and the beauty of nature, and to my wife, Joelle (Cotter) Gunther, who sustains me with her love and wisdom

Preface

This textbook has its roots in a course that was first given by Gary Goldstein and me at Tufts University in 1971. Both of us are theoretical physicists, with Gary focusing on the study of elementary particles and me focusing on condensed matter physics, which is the study of the fundamental behavior of various types of matter – superconductors, magnets, fluids, among many others. However, in addition, we both have a great love and appreciation for the arts. This love is fortunately also manifested in our involvement therein: Gary has been seriously devoted to oil painting. I have played the violin since I was seven and played in many community orchestras. I am also the founder and director of a chorus. Finally, I am fortunate to have a brother, Perry Gunther, who is a sculptor and my inspiration and mentor in the fine arts.

It is common to have a course on either the Physics of Music or the Physics of Color. Numerous textbooks exist, many of which are outstanding. Why did we choose to develop a course on both music and color? There are a number of reasons:

1. The basic underlying physical principles of the two subjects overlap greatly because both music and color are manifestations of wave phenomena. In particular, commonalities exist with respect to the production, transmission, and detection of sound and light. Our decision to include both music and color was partly due to the fact that some wave phenomena are relatively easy to demonstrate for sound but not for light; they are experienced in every day life. Examples include diffraction and the Doppler effect. Thus, the study of sound helps us understand light. On the other hand, there are some wave phenomena – common to both sound and light – that are more easily observed for light. An example is refraction, wherein a beam of light is traveling through air and is incident upon a surface of glass. Refraction causes the beam to bend upon passing into the glass. Refraction is the basis for the operation of eyeglasses. And finally, there are wave phenomena that are easily observable for both sound and light. Interference is an example.

Two stereo loudspeakers emitting a sound at the same single frequency produce dead (silent) regions within a room as a result of the interference between the two sound waves produced by the two loudspeakers; the colors observed on the CDs of the photo in the frontispiece are a result of the interference of light reflected from the grooves within the CDs.

2. The production of music and color involves physical systems, whose behavior depends upon a common set of physical principles. They include vibrating mechanical systems (such as the strings of the violin or the drum, vibrating columns of air in wind instruments and the organ), electromagnetic waves such as light, the rods and cones of the eye, and the atom. All manifest the existence of **modes** and the phenomena of excitation, resonance, energy storage and transfer, and attenuation.

   CDs "produce" sound through a series of processes that involve many distinct physical phenomena. First, the CD modulates a laser beam that excites an electronic device into producing an electrical signal. The laser light itself is a manifestation of electric and magnetic fields. The electrical signal is used to cause the cone of a loudspeaker to vibrate and produce the motion in air that is none other than the sound wave that we hear.
3. The course that led to the writing of this book offers us the opportunity to study a major fraction of the basic principles of physics, with an added important feature: Traditionally, introductory physics courses are organized so that basic principles are introduced first and are then applied wherever possible. This course, on the other hand, is based on a motivational approach: Because of the ease of observing most phenomena that is afforded by including both light and sound, we are able to introduce the vast majority of topics using class demonstrations.

   We challenge ourselves by calling for a physical basis for what we observe. We turn to basic principles as a means of understanding the phenomena. A study of both subjects involves pretty nearly the entire gamut of the fundamental laws of classical as well as modern physics. (The main excluded areas are nuclear and particle physics and relativity.)

Ultimately, our approach helps us appreciate a central cornerstone of physics – to uncover a minimal set of concepts and laws that is adequate to describe and account for all physical observations. Simplification is the motto. We learn to appreciate how it is that because the laws of physics weave an intricate, vast web among physical phenomena, physics (and science generally) has attained its stature of reflecting what some people refer to as "truth" and, much more significantly, of having an extraordinarily high level of dependability.

The prerequisites for the associated course are elementary algebra and a familiarity with the trigonometric functions. The only material in the textbook that requires a higher level of mathematics is the appendix on the Transformation of Color Matching Functions (Appendix I) from one set of primaries to another – the analysis requires a good understanding of matrices. I have never included this appendix in my course; it is available for those who might be interested in it. The level of the textbook is such as to produce questions as to whether a student

without inclinations to major in the sciences can handle the material. It has been my experience in teaching the associated course at Tufts University for over 35 years, that very few such students have failed to do well in the course. In the Fall, 2009 semester, in particular, the 15 students who took the course were all majoring in the Arts, Humanities, or Social Sciences or as yet had not declared a major. The average score on the Final Exam was a respectable 73%, with a range from 61% to 94%.

When I have taught the course using this textbook, I have often had to omit the section on Polarized Light for want of time. Sections that can be skipped without loss of continuity for the remaining material are marked with a double asterisk (**).

Note on problems and questions: Whether you are reading this book in connection with a course you are taking or reading it on your own, I strongly urge you to take the questions and problems in the book very seriously. To test your understanding and to measure your level of understanding, you have to do problems. In all my more than 50 years of studying physics, I have never truly appreciated a new subject without doing problems.

There are many fine books already available that cover either the physics of sound and music or the physics of light and color. Some of these books go into great depth about a number of the subjects, way beyond the depth of this book. For example, you will not find details on the complex behavior of musical instruments in this book. The book by Arthur Benade, listed in the Appendix of references D, is a great resource on this subject, even though it is quite dated. And, you will not find in-depth coverage of the incredibly rich range of light and color phenomena that is treated in the wonderful book by Williamson and Cummins. Their section on oil paint is outstanding. Instead, you should look on this book as a resource for gaining an in-depth understanding of the relevant concepts and learning to make simple calculations that will help you test hypotheses for understanding phenomena that are not covered in this book. You will be able to read other books and articles on the web empowered with an understanding that will help you appreciate the content. One of the problems raging today (2011) is the proliferation of information. Ah yes, you can look up on the Web any topic in this book. Unfortunately, a huge fraction of the information is incorrect or unreliable.[1] How can you judge what you read? The

[1]Recently, the SHARP Corporation announced that it was going to make available a color monitor and TV that has **four primary colors** among the color pixels, in contrast to the three primaries currently used. As a result, it claimed that the number of colors available would approach one trillion. (See their website: http://www.sharpusa.com/AboutSharp/NewsAndEvents/PressReleases/2010/January/2010_01_06_Booth_Overview.aspx) Yet you will learn in Chap. 14 that human vision can differentiate only about ten million colors. Therefore, even if the Sharp monitor were able to produce one trillion colors, viewers would not be able to benefit from this great technology. We can still ask what can possibly be the gain in adding a yellow primary? Is their chosen color yellow for the fourth primary the best one to choose to improve our color vision? See Chap. 14 for information on this question. Websites abound dealing with the significance of Sharp's new technology; this book will help you analyze and judge what you read.

only solution is for you to accumulate knowledge and understanding of the basics and to criticize what you read.[2]

Acknowledgements First and foremost I am indebted to Gary Goldstein, who was a co-developer of the original course on The Physics of Music and Color. Gary's contributions in teaching a number of the subjects in a clear way were invaluable. Most noteworthy were his ideas for teaching color theory. I am grateful to my daughter, Rachel Gunther, for producing the first word-processed draft of the book. I am deeply indebted to Stephan Richter, one of my graduate students, who was a driving force and indefatigable in producing a Latex copy of the book, worked over numerous figures, and is responsible for the layout of the book. I had a number of teaching assistants over the years who made very valuable contributions in teaching the course, most notably Stephan Richter and Rebecca Batorsky. Both Stephan and Rebecca are gifted teachers and frequently shared productive advice for me. My long time friend and violin teacher, Wolfgang Schocken, was a well-known teacher of the violin. He was also extremely knowledgeable about the numerical issues involved in intonation, which he shared with me. In spite of my familiarity with resonances and overtones of a vibrating string, it was he who taught me to listen carefully to the resonant vibration of unbowed strings to vastly improve my intonation. My son, Avi Gunther, who got his Bachelor's degree from the Berklee College of Music in Boston with a major in Music Production and Engineering, was often extremely valuable in advising me on many aspects of music and on sound production.

I benefitted greatly from two readers of this book: The first reader was my personal opthalmologist, Dr. Paul Vinger, who pointed out numerous typos and provided me with questions that he suggested be addressed in the book. My second reader was a student of mine, Bryce Meyer, who did an incredibly dedicated job reading carefully through the book – finding typos and making countless suggestions for improving the clarity of various passages in the text. Bryce also helped me with some figures.

Many individuals have helped me in one way or another toward the writing of this book. I list the following with apologies those who should be here but are omitted: Paavo Alku, Anandajoti Bhikku, Bruce Boghosian, Andrew Bregman, Andrew Clarke, David Copenhagen, Tom Cornsweet, Russ Dewey, Marcia Evans, Oliver Knill, Paul Lehrman, Ken Lang, Jay Neitz, Donna Nicol, Ken Olum, Charles Poynton, Jeffrey Rabin, Brian Roberts, Judith Ross, Eberhard Sengpiel, George Smith, and Raymond Soneira. This book would not have been published were it not for the strong support and help of my editors, Christopher Coughlin and HoYing

[2]What applies to information on science applies to all subjects. If you are given a multitude of conflicting **expert** opinions on a subject, you will tend to choose one expert who is closest to your point of view or you will want to throw all the sources out the window with the conclusion that reliable information not only cannot be found but has no meaning. The fascinating book by Neil Postman – *Amusing Ourselves to Death* [Penguin Books, N.Y, 1986] – discusses some related problems connected with this proliferation of information.

Fan. I want to pay special attention to Kaća Bradonjić, who produced tens of figures with great finesse, especially those in Chap. 5 that are based on my crude hand drawings.

This book has been a work in progress for more than 35 years. It has had many drafts. I need to share with you my deep appreciation for my loving wife, Joelle, for supporting me in this effort. Whenever I needed encouragement to sustain my spirits and energy, Joelle was there for me.

Questions Discussed in This Book

1. Why is the sky blue and the setting sun red?
2. How does the rainbow get its colors?
3. How is it that all light is a mixture of the colors of the rainbow? Yet the color brown is not simply a mixture of these colors?
4. How is it that sound can bend around corners?
5. Does light bend around corners?
6. What simple mathematical relationships form the bases of the musical scales of most of the world's cultures? Are these relationships unique?
7. Are there three primary colors?
8. What are the colors white, black, gray, and brown?
9. How is the eye like a camera?
10. How is it that the ear can perceive two distinct musical tones, yet the eye perceives a mixture of two colors as a single color?
11. How can we get color from purely black and white images?
12. How does the brain determine the direction of a source of sound?
13. What is noise?
14. Why does the trumpet sound different from the violin?
15. What is a mirage?
16. Why do stars seem to twinkle?
17. How do color prints, color slides, and color TV work?
18. Can a soprano really break glass?
19. Why does a flutist have to retune his or her flute a while after having begun playing?
20. How is sound transmitted electrically?
21. How does the ear provide us with a sense of pitch?
22. Can a fish hear a fisherman talking?
23. Why do some automobiles rattle at a speed of about 55 mph?
24. How can we hear sounds which are not in the air? How is this phenomenon related to the blue color of the ocean?

25. How can a person hear a clock ticking at a frequency of one tick per second, while it is said that the lowest frequency that can be heard is about 20 cycles per second?
26. How can we estimate the speed of an overhead propeller-driven airplane from the sound it emits?
27. How does the vibrato of a violin help improve our perception of consonance among groups of notes?
28. Why does it become more difficult to perceive a sense of pitch as we play ever lower-pitched notes on a piano?

Contents

1 Introductory Remarks 1
1.1 The Legend of the Huang Chung 7

2 The Vibrating String 11
2.1 Waves Along a Stretched String 11
2.2 A Finite String Can Generate Music! 13
2.3 Pitch, Loudness, and Timbre 16
2.4 The Relation Between Frequency and Pitch 17
2.5 The Wave Motion of a Stretched Rope 18
2.6 Modes of Vibration and Harmonics 20
2.7 The Sine Wave 23
2.8 The Simple Harmonic Oscillator 26
2.8.1 The Vibration Frequency of a Simple Harmonic Oscillator 28
2.9 Traveling Sine Waves 29
2.9.1 Applications 31
2.10 Modes of Vibration: Spatial Structure 32
2.11 The Wave Velocity of a Vibrating String 34
2.11.1 Application of the Above Relations to the Piano 37
2.12 The Connection Between an SHO and a Vibrating String 38
2.13 Stiffness of a String 41
2.14 Resonance 43
2.15 General Vibrations of a String: Fourier's Theorem 45
2.15.1 Frequency of a Wave with Missing Fundamental 51
2.16 Periodic Waves and Timbre 52
2.17 An Application of Fourier's Theorem to Resonance Between Strings 52
2.18 A Standing Wave as a Sum of Traveling Waves 55
2.19 Terms 55
2.20 Important Equations 57
2.21 Problems for Chap. 2 58

3 The Vibrating Air Column 63
3.1 The Air of Our Atmosphere 63
3.1.1 Generating a Sound Pulse 66
3.1.2 Digression on Pushing a Block of Wood 67
3.2 The Nature of Sound Waves in Air 67
3.3 Characterizing a Sound Wave 69
3.4 Visualizing a Sound Wave 70
3.5 The Velocity of Sound 71
3.5.1 Temperature Dependence of Speed of Sound in Air 72
3.6 Standing Waves in an Air Column 73
3.6.1 Standing Waves in a Closed Pipe 76
3.6.2 End Correction for Modes in a Pipe 79
3.7 Magic in a Cup of Cocoa** 79
3.8 Terms 80
3.9 Important Equations 80
3.10 Problems for Chap. 3 81
3.10.1 Derivation of the Helmholtz Formula 84

4 Energy 87
4.1 Forms of Energy and Energy Conservation 88
4.1.1 Fundamental Forms of Energy 89
4.1.2 "Derived" Forms of Energy 93
4.1.3 The Energy of Cheerios 94
4.2 The Principle of Conservation of Energy, Work, and Heat 95
4.3 Energy of Vibrating Systems 96
4.3.1 The Simple Harmonic Oscillator 96
4.3.2 Energy in a Vibrating String 98
4.3.3 Energy in a Sound Wave 99
4.4 Power 99
4.5 Intensity 101
4.6 Intensity of a Point Source 103
4.7 Sound Level and the Decibel System 105
4.7.1 Logarithms 105
4.7.2 Sound Level 107
4.7.3 From Sound Level to Intensity 108
4.8 Attenuation 110
4.8.1 Attenuation in Time 110
4.8.2 Resonance in the Presence of Attenuation 113
4.8.3 Attenuation of Travelling Waves: Attenuation in Space 114
4.9 Reverberation Time 118
4.10 Terms 120
4.11 Important Equations 121
4.12 Problems for Chap. 4 122

5 Electricity, Magnetism, and Electromagnetic Waves 127
5.1 The Fundamental Forces of Nature 127
5.2 The Electric Force 129
5.3 Electric Currents in Metal Wires 130
5.4 The Magnetic Force 131
5.5 Magnetic Forces Characterized 133
5.6 Is There a Connection Between Electricity and Magnetism? 135
5.6.1 Action–Reaction Law and Force of Magnet on Current-Carrying Wire 138
5.7 The Loudspeaker 141
5.8 The Buzzer 141
5.9 The Electric Motor 142
5.10 Force Between Two Wires Carrying an Electric Current 143
5.11 The Electromagnetic Force and Michael Faraday 143
5.12 Applications of Faraday's EMF 147
5.13 A Final "Twist" 148
5.14 Action-at-a-Distance and Faraday's Fields 149
5.15 The Electric Field 150
5.16 The Magnetic Field 154
5.17 Magnetic Force on a Moving Charge 157
5.18 Force Between Two Parallel Wires Carrying Currents 158
5.19 Generalized Faraday's Law 158
5.20 What Do Induced Electric Field Lines Look Like? 163
5.21 Lenz's Law 164
5.22 The Guitar Pickup 166
5.23 Maxwell's Displacement Current 167
5.24 Electromagnetic Waves 169
5.25 What Is the Medium for Electromagnetic Waves? 174
5.26 The Sources of Electromagnetic Waves 175
5.27 Terms 177
5.28 Important Equations 178
5.29 Problems for Chap. 5 178

6 The Atom as a Source of Light 179
6.1 Atomic Spectra 179
6.2 The Hydrogen Spectrum of Visible Lines 181
6.3 The Bohr Theory of the Hydrogen Atom 184
6.4 Quantum Theory 190
6.5 Complex Scenarios of Absorption and Emission 195
6.5.1 Rayleigh Scattering 196
6.5.2 Resonance Fluorescence 196
6.5.3 General Fluorescence 196
6.5.4 Stimulated Emission 197
6.6 Is Light a Stream of Photons or a Wave? 199
6.7 The Connection Between Temperature and Frequency 200

6.8 Terms 202
6.9 Important Equations 202
6.10 Problems for Chap. 6 203

7 The Principle of Superposition 205
7.1 The Wave Produced by Colliding Pulses 205
7.2 Superposition of Two Sine Waves of the Same Frequency 207
7.3 Two Source Interference in Space 209
7.3.1 Sound Level with Many Sources 216
7.3.2 Photons and Two-Slit Interference 216
7.4 Many-Source Interference 217
7.4.1 Gratings 217
7.4.2 Diffraction Through a Mesh** 218
7.4.3 X-ray Diffraction of Crystals 220
7.5 Beats 221
7.6 Terms 224
7.7 Important Equations 224
7.8 Problems for Chap. 7 225

8 Propagation Phenomena 231
8.1 Diffraction 231
8.1.1 Scattering of Waves and Diffraction 238
8.1.2 Why Is the Sky Blue? 240
8.2 Reflection 241
8.2.1 A Complex Surface: A Sand Particle 244
8.3 Reflection and Reflectance 245
8.3.1 The Reflectance for a Light Wave 246
8.3.2 The Reflectance for a Sound Wave 248
8.4 Refraction 249
8.5 Total Internal Reflection 251
8.6 The Wave Theory of Refraction 252
8.7 Application to Mirages 255
8.8 The Prism 256
8.9 Dispersion 257
8.9.1 Effect of Dispersion on a Prism 257
8.9.2 Effect of Dispersion on Fiber Optics Communication . 258
8.10 Lenses 259
8.10.1 The Converging Lens 259
8.10.2 Lens Aberrations 260
8.10.3 Image Produced by a Converging Lens 264
8.10.4 Magnification 266
8.10.5 Reversibility of Rays: Interchange of Object and Image 269
8.10.6 The Diverging Lens 269
8.10.7 Determining the Focal Length of a Diverging Lens 271

8.11 The Doppler Effect 272
8.11.1 Doppler Effect for Waves in a Medium 273
8.11.2 Doppler Effect for Electromagnetic Waves in Vacuum 277
8.11.3 Applications of the Doppler Effect 278
8.12 Polarized Light** 280
8.12.1 How Can We Obtain a Beam of Polarized Light? 281
8.12.2 Series of Polarizers 282
8.12.3 Ideal vs. Real Polarizers 283
8.12.4 Sample Problems 284
8.12.5 Partial Polarization of Reflected Light 285
8.12.6 The Polarization of Scattering Light 286
8.12.7 The Polarizer Eyes of Bees 286
8.12.8 Using Polarization of EM Radiation in the Study of the Big Bang 287
8.12.9 Optical Activity 287
8.12.10 Our Chiral Biosphere 291
8.13 Terms 293
8.14 Important Equations 293
8.15 Questions and Problems for Chap. 8 294

9 The Ear 305
9.1 Broad Outline of the Conversion Process 306
9.2 The Auditory Canal 310
9.3 The Eardrum 310
9.4 The Ossicles 311
9.5 Improving on the Impedance Mismatch: Details** 313
9.6 The Cochlea 315
9.6.1 Summary 318
9.7 Pitch Discrimination 319
9.7.1 Some Mathematical Details on Pitch vs. the Peak of the Envelope** 322
9.7.2 Mach's Law of Simultaneous Contrast in Vision 322
9.7.3 Rhythm Theory of Pitch Perception 324
9.8 Terms 325
9.9 Problems for Chap. 9 326

10 Psychoacoustics 327
10.1 Equal Loudness Curves 329
10.2 The "Sone Scale" of Expressing Loudness** 331
10.3 Loudness from Many Sources 334
10.4 Combination Tones and the Nonlinear Response of the Cochlea 335
10.5 The Blue Color of the Sea and Its Connection with Combination Tones 341
10.6 Duration of a Note and Pitch Discrimination 342
10.7 Fusion of Harmonics: A Marvel of Auditory Processing 344
10.7.1 Mathematica File 346

10.8 Additional Psychoacoustic Phenomena 348
10.9 Terms 349
10.10 Important Equations 349
10.11 Problems for Chap. 10 349

11 Tuning, Intonation, and Temperament: Choosing Frequencies for Musical Notes 353
11.1 Musical Scales 355
11.2 The Major Diatonic Scale 358
11.3 Comments Regarding **Western Music** 360
11.4 Pythagorean Tuning and the Pentatonic Scale 362
11.5 Just Tuning and the Just Scale 363
11.6 The Just Chromatic Scale 365
11.7 Intrinsic Problems with Just Tuning 367
11.8 Equal Tempered Tuning 369
11.9 The Cents System of Expressing Musical Intervals 371
11.10 Debussy's Six-Tone Scale 373
11.11 Terms 374
11.12 Important Equations 374
11.13 Problems for Chap. 11 375

12 The Eye 383
12.1 The Cornea and Lens 383
12.2 The Iris 386
12.3 The "Humorous" Liquids of the Eye 387
12.4 The Retina 387
12.5 Dark Adaptation 391
12.6 Depth Perception 391
12.7 Terms 393
12.8 Problems for Chap. 12 393

13 Characterizing Light Sources Color Filters and Pigments 397
13.1 Characterization of a Light Beam 397
13.1.1 Spectral Intensity vs. Intensity 402
13.2 Color Filters 403
13.2.1 Stacking Filters (Filters in Series) 405
13.3 Pigments 409
13.4 Summary Comments on Filters and Pigments 409
13.5 Terms 410
13.6 Important Equations 411
13.7 Problems for Chap. 13 411

14 Theory of Color Vision 413
14.1 A Simplified Version of the Three-Primary Theory 414
14.2 Exploration of Color Mixing with a Computer 416
14.3 Introduction to the Chromaticity Diagram 419
14.4 Metamers 420
14.5 A Crude Chromaticity Diagram 421

14.6 A Chromaticity Diagram of Practical Use 423
14.6.1 The Units for the Admixture of the Three Primaries... 424
14.6.2 Tristimulus Values 425
14.6.3 Color Coordinates..... 426
14.6.4 On the Significance of the Chromaticity Diagram 426
14.7 The Calculation of Color Coordinates 432
14.7.1 Color Coordinates of Butter 435
14.8 Using a Different Set of Primaries 436
14.8.1 General Features of a Different Set of Primaries..... 437
14.9 The Standard Chromaticity Diagram of the C. I. E..... 439
14.10 From Computer RGB Values to Color** 443
14.11 How Many Colors Are There? 445
14.11.1 Limitations of a Broadened Gamut of a Monitor..... 452
14.12 A Simple Physiological Basis for Color Vision 453
14.13 Color Blindness 458
14.14 After-Images 459
14.14.1 Questions for Consideration..... 461
14.15 Terms 462
14.16 Important Equations 462
14.17 Problems on Chap. 14..... 463

A Symbols..... 473

B Powers of Ten: Prefixes 477

C Conversion of Units and Special Constants..... 479

D References for The Physics of Music and Color..... 481

E A Crude Derivation of the Frequency of a Simple Harmonic Oscillator** 485

F Numerical Integration of Newton's Equation for a SHO** 489

G Magnifying Power of an Optical System** 495
G.1 Image with the Naked Eye and with a Magnifying Glass..... 496
G.2 The Microscope 499
G.3 Problems on Magnifying Power..... 500

H Threshold of Hearing, Threshold of Aural Pain, and General Threshold of Physical Pain** 501

I Transformation Between Tables of Color-Matching Functions for Two Sets of Monochromatic Primaries** 507
I.1 Application of the Transformation: Determining an Ideal Set of Primaries..... 509
I.2 Proof of Equations (I.1) and (I.6) 512
I.3 Problems on the Transformation of TCMFs..... 518

J Hommage to Pierre-Gilles de Gennes: Art and Science..... 521

K MAPPINGS as a Basis for Arriving at a Mutually Agreed Upon Description of Our Observations of the World – Establishing 'Truths' and 'Facts' 525
K.1 MAPPINGS as Central to Organizing Human Experience 527
K.2 NUMBERS as a Mapping 527
K.3 The Concept of TIME as a Mapping 528
K.4 Mappings as the Essential Goal of Physics 530

Index 533

Chapter 1
Introductory Remarks

Why should someone be attracted to a book on the Physics of Music and Color? For those people who are well versed in both the sciences and the arts, the question would very likely not arise. But for those who are well versed in but one of these areas, the relationship between the two is probably unclear, if not a total mystery. Let us consider two contrary attitudes to the role the study of physics can make with regards to our sense of the world about us. One is by the great poet Walt Whitman, and the other by the renowned physicist Richard Feynman (Fig. 1.1).

Here is **Walt Whitman's** attitude toward Astronomy. His poem "When I Heard the Learn'd Astronomer" is sardonic:

> When I heard the learn'd astronomer,
> When the proof, the figures, were ranged in columns before me,
> When I was shown the charts and diagrams, to add, divide, and measure them,
> When I sitting heard the astronomer where he lectured with much applause in the lecture-room,
> How soon unaccountable I became tired and sick,
> Till rising and gliding out I wander'd off by myself,
> In the mystical moist night-air, and from time to time,
> Look'd up in perfect silence at the stars.

I wonder whether Whitman would have reacted the same way to the documentary film on the work of **Louis Leakey**, who discovered the remains of **Australopithecus bosei**, a prehistoric form of man that was dated to have existed about one and three-quarter million years ago. Leakey has been described as having worked persistently but unrewardingly for 28 years at the site, before the discovery was made.

There is a scene wherein Leakey is standing on a hilltop overlooking the **Olduvai Gorge** in Kenya. The terrain is devoid of greenery, in fact, lifeless in appearance. Still, Leakey passionately paints word images of the life of the prehistoric people who lived and died in that valley as if they were alive that very day the filming took place. Upon what information were these images based? Merely upon dry pieces of bone and artifacts, most of which would barely be noticed by the average passerby.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_1,

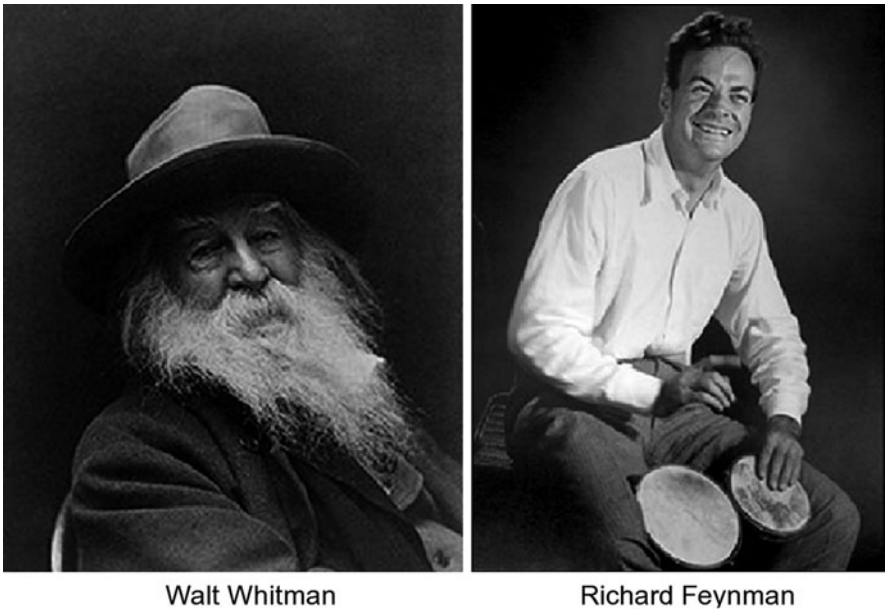


Fig. 1.1 Whitman and Feynman (Whitman photo from http://en.wikipedia.org/wiki/Walt_Whitman; Feynman photo credit: Tom Harvey)

The same can be said of the work of astronomers, astrophysicists, and cosmologists. They have provided us with the images of our solar system, our galaxy, and our Universe, revealed the detailed workings of the stars, charted their life history, and deduced a possible history of the Universe starting with the Big Bang theory – but only after painstaking patient mathematical analysis of astronomical data, an activity that is fuelled by irresistible curiosity, and by egos too!

Still, one need not know any physics to be a successful professional musician or artist, although currently, many artists are making use of physics in their work. The musician must understand the relationships among the various elements that make for a great musical composition, such as musical notes. The musician understands that in some, oftentimes mysterious way, our perception of the specific relationships among these elements exists at various levels, from the subconscious to the conscious levels, so as to produce a sense of esthetic beauty and a variety of emotional responses. There is an obvious underlying degree of order among these elements. The same can be said for the visual artist with respect to a great work of art.

What turns some people off from science? Is it boredom with the subject matter or boredom that is due to an inability to appreciate the content of science? Is there a fear that science will remove the element of mystery, upon which much of our

Fig. 1.2 A photograph of the Langlois Bridge outside Arles (photo credit: **stock.xchng**)

pleasure of music and art is based? Consider the viewpoint of the great physicist **Richard Feynman**, as quoted from his book *What Do You Care What People Think?*:

> I have a friend who's an artist, and he sometimes takes a view which I don't agree with. He'll hold up a flower and say, "Look how beautiful it is", and I'll agree. But then he'll say, "I, as an artist can see how beautiful it is. But you, as a scientist, take it all apart and it becomes dull." I think he's nutty.
>
> First of all, the beauty that he sees is available to other people – and to me, too, I believe. Although I might not be refined aesthetically as he is, I can appreciate the beauty of a flower. But at the same time, I see much more in the flower than he sees. I can imagine the cells inside, which also have a beauty. There's beauty not just at the dimension of one centimeter; there's also beauty at a smaller dimension.
>
> There are the complicated actions of the cells, and other processes. The fact that the colors in the flowers have evolved in order to attract insects to pollinate it is interesting; that means that insects can see the colors. That adds a question: Does this aesthetic sense exist in lower forms of life? There are all kinds of interesting questions that come from a knowledge of science, which only adds to the excitement and mystery and awe of a flower. It only adds. I don't understand how it subtracts.

The fact is that in many ways, the work of the physicist is similar to that of the impressionistic painter. While people marvel at the visual relationships in art, physicists marvel, in addition, at conceptual relationships in theories that describe natural phenomena as revealed by experimental and theoretical analysis.

Consider the Langlois bridge at Arles, France, as shown in the photograph in Fig. 1.2. As seen in the photograph, the bridge normally would not attract much attention to people. Yet Googling this bridge results in quite a number of hits. Many people take the trouble to go out of their way to visit this bridge. Why is this so? Because the painter van Gogh produced a number of paintings of this bridge. A print

Fig. 1.3 A painting of the Langlois Bridge by van Gogh (photo source: http://www.wallpapers-free.co.uk/background/paintings/vincent_van_gogh/the-langlois-bridge-at-arles/)

of one of these paintings is shown in Fig. 1.3. The photograph indicates that the bridge and its surroundings have probably deteriorated quite a bit since van Gogh produced his paintings. Thus, we cannot expect a photographer to be dishonest in presenting a bridge without the color it once had. However, there is an important reason for the interest and attraction in van Gogh's painting.

I suggest the following as a modest response to this question: The human mind cannot absorb and integrate all the information that is transmitted to it by the senses. Nature is too complicated. Van Gogh chose certain elements of the visual field and emphasized them with well-chosen strokes of the brush. Viewing the painting helps you to become more sensitive to and more aware of these elements, so that once you have been "impressed" by the painting, bridges and streams will forever appear very different to you, certainly more alive and vibrant. Thus, I expect that my having appreciated impressionistic paintings for many years have reduced the difference between the visual reality and the painting.

Here is an experiment that I recommend for the reader that confirms this idea for me: Stare at the photograph for about 15 s. Then close your eyes and work to picture the photograph in your mind. Do the same for the painting. When I do so, I find that I can much more easily visualize the painting than I can visualize the photograph, indicating that the reduced focused information in the painting is the reason for this experience. And the particular reduced information selected by the artist makes an intense 'impression' upon us that the photograph cannot provide.

Fig. 1.4 Ravel and Einstein (photo credits: Ravel – http://en.wikipedia.org/wiki/File:Maurice_Ravel_1912.jpg; Einstein – http://commons.wikimedia.org/wiki/File:Albert_Einstein_violin.jpg)

NOTE: My comments are not at all intended to demean the art of photography! The photograph of Feynman at the beginning of this chapter is an example of how a good photographer can capture a moment like nothing else can. One look at this photograph leaves you with a permanent memory of a piece of Feynman's appearance and personality.

What can the study of Physics contribute? Music has significance only as change in TIME, with sound being the only element. On the other hand, for the most part over the ages, artists have focused on static representations of the visual world about us – that is, on SPACE alone. Only in the past century, have visual artists included change in time of the visual field; SPACE and TIME have been united (Fig. 1.4).

It is interesting to consider how **Albert Einstein** viewed the relationship between science and art or music[1]:

"All great achievements of science must start from intuitive knowledge. I believe in intuition and inspiration.... At times I feel certain I am right while not knowing the reason." Thus, his famous statement that for creative work in science, "Imagination is more important than knowledge." But how, then, did art differ from science for Einstein? Surprisingly, it was not the content of an idea, or its

[1] Based on the journal article, Physics Today, March 2010 issue, with quotes from Alice Calaprice's *The Expanded Quotable Einstein. [Princeton University Press, Princeton, N. J., 2000].*

subject, that determined whether something was art or science, but how the idea was expressed. "If what is seen and experienced is portrayed in the language of logic, then it is science. If it is communicated through forms whose constructions are not accessible to the conscious mind but are recognized intuitively, then it is art".

Musicologists and composers would well disagree with Einstein with respect to the absence of logical organization in a great piece of music! Consider, for example, an exchange between the composer, **Maurice Ravel** and the French violinist Andre Asselin who asked Ravel about the role of inspiration in Ravel's Sonata for Violin and Piano. Ravel replied as follows: "Inspiration – what do you mean? No – I don't know what you mean. The most difficult thing for a composer, you see, is choice – yes, choice."[2] For me, "choice" represents logical analysis in musical composition – analysis that is necessary for composing a great original piece of music.

In order to appreciate the difference between science and art, consider the following: Imagine yourself standing next to a stream of water in the woods. Consider how we observe a stream flowing with our eyes. We can observe waves moving along the surface of the water. The painter provides us with focused static content. Cartoons can provide us with dynamical representations of our experience but they typically fall short in being convincing in accuracy. Videos can do a better job. Yet both cartoons and videos are two dimensional. How can we extend the focused static information provided by the painter to a focused dynamical level? Physics provides this extension for us. Moreover, the physicist seeks to determine the RELATIONSHIPS that connect all physical phenomena; it is the revelation of these relationships that excites a physicist.

The physicist would seek to understand questions like:

- How does light produce the image of the trees and the bridge on the water?
- What tension must there be in the cables and stresses in the wood to keep the sections at rest. This information can lead to information about how the cable is responding to the tension and how the wood is responding to these stresses. We can compare this study to the interest we have as to how various psychological stresses affect one's emotional state. Scientific study of the wood gives the wood a life of its own.
- What is the nature of the water waves on the stream? How can we characterize their shape and how they evolve, move, and disappear?
- Given that the waves are produced by breezes and wind, what is the relationship between the wind characteristics, such as the wind velocity, and the waves and surface textures produced?
- What determines the apparent color of any object and whether the surface of the object is shiny or dull?

[2]Taken from *A Ravel Reader: Correspondence, Articles, Interviews*, by Arbie Orenstein, (Columbia University Press, New York, 1990).

These are not questions that would necessarily bore an artist. If one learns to synthesize one's knowledge, analysis, through a familiarity with Physics, can only add to one's appreciation of nature.

Often people are turned off by the heavy mathematical analysis that dominates Physics and is its essential language. Yet music and mathematics have been inseparable throughout history. Most significantly, it was recognized long ago that pleasurable music is connected with ratios of small integers. This fact is exemplified by the ancient Chinese Legend of the **Huang Chung** (meaning "yellow bell"), the earliest known account of which is due to **Leu Buhwei** (226 BC). This legend is believed to be over 3,000 years old.

1.1 The Legend of the Huang Chung

Emperor Huang Ti one day ordered Ling Lun to make pitch pipes. Ling Lun needed a mathematical recipe for their construction both to end up with pleasing sounds and to be able to have an instrument that could be played along with other instruments. So Ling Lun went from the West of the Ta Hia country to the north of Yuan Yu mountain (see Fig. 1.5). Here Ling Lun took bamboos from the valley Hia Hi. He made sure that the sections were thick and even, and he cut out the nice sections. Their length was 81 lines, that is, about 9 in.

He blew them and made their tone the starting note, the huang chung, of the scale. (The huang chung had the same pitch as Ling Lun's voice when he spoke without passion.)

He blew them and said: "That is just right." Then he made 12 pipes. With what notes? Well, he heard Phoenix birds singing at the foot of the Yuen Yu mountain. From the male birds he heard six notes and from the female birds he heard six notes.

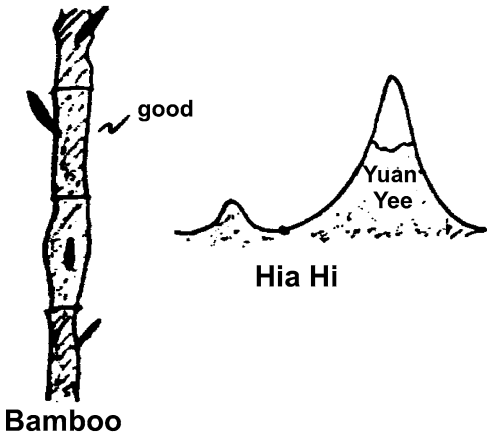


Fig. 1.5 Bamboo from the Ta Hia country

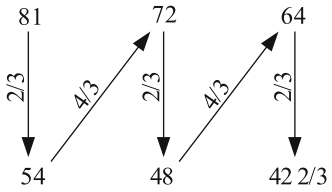


Fig. 1.6 Generating the Chinese scale from the huang chung generator

These are the lengths of the pipes. (Our current notation for the notes is added for reference purposes.)

Male Pipes:	F	G	A	B	C#	D#
	81	72	64	57	51	45

Female Pipes:	C	D	E	F#	G#	A#
	54	48	43	38	34	30

- F is the "huang chung" (yellow bell)
- G is the "great frame"
- A is "old purified"
- C is the "forest bell"
- D is the "southern tube"

What is the basis for these numbers? Here is the recipe for the Chinese scale as recorded in China:

"From the three parts of the 'huang chung generator' reject one part, making the 'inferior generator' (hence equal to 2/3 of the huang chung generator). Next, take three parts of the new (i.e. inferior) generator and **add** one part, making the 'superior generator' (hence equal to $4/3$ of the inferior generator)...," and so forth.

The lengths of the pipes are based on repeated applications of the factor $2/3$ and $4/3$ on the basic length of the huang chung generator. THUS:

The coincidence between what was considered esthetically pleasing musically and the role of ratios of small integers and hence mathematics, or as the sixth century AD Roman philosopher **Boethius** put it, the coincidence between "sensus and ratio" (senses and reason) had a significant, meaningful effect on people. The pre-Socratics began a tradition of lack of trust in the senses as not providing truth about reality. **Truth** is obtained from **thought**. Thus, one should not trust the senses to produce an acceptable version of the musical interval called the "fifth"; one should use an exact ratio of 3:2 of string lengths or pitch pipe lengths.[3] It should not be surprising that people would be very curious as to why the two – mathematics and music – should be connected. The answer must necessarily lie in mathematics and physics and their ramifications in the nature of the human body and mind (Fig. 1.6).

[3]How interesting it is that in recent times, a large fraction of society abhors the possible squelching of the senses by excessive thought.

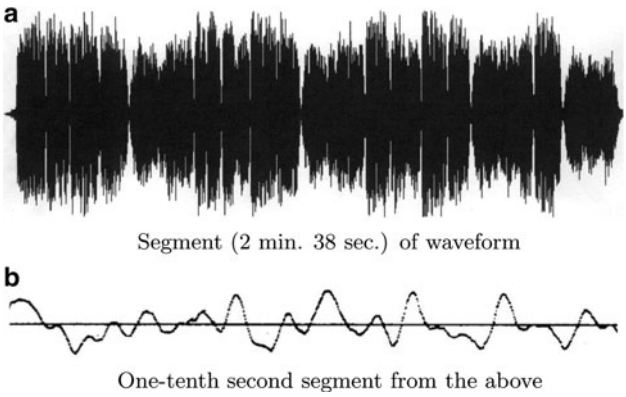


Fig. 1.7 Waveform of *Adon Olam*, by Salomon de Rossi. (**a**) Segment (2 min 38 s) of waveform. (**b**) One-tenth second segment from the above

Later on, armed with some background physics, we will try to provide answers to this question. In particular, in the context of the Legend of the Huang Chung, we will discuss the possible choices of the pipe lengths. In Chapter 11, TUNING, INTONATION, AND TEMPERAMENT: CHOOSING FREQUENCIES FOR MUSICAL NOTES, we will demonstrate that within the framework of the level of complexity of the classical music of these past few hundred years, the desire for an omnipresence of ratios of small integers, which is connected with consonant musical intervals, cannot possibly be satisfied for purely mathematical reasons.

In our study of the Physics of music and color, we will study the nature of sound and light. Analysis will be our focus. Many people find too detailed an analysis destructive to our ability to appreciate music and art. Interestingly, analysis within the framework of music and art proper seems to be acceptable. Fortunately, analysis leads to a richer synthesis. I hope that the reader will discover that analysis within the framework of Physics enriches our experience and need not be destructive either.

In order to analyze sound and light, we must learn how to characterize sound and light. The sound of music is by far the easier of the two because it is characterized by a series of events in time. The sound that strikes our ears can be represented simply by a graph. We see in Fig. 1.7a a graph of the wave of a short piece of music, 2:38 min in duration, composed by the Italian Renaissance composer **Salomone de Rossi** for five voices. It is difficult to see the details of the graph because of the extreme compression. To appreciate the content, Fig. 1.7b provides us with a magnification of an excerpt lasting about one-tenth of a second.[4]

Such a graph might seem to trivialize human experience. Alternatively, one might be amazed at how such a simple graph can fully represent something so powerful! The human mind is wonderful.

[4]The graph represents the output of a single loudspeaker; for stereophonic sound, we would simply need two such graphs.

Art is by far more complex and varied. Typically, it is two- or three-dimensional (2D and 3D) and is static in time. Modern art includes dynamic visual works too. In this book, our study of the place of Physics as it relates to art will be extremely limited. We will study the nature of light and its relation to our perception of color. We will not go much beyond 2D images, with a focus on simple patches of uniform color and interactions between neighboring patches. A 2D image on a plane can be characterized by specifying the color at each point on the plane. The color can be specified in terms of what is referred to as the **spectral intensity**. We will learn that the spectral intensity gives more information than is necessary. A simpler though **incomplete characterization** of color makes use of a **three-primary representation**. One must specify the intensity of each of three primaries at each point on the image.

Will this text enable you to account for the esthetic pleasures of music and art? Perhaps, only to a small degree. Is there in fact such a connection? I certainly believe so, though I do not expect such a connection to be fully clarified in my lifetime. Perhaps, it never will be. However, I will be satisfied if our study of the Physics of Music and Color reveals new vistas of sound and light, so that your world experience of music and color will be greatly enriched.

Chapter 2
The Vibrating String

The subject of this text is music and color. Music is produced by musical instruments, some occurring naturally – such as the songs of birds – and others produced by man-made instruments – such as stringed instruments, wind instruments, and the percussive instruments of drum sets. Color is produced by sources of light such as natural sunlight and by man-made sources such as the floodlights for a stage.

Essentially, music and color are *subjective* manifestations of the corresponding *objective* physical phenomena – sound and light, respectively. Both sound and light are examples of **wave phenomena**. If we can understand the nature of waves along with the multitude of phenomena associated with waves, we will become more aware of much of the richness of our human experiences with sound and light and hence music and color.

There are many types of waves. We can observe the wave nature of some types of waves with our own eyes – such as waves along a vibrating string or waves on the surface of the ocean. On the other hand, the wave nature of many important waves are invisible; examples are sound waves and light waves. It is therefore reasonable for us to begin our study with waves along a string – the fundamental component of all stringed musical instruments.

2.1 Waves Along a Stretched String

Suppose that we have a long string and stretch it. The string is depicted as the uppermost solid line in Fig. 2.1. The **tension** in the string keeps the string straight. Next, we disturb the string by pulling the string upward a bit at a particular point along the string. The shape of the disturbance is a small triangle. What will happen next? The disturbance will move along the string as shown in the figure at one milli-second (1 ms) intervals: We set the time t equal to 1, 2, 3, 4, and 5 ms. Each of the vertical dotted lines marks a position along the string at a sequence of one-meter (1 m) intervals. We note that after each 1-s interval, the disturbance progresses a distance

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_2,

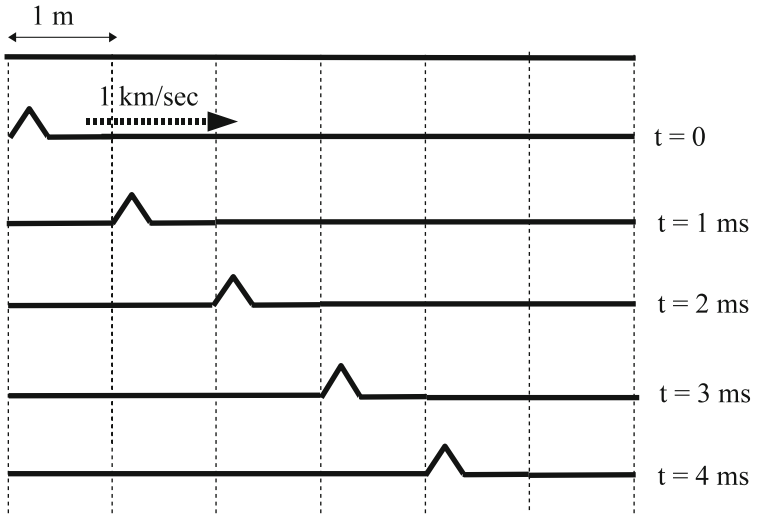


Fig. 2.1 A pulse traveling down the length of the stretched string

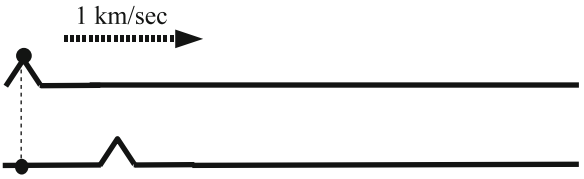


Fig. 2.2 The motion of a point – marked by a dot along the string

1 m to the right. Thus, the disturbance moves at a speed of 1 m/ms. This value is equivalent to 1,000 m/s. Note that this speed is quite large; in common units it is one kilometer per second (1 km/s), which is equivalent to 0.6 miles/s. Nevertheless, this value is close to the speed of a disturbance moving along a typical violin string.

A localized disturbance of this sort is called a **pulse** and is a simple example of **wave propagation**. The speed of the pulse is called the **wave velocity**. Later on in the chapter, we will investigate what determines the wave velocity for a stretched string.

We can easily show that the string itself does not move at a speed of 1 km/s, or 1,000 m/s, nor does the string itself move to the right. In order to see this, suppose we focus our attention on a single point along the string, say the point marked with a dot, shown in Fig. 2.2. We note that while the pulse is moving to the right, this point along the string has moved *downward*! We say that the wave is **transverse**, here meaning *perpendicular*. Suppose next that the height of the pulse is one millimeter (1 mm) (not drawn to scale above). Then the average speed of this point is 1 mm/s, a value much less than the wave velocity.

How can we account for the motion of the pulse? Think of the old familiar "telephone game," wherein we have a string of people. The first person whispers

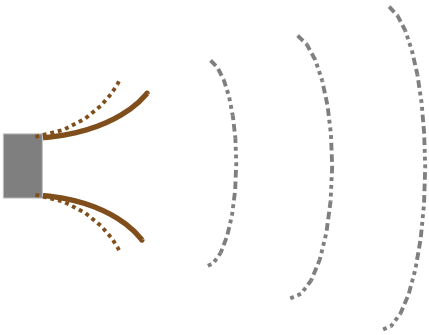


Fig. 2.3 Schematic of a loudspeaker

a message to the second person. The second person whispers the perceived message on to the third person, and so on. The last person announces the message received and the first person reveals the original message. One hopes that the message will not be garbled!

In the case of the string, the initial material of the pulse along the string pulls upward on the neighboring string material. The neighboring material pulls upward on its neighboring material, and so on, leading to the propagation of the pulse.

How does this description relate to other types of waves? The most important wave in the context of music is of course a sound wave – the focus of Chapter 3, THE VIBRATING AIR COLUMN. Sound waves can propagate through a variety of media – such as air or water or a solid. Let us try to produce such a wave: Imagine what would happen if you were to move your hand forward suddenly. You would compress the air immediately in front of your hand. That compressed region of air would compress the air immediately in front of it. This process will continue as in the case of a pulse propagating along a stretched string. You will have produced a **sound pulse**. The wave is said to be **longitudinal**, meaning that the motion of the air is along the same direction as the direction of propagation of the disturbance. Unfortunately, you cannot move your hands fast enough to hear this pulse.

If you were to be able to move your hand forward and backward at a rate that exceeds 20 times per second, you would in fact produce an audible sound. Your hand would be acting essentially like a loudspeaker, as shown in Fig. 2.3. At the left, we see the gray cone of the loudspeaker moving forward and backward. There are two positions shown – one as a pair of solid brown curves, the other as a pair of dotted brown curves. The sequence of three dotted pseudo-vertical curves represent the sound wave traveling through the air.

2.2 A Finite String Can Generate Music!

Consider now a guitar string strung on a guitar. The string considered in the previous section was assumed to be infinite; this string is finite with ends that are held fixed. See the uppermost line segment in Fig. 2.4, where we represent a string of length

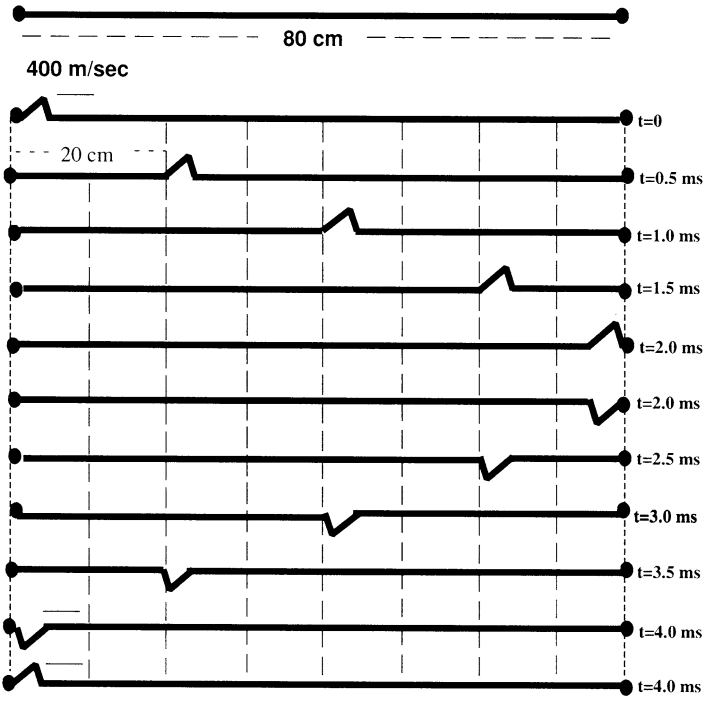


Fig. 2.4 A pulse traveling back and forth along a string with fixed ends

$l = 80\,\text{cm}$. We will assume that the wave velocity is $v = 400\,\text{m/s}$. Imagine what would happen to a pulse that is sent down the string, starting at one end, as in Fig. 2.4. The width of the pulse is exaggerated – the width is understood to be much less than a centimeter, so that it can be ignored in the calculations below.

Let us determine how long it will take for the pulse to reach the opposite end. We will use the relation

$$\text{Speed} = \frac{\text{Distance}}{\text{Time}} \quad \text{OR} \quad \text{Time} = \frac{\text{Distance}}{\text{Speed}}. \tag{2.1}$$

We will carry out the calculation using symbols – t for time, l for distance, and v for speed. We must be careful when we are given quantities that use different units for a given quantity. This issue is exemplified by the current situation, where we have a distance of 80 cm and a speed of 400 m/s. Thus both the centimeter and the meter are used for the dimension of length. In order to use (2.1), we must use the same unit of length for both quantities. We will choose to use the meter for both, recognizing that we could also use the centimeter for both without any error.

Since $1\,\text{m} = 100\,\text{cm}$, the distance is $0.80\,\text{m}$. We then obtain

$$t = \frac{l}{v} = \frac{0.80\,\text{m}}{400\,\text{m/s}} = 0.0020\,\text{s} = 2.0\,\text{ms}. \tag{2.2}$$

We note in the figure that the pulse reaches the opposite end in 2.0 ms. The pulse is then reflected back to the left along the string.

Look closely at the shape of the reflected pulse. Notice that the shape of the pulse is "reversed" in two ways: First, the original pulse approached pointing upward; the reflected pulse is pointing downward. Second, notice that the original pulse is steeper on the right side compared to the left side; on the other hand, the reflected pulse is steeper on the left side.

What will happen next? The pulse will reach the left end and be reflected back to the right. The same reversals as above will take place once again. The pulse is reversed from pointing downward to pointing upward; the steeper edge is reversed from being steeper on the left side to being steeper edge on the right side. The end result is a pulse that is exactly the same as the original pulse! The time for the round trip will be $2 \times 2.0\,\text{ms} = 4.0\,\text{ms}$.

Such a round trip is generally referred to as a **cycle**. Ultimately, the pulse will move back and forth, with one round trip every 4.0 ms. This time interval is called the **period**, with the symbol T. Thus,

$$T = \frac{2l}{v} = \frac{2(0.80)}{400} = 4 \times 10^{-3}\,\text{s} = 4\,\text{ms}. \tag{2.3}$$

The number of cycles per unit time is called the **frequency**, with the symbol f. In the current case, we have

$$f = \text{one cycle per 4 ms} = \frac{1\ \text{cycle}}{4 \times 10^{-3}\,\text{s}} = 250\ \text{cycles per second} \equiv 250\,\text{cps}. \tag{2.4}$$

An alternative term for the cycle per second as a unit of frequency is the **Hertz**,[1] which is abbreviated as Hz. Thus, **one cycle per second** $= 1\,\text{cps} = 1$ **Hertz** $= 1\,\text{Hz}$.

Note that the frequency and the period are inverses of each other:

$$f = \frac{1}{T}. \tag{2.5}$$

In the above case, $250\,\text{Hz} = 1/(4\,\text{ms})$.

One should note that there are many ways that the string could be excited. The most important example for a guitar is the **pluck**, which is shown in Fig. 2.5. The pluck is produced by pulling the string aside at one point and then releasing it from the rest. The figure shows the subsequent motion of the string.

We note that the time for a full cycle, the period T, is again 4 ms. The corresponding frequency is 250 Hz.

[1] Named after Heinrich Hertz (1857–1894). Hertz was a great physicist who first demonstrated the existence of electromagnetic waves, which will be discussed later in this book.

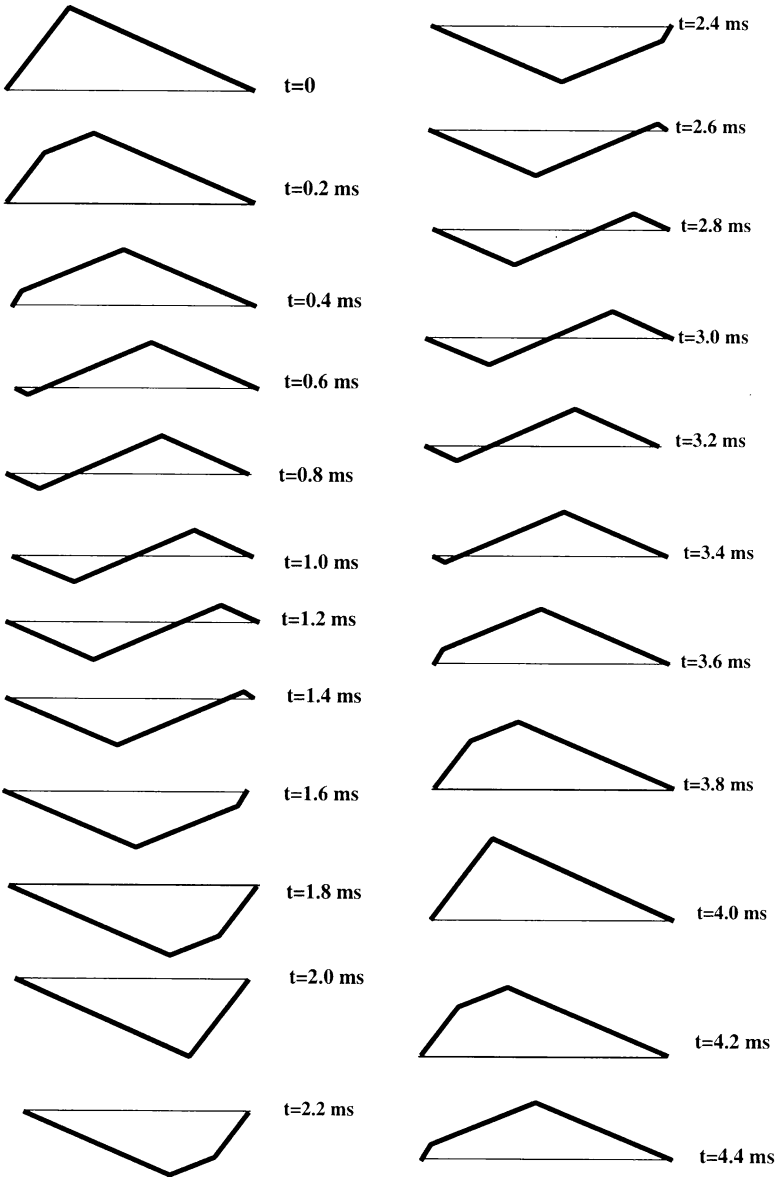


Fig. 2.5 The progressive wave along a plucked string

2.3 Pitch, Loudness, and Timbre

If you pluck a string, a sound is produced. You can identify several attributes of that sound. There is a definite **pitch**. Pitch designates the musical note to which the string is tuned. For example, the so-called *G* string of the violin (which is tuned to

the G below middle C on the piano) produces the pitch G. If you loosen the string, by turning the tuning peg, you will immediately notice that the pitch will change – it will become lower. If you tighten the string, the pitch will become higher.

A second attribute of the sound is its **loudness** . By giving the string a bigger pull when you pluck it, you can produce a louder sound. Furthermore, the loudness decreases after the initial pluck, until the sound is inaudible.

The third attribute is what we identify with the quality of the sound produced by the particular instrument – the **timbre**. Timbre is one of the factors that enables you to distinguish the G played on the violin from an equally loud G played on a piano or a trumpet or any other instrument. You can vary the timbre of the plucked string itself by changing the point at which you pluck as follows: first pluck the string near its center and listen carefully to the quality of the sound. Then pluck the string very near one end, trying to produce the same loudness. The pitch will be the same but there will be a slightly different timbre to the sound. When plucked near the end, the resulting sound has a slight high-pitched ring or "twang," which is not present in the sound produced by plucking near its center. Similarly, if a narrow pulse is cycling back and forth along the string, a sound will be produced having the same pitch but different timbre. A bowed string produces a wave that moves back and forth the length of the string with a different characteristic shape; yet again, we will hear a sound with the same pitch.[2]

We have not been very precise, at this point in defining pitch, loudness, and timbre. To be more precise, you must first understand what physical phenomena give rise to the "perceptual" qualities we have discussed.

2.4 The Relation Between Frequency and Pitch

Recall that in discussing pitch, we said that if the string being plucked was loosened, the pitch would become lower. Imagine loosening the string of Fig. 2.5 and then plucking it, so that at the moment of release it has exactly the same shape as that in the first frame of the figure. However, it will take more time to complete one cycle. The period will increase, with a consequent decrease in the number of oscillations per second; that is, the frequency will decrease. This is in agreement with (2.5).

Let us suppose that the string is loosened just enough to increase the period to 5 ms. Then the new frequency is $f = 1/0.005\,\text{second per cycle} = 200\,\text{cps} = 200\,\text{Hz}$.

How much loosening does this change require? To answer this question, we need a quantitative measure of the "tautness" or tension of the string and how that tension is related to frequency. We will return to this question in Sect. 2.8. What we want you to consider at the moment is the qualitative result of this little experiment.

[2]The sound of the violin is strongly affected by the other physical components of the instruments, along with their respective vibrations.

Loosening the string decreases the frequency of the oscillations, as it lowers the pitch. Correspondingly, tightening the string increases the frequency and raises the pitch. So there is a relation between the physical quantity, frequency, and the psychological attribute, pitch. This relationship is the basis for the tuning of musical instruments. In Chap. 10, we will note that the loudness of a note also affects the sense of pitch.

The strings of a piano are tuned to a definite set of frequencies. First, one sets the A above the middle key on a piano – referred to as "middle-C" – at a frequency of 440 Hz. The "middle C" on a piano is set at a frequency of approximately 262 Hz. The lowest C is set correspondingly to a frequency of approximately 33 Hz, and so on.

We see that one way to produce different frequencies is to vary the tension. Are there other ways? If we combine (2.3) and (2.5), we obtain the relation

$$f = \frac{v}{2\ell}. \tag{2.6}$$

We will see later in the chapter that an increase in the tension produces an increase in the wave velocity. As a consequence, according to (2.6) the frequency will increase and so will the pitch. We will also see later that changing the nature of the string itself will change the wave velocity. Finally, we see that decreasing the length of the string will increase the frequency. All three factors are used to produce the huge range of frequencies of a piano – from 27.5 Hz to ∼4,186 Hz.[3]

Various stringed instruments are tuned accordingly. For example, in order that the A string on the violin be in tune with the corresponding A string on the piano, their frequencies should be equal.

Why the particular frequency of 440 Hz is chosen for the "A" is a matter of history. In fact, this frequency has been rising steadily over the past 200 or more years, so much so that in Bach's time it is believed to have been about 415 Hz. Why the notes of the Western scale have the frequencies to which we have just alluded will be the subject of Chapter 11, TUNING, INTONATION, AND TEMPERAMENT: CHOOSING FREQUENCIES FOR MUSICAL NOTES. The development of scales is a fascinating story of the interdependence of scientific understanding and esthetics.

2.5 The Wave Motion of a Stretched Rope

It is difficult to study the motion of the strings of musical instruments without special equipment because the wave velocities and the frequencies are very large. It is possible to check the relation (2.6) by performing a simple, but illustrative experiment. Get a long piece (2 or 3 m) of heavy rope or clothesline or a long tightly

[3]The lower frequency is precisely four octaves (a factor of $2^4 = 16$ below 440 Hz), while the latter frequency corresponds to tuning according to equal temperament. (See Chap. 11)

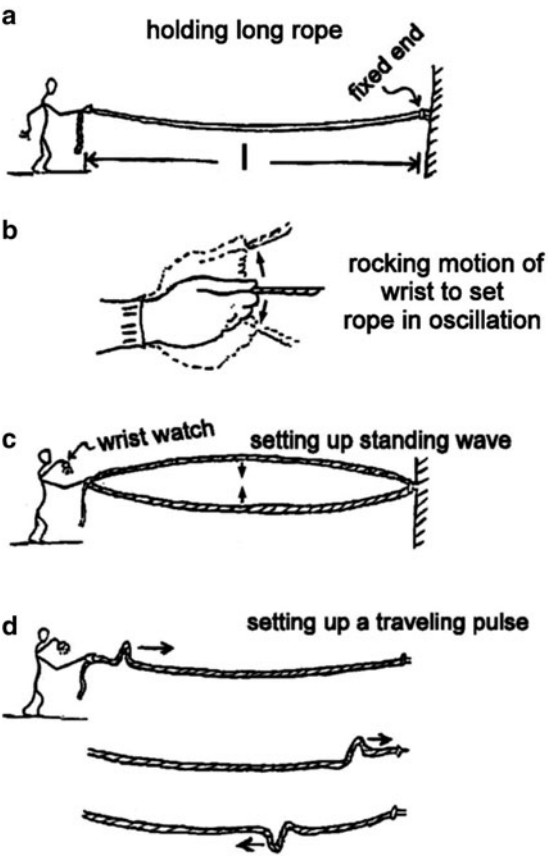


Fig. 2.6 Exciting a long rope. (drawing by Gary Goldstein)

wound spring (as used to close screen doors). Secure one end to a fixed point – say a doorknob on a closed door. Pull the free end so that the rope is stretched loosely to its full extent, as shown in Fig. 2.6a. Estimate the length, ℓ, of the stretched rope.

You are going to set up wave motion of the rope by shaking the held end up and down while using your wrist as a pivot, as shown in Fig. 2.6b. By shaking very slowly at first and gradually increasing the rate of shaking, you will soon reach a rate that sets up a wave of the form shown in Fig. 2.6c. The whole rope will be oscillating up and down at that rate. Notice that once you set up that motion, it is easy to maintain the motion. It is as if the system has "locked in" to that mode of oscillation.

While maintaining the motion of Fig. 2.6c, use the seconds hand on your wrist watch to determine the period. (You might have a friend to assist you.) This can be done easily by counting, say ten cycles and observing how many seconds have elapsed. Remember that a cycle is completed when the rope has returned to some initial configuration, so whenever it reaches the lowest point in its motion it has

completed a cycle. If, for example, the rope completes ten cycles in 8 s, the period would be $8/10$ s. The frequency would be $10/8 = 5/4$ Hz.

Next, let the rope return to rest. It is important not to vary the tension, so do not change your position. Now you are going to set up a disturbance in the rope of the form shown in Fig. 2.6d. This is accomplished by very quickly jerking your hand up and down while quickly returning to the starting position. It is best to keep your hand as rigid as possible. Observe what happens. The short disturbance or **pulse** moves rapidly to the end of the rope, is reflected, and returns to your hand upside down. If your hand remains rigid the pulse will reflect at your hand, turn right side up and move to the far end again. The pulse might make many round trips before it disappears. You have set up a **traveling wave**. (A pulse travels across the string.)

Note that any particular segment of rope material moves up and down, while the wave pattern, the pulse in this case, moves down the length of the rope. These two directions are perpendicular to each other. The waves are **transverse**.

Now time the pulse by measuring the time required for the pulse to complete several round trips. For example, if the pulse makes five round trips in 4 s, then the time for a single circuit would be $4/5$ s or 0.8 s. If you are careful, you will find that the time required for one round trip is the same as the period of oscillatory wave motion that you determined before, for the standing wave.

Measuring the length of the stretched rope will then enable you to determine the velocity of propagation for the traveling wave. In our example, the round trip time and the period were 0.8 s. If the rope were 2 m long, a round trip would be 4 m and the velocity of propagation would be $4\,\text{m}/0.8\,\text{s} = 5\,\text{m/s}$. Determine the velocity for your rope, using (2.3).

2.6 Modes of Vibration and Harmonics

One might ask whether the string can be excited so as to produce a vibration that does not have a frequency of 250 Hz. The answer is yes. In the course of demonstrating this fact, we will describe what are referred to as the modes of vibration of the string.

Now that you have become familiar with working the rope you can learn how to excite its modes of vibration. Start by exciting the same standing wave that you did before (Fig. 2.6c). Count the cycles rhythmically while the rope is oscillating. That is, say the numbers out loud every time the rope reaches bottom – "one-two-three-four-one-two…". Now start shaking your hand at twice the original tempo. You will have doubled the rate of oscillations and hence the frequency. The rope will "lock into" a different mode of oscillation. It will now appear as in Fig. 2.7a. The period of this oscillation is one-half the period of the preceding oscillation, wherein a pulse is traveling back and forth along the rope, as in Fig. 2.6.

For the sake of identification, we call the mode of Fig. 2.6c the **fundamental mode** or the **first harmonic** of the string. The mode of oscillation you are now producing is called the **second harmonic** (Fig. 2.7a). While the rope is oscillating

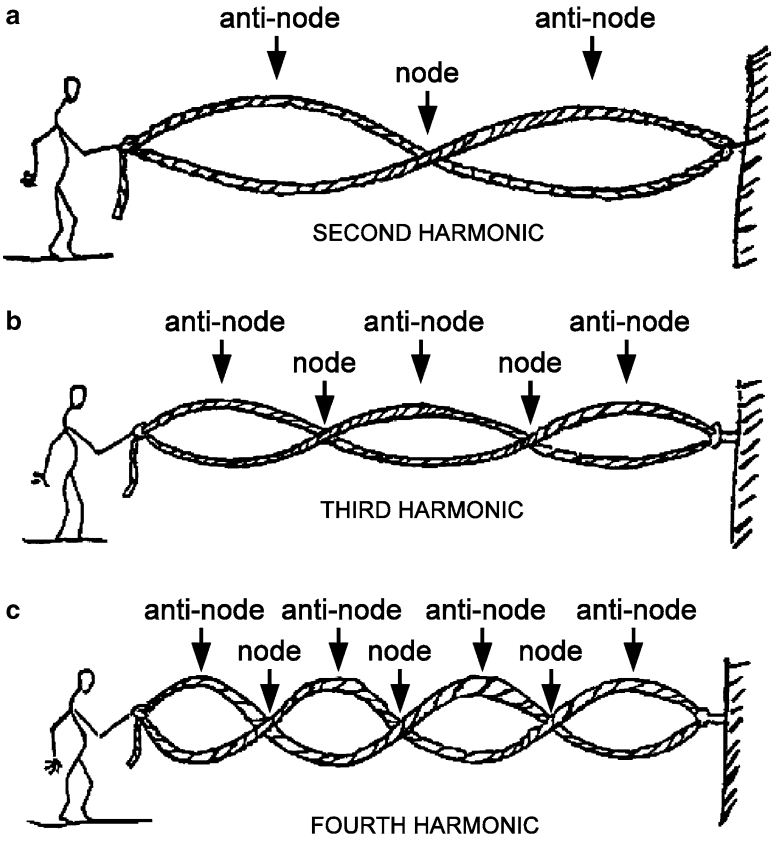


Fig. 2.7 Higher harmonics of the vibrating string (drawing by Gary Goldstein)

in this mode, notice that near the midpoint the rope is hardly moving at all. This point at which no motion occurs is called a **node**. For the second harmonic, there is one node between the end points, whereas the fundamental mode (Fig. 2.6c) had no nodes between the end points.

Observe also that there are two points along the rope which achieve the greatest displacement from equilibrium (either above or below), one at about $1/4$ the distance from your hand, the other at $3/4$ the distance. These points along the rope at which the maximum displacement occurs are called **antinodes**. The second harmonic has two antinodes, whereas the fundamental mode has one antinode at the midpoint of the rope (see Fig. 2.6 again).

Now, by shaking your hand at triple the rate for the fundamental mode you can excite the mode shown in Fig. 2.7b, the third harmonic. This is somewhat harder to excite than the preceding mode, but once you get near the right rate of shaking, the rope will respond very strongly and will "lock in" to that mode. The third harmonic has three times the frequency of the fundamental. You will observe that there are

Table 2.1 Harmonics

Mode	Frequency	No of nodes	No of antinodes
Fundamental = 1st harmonic	f_1	0	1
2nd harmonic	$f_2 = 2f_1$	1	2
3rd harmonic	$f_3 = 3f_1$	2	3
4th harmonic	$f_4 = 4f_1$	3	4
5th harmonic	$f_5 = 5f_1$	4	5
6th harmonic	$f_6 = 6f_1$	5	6
nth harmonic	$f_n = nf_1$	$n-1$	n

two nodes in this mode – one at $1/3$ the distance to the fixed end, the other at $2/3$ that distance. There are three antinodes.

You should now see the pattern. By exciting the fourth harmonic (Fig. 2.7c), which has a frequency four times the fundamental frequency, you will produce a mode having three nodes and four antinodes. (It is appreciably harder to excite this mode; the higher modes are progressively more difficult.) The fifth harmonic would have a frequency five times the fundamental frequency, and the wave pattern would have four nodes and five antinodes. We summarize this information in Table 2.1.

The frequencies of the harmonics are written as multiples of the fundamental frequency f_1. We have included a general mode, the nth harmonic, where n symbolizes any integer (1, 2, 3, 4, ...). Letting $n = 7$, for example, tells you that the 7th harmonic has frequency $7f_1$, $(7-1)$ = six nodes and seven antinodes.

From all of the preceding you now see that the rope, or a stretched string, has many different modes of vibration. These modes of vibration have frequencies which are integral multiples of the fundamental frequency – the modes are harmonic. Then the periods for each of the modes will be different from one another. Recall, however, that the time required for a traveling pulse to make a round trip (Fig. 2.6d) was equal to the period of oscillation of the rope in the fundamental mode (Fig. 2.6c). Therefore, the relation between wave velocity, length, and frequency (2.6) should be rewritten to show explicitly that the fundamental frequency is involved.

We have

$$f_1 = \frac{v}{2\ell}. \tag{2.7}$$

For the other harmonics, the frequencies are multiples of the fundamental frequency, so

$$\begin{aligned} f_2 &= 2f_1 = 2 \times \frac{v}{2\ell} \\ f_3 &= 3f_1 = 3 \times \frac{v}{2\ell} \\ f_4 &= 4f_1 = 4 \times \frac{v}{2\ell} \end{aligned} \tag{2.8}$$

and so on. This series is summarized by writing the frequency of nth harmonic, f_n, as

$$f_n = n\ f_1 = n \times \frac{v}{2\ell}. \tag{2.9}$$

The fact that the rope or stretched string can be set into oscillation in many different modes will be of continual importance. It forms the basis for much of the subsequent discussion.

Notice that the wave patterns for these modes do not move either to the right or to the left. We refer to such a wave as a **standing wave**. In contrast, the wave described initially in this chapter that moves along an endless string is called a **traveling wave**. We will discuss such waves more fully in the next section.

We close this section by introducing other widely used terms – the **overtone** and the **partial**. By definition, the first overtone is the second harmonic; the second overtone is the third harmonic; and so on. The term **partial** is used to refer to one of the modes of a musical instrument **whether or not** the frequencies form a harmonic series. An example is the sound of a gong, whose mode frequencies do not form a harmonic series.

2.7 The Sine Wave

The shape of the pattern along a string that is vibrating in one of its modes is very specific – being the curve produced by plotting the trigonometric **sine function**. In addition, if we plot the displacement of any point along the string vs. time, we will obtain a graph of the sine function. In fact, of all periodic curves in nature, the sine curve is very unique in its physical ramifications, as we will see many times in the course of our study of sound and light and therefore of music and color. Thus, we now turn to an examination of the sine curve.

You probably have paid attention to how various radio stations are identified. For example, a popular radio station for classical music in the Boston area is WCRB 102.5FM. The number 102.5 stands for a frequency of 102,500,000 Hz. (The letters 'FM' stand for 'frequency modulation', which is a special means of transmitting information using waves; it will be discussed later in the text.) Or, as another example, you might have heard that most current symphonic orchestras are tuned to a frequency of 440 cycles per second. What these numbers fully represent is the subject of this section.

Let us begin by returning to the long stretched string. Suppose that you were to take hold of the string and move it up and down repeatedly at a constant rate in time. If the pattern of your motion is repeated again and again, we say that the pattern is periodic. As an example, let us display the pattern of motion for the most important such motion; it is called a **sine wave** pattern. See Fig. 2.8.

The graph displays the displacement of the hand as it varies in time. Note that there is a pattern that extends over a 1-s interval. It is repeated three times over the entire 3 s interval. This interval is called the **period** or the motion. The maximum

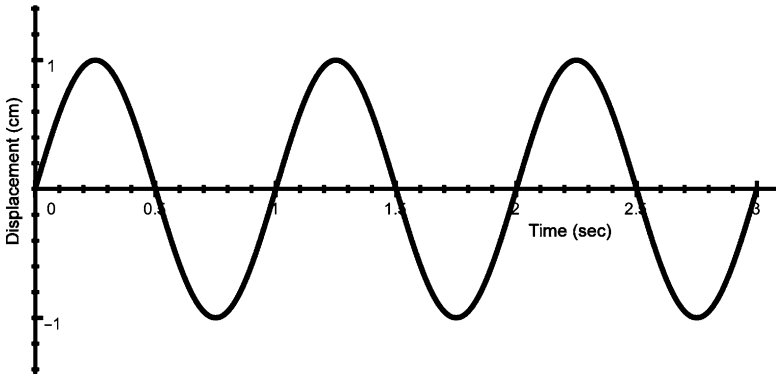


Fig. 2.8 Sine wave of displacement vs. time

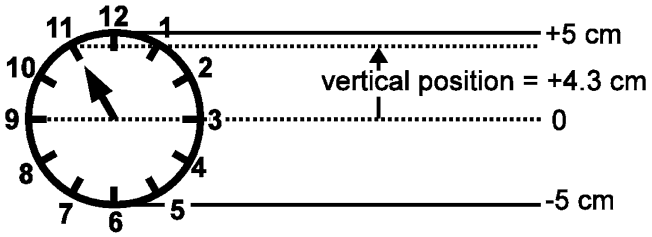


Fig. 2.9 Sweeping hand of a clock defining a sine wave

displacement is 1 cm and is called the **amplitude** of the motion. The pattern is **sinusoidal** and represents the **sine function** of trigonometry. Let us review the nature of the sine function.

You might recall that the sine of an angle is the ratio of the "side opposite" to the hypotenuse of a right triangle. Thus,

$$\sin\theta = \frac{b}{c}. \tag{2.10}$$

We can produce a graph of the sine function by a simple method involving the constant circular motion of the seconds hand of a clock. The seconds hand sweeps around, making a full circle every 60 s. Let us measure the vertical position of the tip of the hand as it sweeps around. We do this by first drawing a base line across the face passing through the center and the 3 and 9 o'clock marks, as shown in Fig. 2.9. Suppose the hand extends 5 cm from the center. Then the *vertical position* of the tip, relative to the base line, will vary from the lowest point (at the 6 o'clock mark) of -5 cm, to the highest point (at the 12 o'clock mark) of $+5$ cm. When the hand points at 11 o'clock, for example, as shown in Fig. 2.9, the vertical position will be $+4.3$ cm.

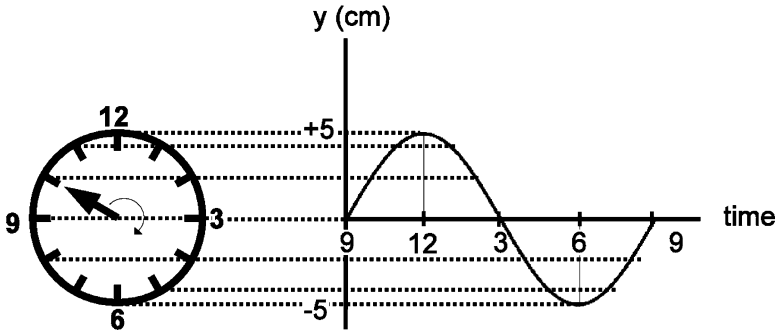


Fig. 2.10 Clock defining one cycle of a sine wave

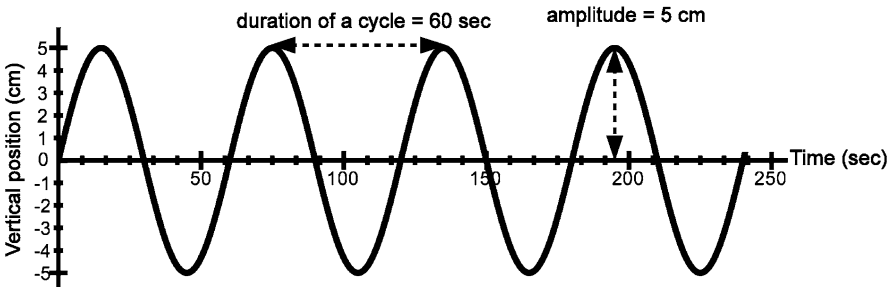


Fig. 2.11 Vertical position of clock hand vs. time elapsed

Now we will plot the vertical position as the hand sweeps around, starting at the 9 o'clock position when the vertical position is 5.0 cm. Five seconds later, the hand will be at 10 o'clock and the vertical position will be +2.5 cm. Then in 5 s more, the hand will be at 11 o'clock and the vertical position will be +4.3 cm, and so on. The procedure of plotting the vertical position as a function of time elapsed is illustrated in Fig. 2.10, for the first 60 s; we obtain one cycle of the sine wave. Continuing this plotting gives the curve in Fig. 2.11.

The form of the curve repeats exactly every 60 s, when the hand has returned to its initial position. If we continued plotting the vertical position of the hand indefinitely, the curve would continue on repeating itself indefinitely (or until the clock stopped). The full curve is the sine function. It is **periodic** in that it repeats itself indefinitely. Note that the angle changes steadily, going through a full circle of 360° in 60 s. Thus, the rate of change of the angle is 360°/min, or 6°/s. Then, the value of the sine function after 5 s will be $\sin 30° = 0.5$. The result is a vertical position equal to $0.5(5.0) = 2.5$ cm.

Although we have obtained this curve by a particular procedure, its significance is far more general. Being a graphical representation of a function, it represents a mathematical prescription: Given some numerical value of the variable, the sine of

that variable has a definite numerical value. The variable may represent a time (as in the example we have used), or it may represent a position along a string, or it may represent an angle. What is important is the shape of the curve and its periodicity.

The particular sine function of Fig. 2.10 can be characterized by two numbers. The first of these is the maximum height of the curve, called the **amplitude**, which is 5 cm for this case. The second is the length of one cycle, which is a time interval of 60 s for this example. When the variable is time, the length of one cycle, its duration, is the **period** of the motion. We will soon consider examples of sine functions representing some vertical position as a function of distance rather than time. In that circumstance the length of one cycle will be a distance and is called the **wavelength**.

2.8 The Simple Harmonic Oscillator

While the sine curve is central to the behavior of the modes of a vibrating string, it shows itself in a simpler way in the behavior of a **simple harmonic oscillator** (**SHO** for short). The SHO is a system that is fundamental for understanding all vibrating systems and therefore deserves significant attention. It consists of a spring having negligible mass that is attached at one end to some fixed support and a rigid object that has the essential mass of the system – referred to as the **mass** of the SHO – at the other end. See Fig. 2.12.

When isolated, an SHO will come to rest at its **equilibrium state**, in which the spring is neither stretched nor compressed. To displace the mass from its equilibrium position, a force must be applied. That force F is proportional to the displacement y from the equilibrium position of the mass, as shown in Fig. 2.12b, in which a **downward** displacement corresponds to positive y, while an **upward** displacement corresponds to a **negative** y. Because the graph of y vs F is a straight line, we say that the relation between y and F is **linear**. We have

$$\text{Displacement} \quad \propto \quad \text{Force}. \tag{2.11}$$

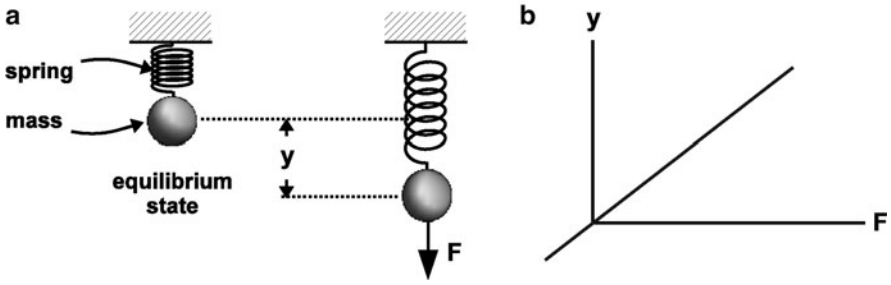


Fig. 2.12 The Simple Harmonic Oscillator

Mathematically we write

$$y = \frac{1}{k}F = \frac{F}{k}, \tag{2.12}$$

where the constant k is known as the **spring constant** or the force constant. The relation is known as **(Robert) Hooke's Law**.

If the force is measured in lbs and the displacement in inches, the spring constant is expressed in lbs-per-in, which will be written as "lbs/in." In this text, the force will often be expressed in Newtons (abbreviated as N) (one Newton is about 4.5 lbs) and the displacement will be expressed in meters. Then the spring constant will be expressed in Newtons per meter, or N/m.

Sample Problem 2-1

Suppose it takes a force of 5 N to stretch a given spring 2 m. Find the spring constant.

Solution

The spring constant is given by

$$k = \frac{F}{y} = \frac{5}{2} = 2.5\,\text{N/m}. \tag{2.13}$$

Note that if that same spring is stretched by a force equal to 10 N, the displacement will be

$$y = \frac{F}{k} = \frac{10}{2.5} = 4\,\text{m}. \tag{2.14}$$

Doubling the force leads to a doubling of the displacement.

If the mass is pulled from its equilibrium position as in Fig. 2.12 and released, it will oscillate in time at a certain frequency.

The *linear* relation between y and F is **unique** in leading to two characteristics:

1. A **sinusoidal** displacement in time, as shown in Fig. 2.13
2. A frequency that is **independent of the amplitude of oscillation**

Let us imagine suspending a mass to a spring and letting it oscillate. We see in Fig. 2.13 that the displacement of the mass exhibits a sine wave pattern. Its amplitude is 2 cm. The period is 2 s. The corresponding frequency is $f = 1/T = 1/2 = 0.5$ Hz.

In fact, the two characteristics of oscillatory motion automatically imply a linear relation. Real springs do not obey **Hooke**'s Law precisely, as shown in Fig. 2.14. However, they do so approximately for small enough displacements.

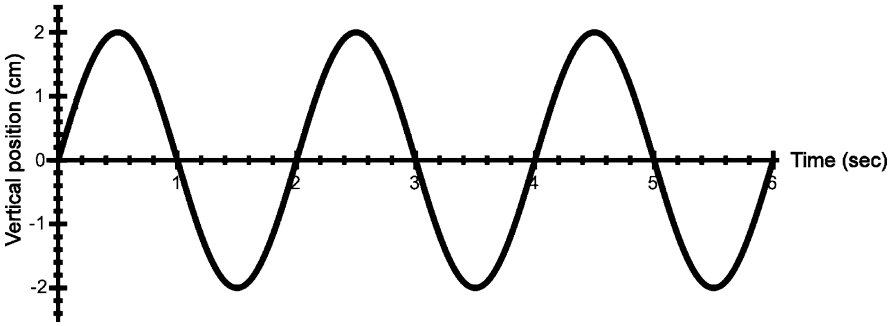


Fig. 2.13 Displacement of an SHO vs. time

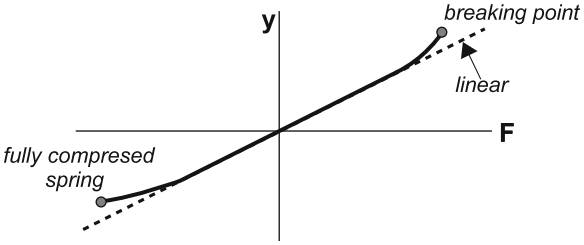


Fig. 2.14 Displacement vs. force for a real spring

2.8.1 The Vibration Frequency of a Simple Harmonic Oscillator

The spring constant and the mass of an SHO determine its vibration frequency. We will see later that the fundamental frequency of a vibrating string is proportional to the square root of the ratio of a restoring force to a mass. This qualitative relationship holds for an SHO too. It can be shown that the frequency of vibration of the SHO is given by[4]

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \qquad \text{Frequency of SHO.} \tag{2.15}$$

The spring constant reflects the restoring force.

In using this formula, we are not free to express the units of k and mass independently. The choice of units must be consistent. Thus, if we express the spring constant in N/m, the mass must be expressed in kilograms (abbreviated as "kg"). Then the frequency we obtain is expressed in Hz.

[4]This expression for the frequency as well as the fact that the motion is sinusoidal can be derived rigourously mathematically using a combination of Hooke's Law $F = kx$ and Newton's Second Law of Motion $F = ma$, where a is the acceleration of the mass. Eliminating the force leads to a direct relation between the displacement and the acceleration: $x = ma/k$. This subject is discussed in Appendix E. In Appendix F, you can see how the sinusoidal motion evolves by using a technique of Numerical Integration.

Sample Problem 2-2

Suppose that an SHO has a spring constant equal to 25 N/m and a mass of 500 g. Find the frequency and period of vibration of the SHO.

Solution

We must first express the mass in kg: 500 g = 0.500 kg. Then

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{25}{0.500}} = 1.1\,\text{Hz}. \tag{2.16}$$

Correspondingly, the **period** of vibration is given by

$$T = \frac{1}{f} = 2\pi\sqrt{\frac{m}{k}} \qquad \text{Period of SHO.} \tag{2.17}$$

so that $T = 1/1.1 = 0.9\,\text{s}$.

We note that generally, the frequency increases if the spring constant increases and/or the mass decreases. However, if the spring constant is doubled or the mass is halved, the frequency is not doubled. Instead, it is increased by a factor of $\sqrt{2}$: Taking off from the previous numerical example, suppose that the spring constant is 50 N/m and the mass is 0.500 kg. A simple calculation leads to a frequency of 1.6 Hz, which equals $\sqrt{2}$ multiplied by the frequency of the previous example.

We can obtain an estimate of the amplitude of the velocity – the maximum speed – of an SHO over the course of one oscillation. It is on the order of the average speed. The latter is simply the total distance traveled by the mass in one cycle divided by the period. Thus, with A = amplitude

$$\text{Average speed} = \frac{4A}{T} = 4Af. \tag{2.18}$$

The actual velocity amplitude is $2\pi Af$, which is a bit greater, as it should be.

2.9 Traveling Sine Waves

The modes of vibration of a string of fixed length are such that there is a pattern that remains stationary except for oscillations in the overall amplitude. The pattern does not move to the right or the left. To the contrary, a wave that progresses in one or the other direction is referred to as a **traveling wave**. Here is a simple way to produce a traveling sine wave.

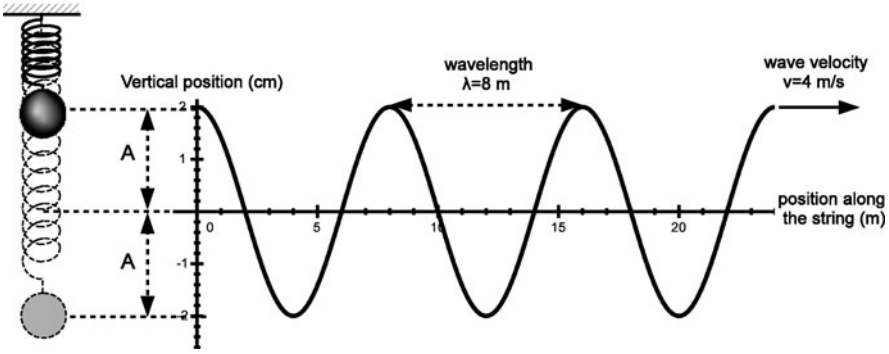


Fig. 2.15 Traveling wave produced by an SHO

Suppose that we attach the mass of an SHO to the end of a long stretched string. The mass is set into oscillation with an amplitude of 2 cm, as above. In Fig. 2.15, we see the mass at an instant when it has its maximum upward displacement of 2 cm. To its right, we see the sinusoidal pattern of the wave along the string. Note that this curve represents the actual material of the string at this instant in time. Furthermore, **whereas the displacement of the oscillator is sinusoidal in *time*, the pattern of the string is sinusoidal in *space*.**

We have assumed that the wave velocity is 4 m/s. As a consequence, during one cycle of oscillation lasting 2 s, the wave moves along the string a distance $x = vt = 4(2) = 8$ m. Each cycle that has its left end in contact with the mass is replaced by another such cycle. The three cycles along the string were produced by three cycles of oscillation of the mass. The period in space is called the **wavelength** and is here equal to 8 m.

We see that the sinusoidal wave in space is characterized by the following four parameters:

1. The **amplitude** A – here equal to 2 cm
2. The **wave velocity** v – here equal to 4 m/s
3. The **wavelength** λ – here equal to 8 m
4. The **period** T – here equal to 2 s

Clearly, we have a simple relation among the velocity, the wavelength, and the period:

$$v = \frac{\lambda}{T}. \tag{2.19}$$

The frequency and period are inverses of each other: $f = 1/T$. Therefore, we have the relation

$$\lambda f = v. \tag{2.20}$$

Notice that this equation can be rewritten as $\lambda = v/f$. As a consequence, for a given velocity, the wavelength decreases as the frequency increases. Alternatively, if the frequency is constant and the velocity decreases, the wavelength must decrease.

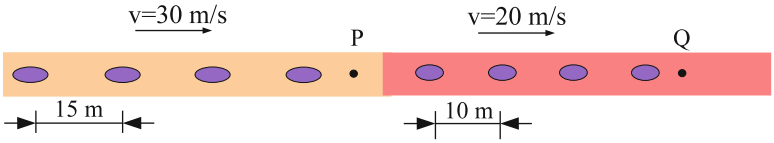


Fig. 2.16 Spacing vs. speed of cars in traffic

This latter result can be understood in terms of the following traffic situation. Suppose a line of cars is traveling along a one-lane road over a long time so that the traffic flow is stationary. The number of cars passing a given point must then be constant so that there is neither a pile up of cars someplace nor any buildup of empty space. Then, if the cars speed up, they must be further apart. Similarly, if the cars slow down, the space between the cars must decrease. A consequence of this decrease in space and the need for safety is that the cars usually slow down even more. Figure 2.16 illustrates how the spacing between neighboring cars decreases when the speed decreases. The rate at which cars pass point P is $30/15 = 2$ cars per second, which is equal to the rate at which cars pass point Q, that is, $20/10 = 2$ cars per second.

2.9.1 Applications

1. In the case of the standing wave, we recall that $\lambda = 2\ell$. Then, as we have already shown.
$$v = 2\ell f_1 \quad \text{or} \quad f_1 = \frac{v}{2\ell}.$$
2. The audible range of frequencies is from 20 to 20,000 Hz. In the case of a traveling sound wave in air, with $v = 340$ m/s, the corresponding range of wavelengths is
$$\max\lambda = 340/20 = 17\,\text{m} \quad \textbf{to} \quad \min\lambda = 340/20{,}000 = 0.017\,\text{m} = 1.7\,\text{cm}.$$
3. As we will see in Chapter 5, ELECTRICITY, MAGNETISM, AND ELECTROMAGNETIC WAVES, light is a visible electromagnetic wave having a range of frequencies from 4.0×10^{14} Hz to 7.0×10^{14} Hz. In the case of a light wave traveling in vacuum, the wave velocity is given the symbol c and is equal to 3.0×10^8 m/s. The corresponding range of wavelengths is
$$\max\lambda = \frac{3.0 \times 10^8}{4.0 \times 10^{14}} = 7.5 \times 10^{-7}\,\text{m} = 750\,\text{nm}$$

to

$$\min\lambda = \frac{3.0 \times 10^8}{7.0 \times 10^{14}} = 430\,\text{nm}.$$

2.10 Modes of Vibration: Spatial Structure

The modes of vibration of a stretched string are related to traveling sine waves. Later on in this chapter we will see that when two sine waves of the same wavelength head toward each other, they superpose to produce a standing wave with the same wavelength. The standing waves of the modes of vibration are portions of sine waves. An examination of their shapes reveals the relationship between these shapes and the corresponding wavelengths. In Fig. 2.17 we show the first six harmonics of the vibrating string. The patterns display the extreme shapes at two times, one-half cycle apart.

The second harmonic is a full cycle of a sine wave. Thus, the length of the string is equal to the wavelength. The fundamental (first harmonic) is a half cycle, so that the length is equal to one-half of a wavelength.

The wavelength for the fundamental is

$$\lambda_1 = 2\ell. \tag{2.21}$$

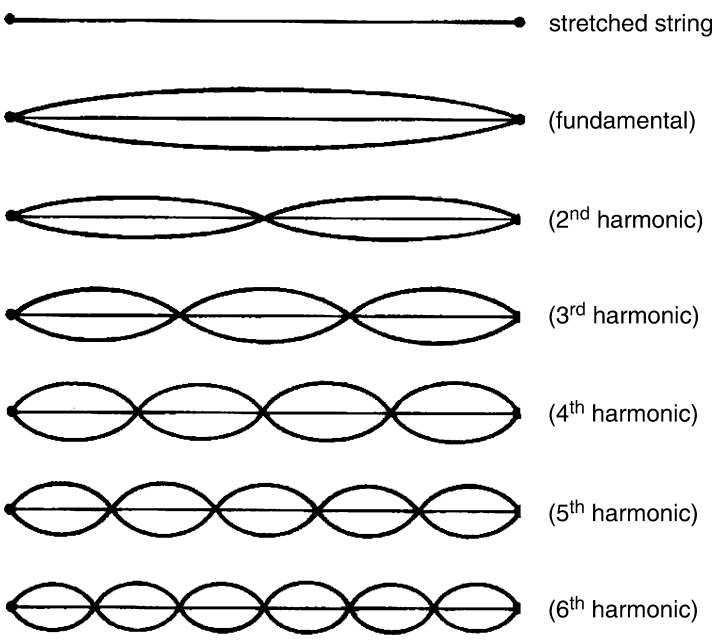


Fig. 2.17 Modes of vibration of the stretched string

For the second harmonic, the shape of the string encompasses a full cycle of a sine wave, so the second harmonic has a wavelength

$$\lambda_2 = \frac{\lambda_1}{2} = \ell. \tag{2.22}$$

In the third harmonic mode, the string has the form of one and one-half cycles of a sine wave. Thus

$$\lambda_3 = \frac{2}{3}\ell \quad \text{or} \quad \frac{3}{2}\lambda_3 = \ell. \tag{2.23}$$

Lastly, the fourth harmonic has the wavelength

$$\lambda_4 = \frac{1}{2}\ell. \tag{2.24}$$

This sequence of wavelengths can be rewritten in a way that allows us to generalize these examples:

$$\begin{aligned} \lambda_1 &= \frac{2\ell}{1} = 2\ell \\ \lambda_2 &= \frac{2\ell}{2} = \ell \\ \lambda_3 &= \frac{2\ell}{3} = \frac{2}{3}\ell \\ \lambda_4 &= \frac{2\ell}{4} = \frac{1}{2}\ell. \end{aligned} \tag{2.25}$$

Written in this way it is obvious that the fifth harmonic will have wavelength $\lambda = 2\ell/5$, and so on. For the nth harmonic, then, we will have the relation

$$\lambda_n = \frac{2}{n}\ell. \tag{2.26}$$

Now recall that according to (2.9) the frequency for the nth harmonic is given by $nv/2\ell$. The formula for the frequency is then

$$f_n = \frac{v}{\lambda_n}. \tag{2.27}$$

We can write this equation also as

$$\lambda_n f_n = v. \tag{2.28}$$

The equation becomes identical to (2.10), which applies to the two sine waves that when added together produce the standing wave.

We can appreciate this process if we realize that when we excite a standing wave by shaking one end of the string, we send a sine wave down the string. This wave is reflected off the other fixed end. Upon reflection, we have a second sine wave with the identical wavelength traveling back toward the hand. This sine wave adds together with the sine wave that we are sending with our hand to produce the standing wave.

2.11 The Wave Velocity of a Vibrating String

It is well known to players of string instruments that the pitch and hence fundamental frequency of a string increases with increasing tension. This fact is connected with the increase in the wave velocity with increasing tension. Similarly, one notices that for given string material, the pitch of a string decreases with increasing string thickness. This fact is connected with the decrease in the wave velocity with increasing mass of string, for given length of string. This section is concerned with the parameters that determine the wave velocity and the precise relationship among them.

It turns out that the wave velocity depends upon two parameters that characterize the string. First, we have the **Tension** , with the symbol $\mathcal{T}$. The tension acts to restore the string to its equilibrium shape and favors a greater wave velocity. The second parameter is the **mass per unit length**, with the Greek letter μ as a symbol. This parameter is also called the **linear mass density**. The mass of an SHO reflects its resistance to having its velocity change – that is, being accelerated. Similarly, the linear mass density of a string reflects the string's resistance to having any point along the string undergo a change in velocity. This set of changes is what constitutes a wave. Just as an increase in the mass of an SHO decreases its vibration frequency, an increase in the linear mass density decreases the wave velocity.

We will now discuss these two parameters in greater detail.

Let us turn our attention to tension. This parameter is measured in units of force such as the "pound" (lb) or the **Newton**(named after Isaac Newton), abbreviated as N. (The two are related as follows: 1 lb = 4.5 N.) A common device for measuring tension is a "spring scale".[5] On average, a string of a stringed instrument is under a tension of about 50 lbs (thus about 200 N). We will later show that the total tension of the strings of a piano is on the order of 70,000 lbs!

The physical parameter **force** has direction as well as magnitude. Thus, for example, the gravitational force of the earth on a person is downward, toward the center of the earth. On the other hand, *tension has no directionality*. This fact is illustrated in the following Fig. 2.18, wherein a string is being pulled on by two spring scales, one to the right and one to the left.

[5] If you attach a spring to one end of a string that is under tension, the tension is proportional to the consequent displacement of the spring.

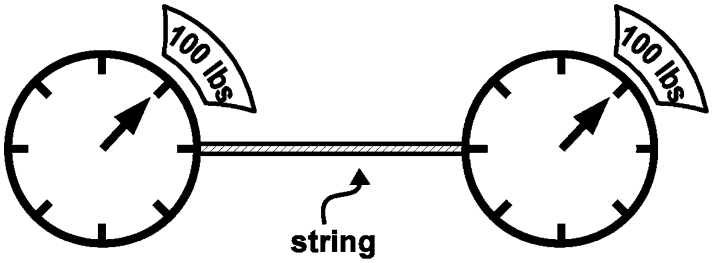


Fig. 2.18 String pulled from two directions

Both scales read a force of 100 lbs. This implies that the right scale pulls on the string with a force of 100 lbs to the right, while the left scale pulls on the string with a force of 100 lbs to the left. These two forces cancel each other out, so that the net force acting on the string is zero and the string can remain stationary in this situation.

The resulting tension, $\mathcal{T}$, in the string might not be obvious: It might seem that the two applied forces add to produce a tension of 200 lbs. In fact the tension is 100 lbs. We write

$$\mathcal{T} = 100 \text{ lbs}. \tag{2.29}$$

Note

The Nature of Tension

How can we understand the above enigma, that the tension is not 200 lbs? We will be able to answer this question by examining what the tension represents. Hence, let us imagine two people facing each other with their right arms outstretched and clasping each other's hand. Call them, Richard and Lisa. Richard's shoulder pulls his arm with a force of 10 lbs and so does Lisa's shoulder pull her arm with a force of 10 lbs. Richard's hand pulls Lisa's hand toward him with a force of 10 lbs and correspondingly, Lisa's hand pulls Richard's hand toward her with a force of 10 lbs. We say that there is a tension of 10 lbs at the point where the two hands are clasped.

Now let us turn to the string in Fig. 2.18. We consider two segments of this rope, labeled L and R, respectively, shown at the top of Fig. 2.19(a), whose boundary is marked by the letter "P". At the bottom of Fig. 2.19(a) we focus on segment L, noting the two forces acting on this segment that balance each other – 100 lbs by the scale to its left and 100 lbs by the R segment to its right. [In fact, the L segment pulls on the R segment towards L with a force of 100 lbs and correspondingly, the R segment pulls on the L segment towards R with a force of 100 lbs.] Since the point P is arbitrary, the tension is uniform all along the string and is said to be 100 lbs.

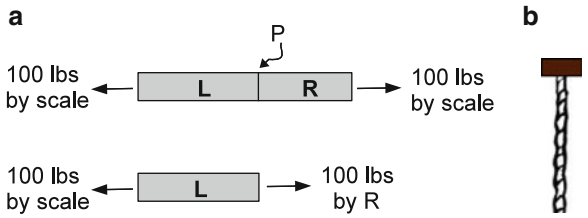


Fig. 2.19 Hanging rope

This is not so for a string with weight that is hanging from a support, as seen in Fig. 2.19(b). At any point along the string, the tension equals the weight of string below that point. Thus, the tension vanishes at the bottom end and equals the total weight of the string at the top end, where the string is supported. **Question**: If the string weighs 8 oz, what is the tension at the midpoint of the string?

We now turn to the **linear mass density**. Suppose we have a spool of string and cut off a meter length of string and find that it has a mass of 5 g. Then the linear mass density of the string in the spool is $5\,\mathrm{g/m} = 0.005\,\mathrm{kg/m}$. A 2-m length of such a string would have a mass of 10 g. As a result, the linear mass density will be $10\,\mathrm{g}/2\,\mathrm{m} = 0.005\,\mathrm{kg/m}$. The result is that there is no change in the linear mass density. Generally, for a length of string ℓ with a mass m, we have the relation

$$\mu = \frac{m}{\ell}. \tag{2.30}$$

The linear mass density of the string of the spool is independent of the length of the string.

We are now ready to reveal the relation between the wave velocity and the two parameters, $\mathcal{T}$ and μ. It is given by

$$v = \sqrt{\frac{\mathcal{T}}{\mu}} \tag{2.31}$$

which is the wave velocity along a string.

Note that the wave velocity involves a square root of a force-like parameter (here $\mathcal{T}$) divided by a quantity which reflects mass or inertia (here μ):

$$\text{Wave velocity} = \sqrt{\frac{\text{Force-like parameter}}{\text{Mass-like parameter}}}. \tag{2.32}$$

The force-like parameter is usually referred to as the **restoring force**.

This result is universal for all types of waves.

In this case, tension is the restoring force.

Now we can see fully how the fundamental frequency depends upon the three parameters of a vibrating string – the length, the tension, and the linear mass density. Using (2.6) and (2.31), we obtain

$$f = \frac{v}{2\ell} = \frac{\sqrt{\frac{\mathcal{T}}{\mu}}}{2\ell}. \tag{2.33}$$

Thus, we have found that **the fundamental frequency is proportional to the square root of the tension, inversely proportional to the square root of the linear mass density, and inversely proportional to the length**.

Sample Problem 2-3

Suppose that a violin string has a length of 33 cm and a linear mass density of 6 g/m, and has a fundamental frequency of 440 Hz. Find the wave velocity, the mass of the string, and the tension in the string.

Solution

$$v = 2f\ell = 2(440\,\text{Hz})(0.33\,\text{m}) = 290\,\text{m/s}$$

$$m = \mu\ell = 6 \times 10^{-3} \times 0.33$$

$$= 1.98 \times 10^{-3}\,\text{kg} = 1.98\,\text{g}.$$

To obtain the tension is a bit more complicated because it appears within a square root:

$$v = \sqrt{\frac{\mathcal{T}}{\mu}}$$

so that

$$v^2 = \mathcal{T}/\mu$$

and

$$\mathcal{T} = \mu v^2 = (0.006\,\text{kg/m})(290)^2 = 506\,\text{N}.$$

2.11.1 Application of the Above Relations to the Piano

We will now review in a bit of detail the methods whereby the huge range of frequencies (and hence pitches) of piano strings, from 27.5 to 4,186 Hz, can be obtained:

To increase the pitch, one can

- Increase the tension
- Decrease the linear mass density, or
- Decrease the length

We can obtain an idea of the **total tension** on the block of metal that supports the strings as follows:

From the relation

$$v = \sqrt{\frac{\mathcal{T}}{\mu}},$$

we obtain $\mathcal{T} = \mu v^2$.

In order to estimate the average wave velocity, we will use the relation $v = 2Rf$. Since the average frequency is about 500 Hz and the average length of the strings is about 0.5 m, we obtain $v \sim 2(0.5)(500) = 500$ m/s. Now for the linear mass density, which in kg/m is the mass in kilogram of a 1 m length of string. That mass is the product of the mass density (8 g/cc $= 8{,}000$ kg/m^3 for the steel of most piano strings) and the volume of 1 m of string. From observation, the strings have an average radius of about 0.5 mm. The above volume V is thus 1 m times the area of a circle of radius 0.5 mm. Since 1 mm $= 10^{-3}$ m, we obtain

$$V = (1\text{ m})\pi R^2 = (1\text{ m})(0.5 \times 10^{-3}\ m)^2 = 8 \times 10^{-7}\ \text{m}^3.$$

Hence, $\mu \sim 8{,}000(8 \times 10^{-7}) = 6.4 \times 10^{-3}$ kg/m^3 $= 6.4$ g/m^3. The average tension is then

$$\mathcal{T} = \mu v^2 \sim (6 \times 10^{-3})(500)2 = 1{,}500\text{ N} \sim 300\text{ lbs}.$$

This is the average tension in a single string. Most keys have a few strings (i.e., there is more than one string per note), so that the total number of keys is about 230. Our estimate for the total tension is then $230 \times 300 = 69{,}000$ lbs.

We have seen that the wave velocity of a vibrating string involves a characteristic force parameter (the tension) and a mass parameter (the linear mass density). We will next see how the behavior of an SHO can help us understand the expression for the wave velocity of a vibrating string. The SHO has a force parameter (the spring constant) and a mass that together determine its vibration frequency.

2.12 The Connection Between an SHO and a Vibrating String**

What has the SHO in common with a vibrating string? To simplify our analysis, we will examine the vibration of a string that is pulled aside at its midpoint and then released and allowed to vibrate.

We have seen that an SHO is characterized simply by a mass that is displaced and experiences a restoring force proportional to its displacement. The entire length of string is the corresponding mass. There is variable displacement all along the length of the string; so the system is more complicated. As an approximation we will let the displacement of the midpoint correspond to the displacement of the SHO. (See Fig. 2.20.)

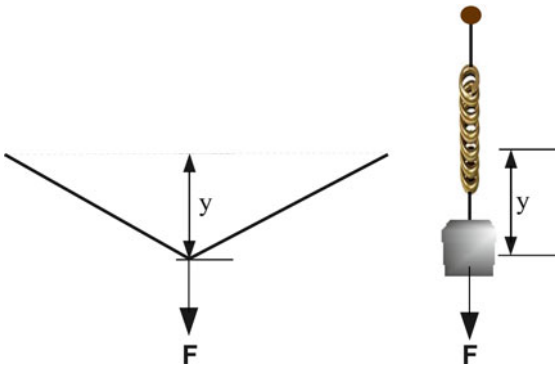


Fig. 2.20 The plucked string vs. the SHO

It can be shown that for small displacements, the restoring force on the string is proportional to the displacement. "Small" means displacement much less than the length of the string. A simpler description of this restriction is that the slope of the string during vibration must be very small. The result is

$$F = \frac{4y}{\ell}\mathcal{T} = \frac{4\mathcal{T}}{\ell}y. \tag{2.34}$$

That is, the restoring force is given by the tension $\mathcal{T}$ reduced by a factor $(4y/\ell)$, which is typically much less than unity. [For example, a guitar string of length 650 mm might vibrate with an amplitude of but a few mm.] Alternatively, we can express this relation as

$$F = \frac{4\mathcal{T}}{\ell}y. \tag{2.35}$$

We see that the restoring force is proportional to the displacement y, as in the case of an SHO. This is the essential reason that a string vibrates sinusoidally. The effective spring constant is defined by the relation $F = ky$. Thus, it is given by

$$k = \frac{4\mathcal{T}}{\ell}. \tag{2.36}$$

Sample Problem 2-4

Suppose that a string has a tension equal to 200 N and a length equal to 33 cm(=0.33 m). Find the effective spring constant.

Solution

$$k = \frac{4\mathcal{T}}{\ell} = \frac{4(200)}{0.33} = 2{,}400\,\text{N/m}.$$

Sample Problem 2-5

Suppose that the previous string is displaced at its midpoint by a distance of 1 mm(=1/1,000 m). Find the restoring force.

Solution

$$F = kx = (2{,}400)(0.001) = 2.4\,\text{N}.$$

We can now combine the exact expression (2.17) for the period of an SHO with our expression (2.36) for the spring constant of the plucked string so as to obtain an expression for the **period of a plucked string**.

$$T_{\text{SHO}} = 2\pi\sqrt{\frac{m}{k}} \qquad \text{along with} \qquad k \approx \frac{4\mathcal{T}}{\ell} \tag{2.37}$$

to obtain

$$T_{\text{SHO}} \sim 2\pi\sqrt{\frac{m}{4\mathcal{T}/\ell}} = \pi\sqrt{\frac{m\ell}{\mathcal{T}}}. \tag{2.38}$$

Finally we are ready to obtain an approximate expression for the wave velocity along a string:

$$v_{\text{approx}} = \frac{2\ell}{T_{\text{SHO}}} = \frac{2\ell}{\pi\sqrt{\frac{m\ell}{\mathcal{T}}}}. \tag{2.39}$$

Substituting μ for the ratio m/ℓ, we obtain:

$$v_{\text{approx}} = \frac{2}{\pi}\sqrt{\frac{\mathcal{T}}{\mu}} \qquad \text{wave velocity along a string.} \tag{2.40}$$

Why the difference between the two equations for the wave velocity, (2.31) and (2.40), amounting to a ratio of $2/\pi \sim 0.6$? The spring constant is the ratio of force to displacement. The displacement of an SHO is well defined with a specific value. In contrast, for a string, the displacement varies from zero at the ends of the string to its maximum value at the string's center. As a consequence, we have overestimated the average displacement and underestimated the effective spring constant. This leads to an overestimate of the fundamental period.[6]

[6]The exact expression for the fundamental period of a plucked string is

$$T = 2\sqrt{\frac{m\ell}{\mathcal{T}}}. \tag{2.41}$$

2.13 Stiffness of a String

So far, we have assumed that the vibrating string is completely flexible. No force is necessary to bend the string. The term **stiffness** is used to characterize the force necessary to bend a string.

The physical parameter that determines stiffness is the same as that which determines the force necessary to stretch a string. It is called **Young's modulus**. Why is this so? Because when a string is bent, one side of the string is stretched while the other side is compressed. This can be seen in Fig. 2.21, where a string of length $L = \pi R$ is bent into a semi-circle. The thickness of the string is $R_2 - R_1$. The outer perimeter has a length πR_2, while the inner perimeter has a length πR_1. While the outer perimeter is stretched by an amount $\pi(R_2 - R)$, the inner perimeter is compressed by an amount $\pi(R - R_1)$. Thus, the difference in the perimeters is $\pi(R_2 - R_1)$, or π times the thickness.

Note: The shape of the wave for the vibrating stiff string in the nth partial is sinusoidal even in the presence of stiffness. It is a portion of a sine wave having a wavelength

$$\lambda = 2\ell/n. \tag{2.42}$$

While we will not discuss Young's modulus because the subject is beyond the scope of this text, we can see qualitatively what effect stiffness might have on the wave velocity along a string and more importantly on the frequency spectrum of the modes.

First, we expect that stiffness contributes to bringing the string back from a curved shape toward a straight shape. It is a restoring force . Therefore, we expect that the wave velocity will increase. Next, as the wavelength decreases, the degree of bending increases. Therefore, we expect the wave velocity to increase with decreasing wavelength – or alternatively, to increase with increasing frequency.[7]

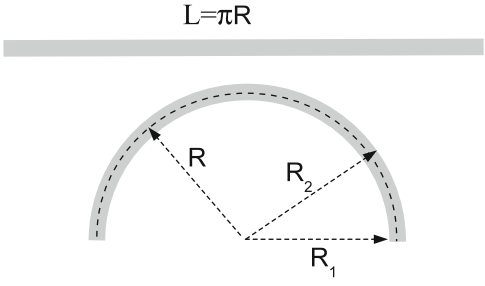


Fig. 2.21 A thick string bent into a semi-circle

[7]Mathematically, the wave velocity can be expressed as

$$v = \sqrt{\frac{\mathcal{T}}{\mu} + \frac{\overline{\mathcal{B}}}{\rho\lambda^2}}, \tag{2.43}$$

Finally, since the wavelength is inversely proportional to n, the effect of stiffness increases with increasing n. In fact, it can be shown that the frequency of the nth partial is given by

$$f_n = n\frac{\sqrt{\frac{\mathcal{T}}{\mu}}}{2\ell}\sqrt{1+\mathcal{B}n^2}, \tag{2.45}$$

where $\mathcal{B}$ is a number that is inversely proportional to the square of the length of the string.[8] Thus, the longer the string, the smaller the effect of stiffness. Relatively, the longer a string, the easier it is to bend it. Therefore, for a given mode, longer strings have less of an effect due to stiffness. Also, the constant is proportional to the fourth power of the radius of the string. As a result, thicker strings are stiffer, as we would expect. Typically, the constant $\mathcal{B}$ is much less than unity, so that the effect is small for $n = 1$. On the other hand, for large n the effect will be much more significant. Alternatively, we can write

$$f_n = nf_0\sqrt{1+\mathcal{B}n^2}. \tag{2.46}$$

Here, the parameter $f_0 = v/2\ell = \sqrt{\mathcal{T}/\mu}/(2\ell)$ is the fundamental frequency f_1 in the absence of stiffness. Note that the fundamental frequency is different in the presence of stiffness: Instead, $f_1 = f_0\sqrt{1+\mathcal{B}}$. The equation for f_n shows us plainly that the frequency spectrum is no longer a harmonic series.

To gain a sense of the order of magnitude of the constant $\mathcal{B}$, we find that for a steel wire of radius 1 mm and a length of 1 m, under a tension of 100 Newtons, $\mathcal{B} = 0.008$. For the fundamental mode, the correction to the frequency is less than 1%. However, as the mode number increases, the correction increases too, and not proportionately. We can see the effect dramatically in the graph of Fig. 2.22. The dark curve represents the spectrum with stiffness, while the straight line in magenta represents the spectrum without stiffness included.

Note that the effect on the first two modes is not great. For the third mode, the corresponding frequencies are 1,500 and 1,600 Hz, respectively – a significant difference. For the fifth mode, the difference is dramatic: 2,500 vs. 3,000 Hz.

---

where $\overline{\mathcal{B}}$ is a constant. If the tension is absent, as is the case for a suspended rod of metal – e.g., one prong of a tuning fork – the speed of transverse vibrations is given by

$$v = \sqrt{\frac{\overline{\mathcal{B}}}{\rho\lambda^2}}. \tag{2.44}$$

We see from this equation that the ratio $\overline{\mathcal{B}}/\lambda^2$ is the restoring force. Sometimes this force is referred to as the **bending force**. This is the equation for the speed of **transverse waves** along a solid rod. The rod also exhibits **longitudinal sound waves**, which will be discussed in the next chapter.

[8] Explicitly, $\mathcal{B} = (\pi Y/\mathcal{T})(\pi a^2/2\ell)^2$, where a is the radius of the string and Y is **Young's modulus**. Young's modulus determines how much of an elongation $\Delta\ell$ results from tension. Thus, if a string is under a tension $\mathcal{T}$, the relative change in its length is given by $\Delta\ell/\ell = \mathcal{T}/(\pi a^2 Y)$.

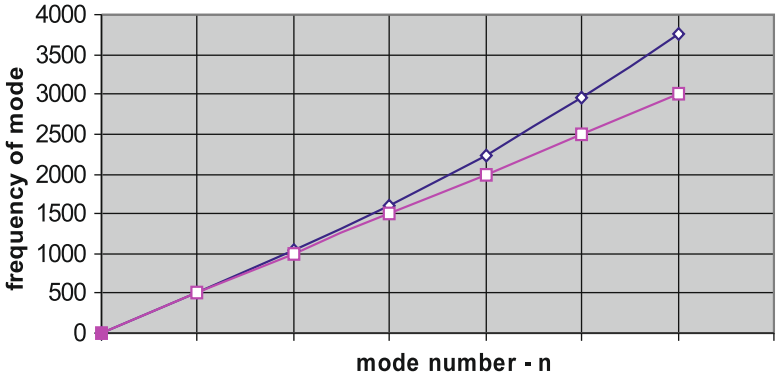


Fig. 2.22 Frequency of modes for string with (*dark curve*) and without stiffness (*magentacurve*)

2.14 Resonance

Consider the process of exciting a mode of vibration of a string. It requires that you move the end of the string up and down at a frequency equal to the frequency of that mode. If your hand moves at a frequency that is different from a mode frequency, the degree of excitation of the mode will not be great. However, the closer the frequency match is, the greater the ultimate amplitude of the excited mode.

We say that there is a **resonance** between the two systems – your hand and the string – when there is a frequency match, or practically speaking, close to a frequency match, so that there is a high degree of excitation.

For a simple example of resonance, consider two SHOs having an identical frequency f. We connect them with a fine string as shown in Fig. 2.23. Clearly, either SHO can excite vibrations in the other through the coupling between them.

Digression on the Modes of Two Coupled SHOs

Suppose that both SHOs are released from rest with the same initial displacement. Clearly, they will oscillate up and down at their common mode frequency f, because the coupling between them will be **inactive**. We say that the SHOs oscillate **in phase**. Now, suppose that the two SHOs are released from rest with the same initial amplitude but NOW with **displacements in opposite directions**. The two SHOs will oscillate up and down, always in opposite directions. We say that they oscillate **out of phase**. In this case, there will be strong coupling between the two SHOs.

What we have described above are the two modes of a pair of coupled SHOs. They are depicted in Fig. 2.24. The IN-PHASE mode has a frequency $f_{\text{in}} = f$,

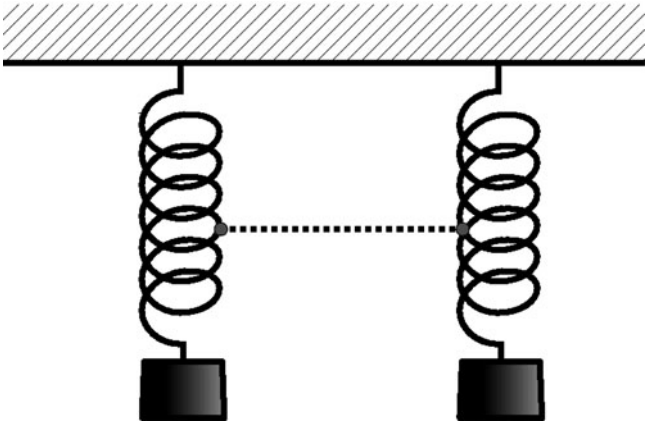


Fig. 2.23 Resonance between two coupled SHOs

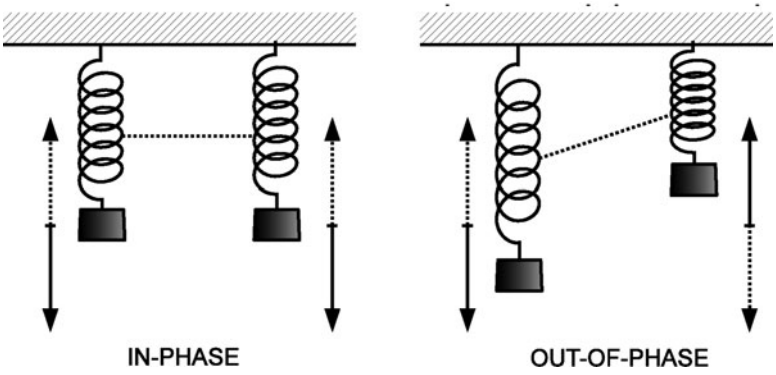


Fig. 2.24 Two modes of two coupled SHOs

while the OUT-OF-PHASE mode has a frequency f_{out} that is slightly larger. The *difference* Δf between these two frequencies increases with increasing coupling between the strings and vanishes in its absence.

It can be shown that any particular motion of the two SHOs can be expressed as a sum of the two modes. A very interesting example is the following:

Suppose that both SHOs are released from rest, with the **left** SHO displaced downward by an amount A from its equilibrium position, while the **right** SHO is kept in its equilibrium position. We see that the initial condition is a sum of the initial conditions described above for the two modes. As a consequence, the subsequent motion is a sum of the in-phase and the out-of-phase modes, with equal amplitudes $A/2$ of each in the sum. The resulting subsequent motion is quite interesting. (We neglect attenuation, for simplicity.)

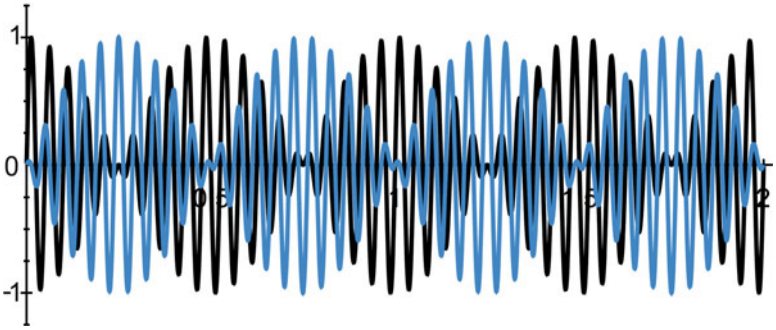


Fig. 2.25 Exchange of oscillation between two coupled SHOs

The left SHO will begin oscillating with an amplitude A. In time, that amplitude will decrease, while the right SHO will begin to oscillate. The amplitude of the left SHO will eventually momentarily vanish, while at the same time, the amplitude of the right SHO will equal A: The left SHO will have passed its energy (initially potential energy) entirely onto the right SHO!

Subsequently, the roles will be reversed, with the right SHO passing its energy back to the left SHO. Ultimately, the two SHOs will exchange energy sinusoidally at an **exchange frequency** f_{ex} that is exactly equal to the frequency Δf! The time dependence of the displacements of the two oscillators is shown in Fig. 2.25. The black curve represents the oscillation of the left SHO, while the blue curve represents the oscillation of the right SHO. Two cycles of exchange are shown. Note how the left SHO comes to rest at the time 0.5 units, where the right SHO has its maximum oscillation.[9]

2.15 General Vibrations of a String: Fourier's Theorem

Suppose that you do not move a string up and down at exactly any of the mode frequencies. The string will vibrate; however, the pattern of vibration will not resemble any one of the modes unless there is a near frequency match. Furthermore, a plucked string does not vibrate with the pattern of any of the modes, even though the pattern vibrates periodically at the fundamental frequency! **How is the general vibration of a string related to the various modes of vibration?** The answer

[9]Note that a musical instrument can have a few vibrating components, such as does the violin – the two components being a vibrating string and a vibrating wooden plate. It can be desirable to have mode frequencies of the components match: the original source of vibration – as from a bowed string – might not transfer the vibration into the air efficiently. The second component – here the wooden plate – might be able to do so efficiently. With efficient transfer of vibration to the air, the second component will not fully return the vibration back to the original source and the above annoying phenomenon will be reduced.

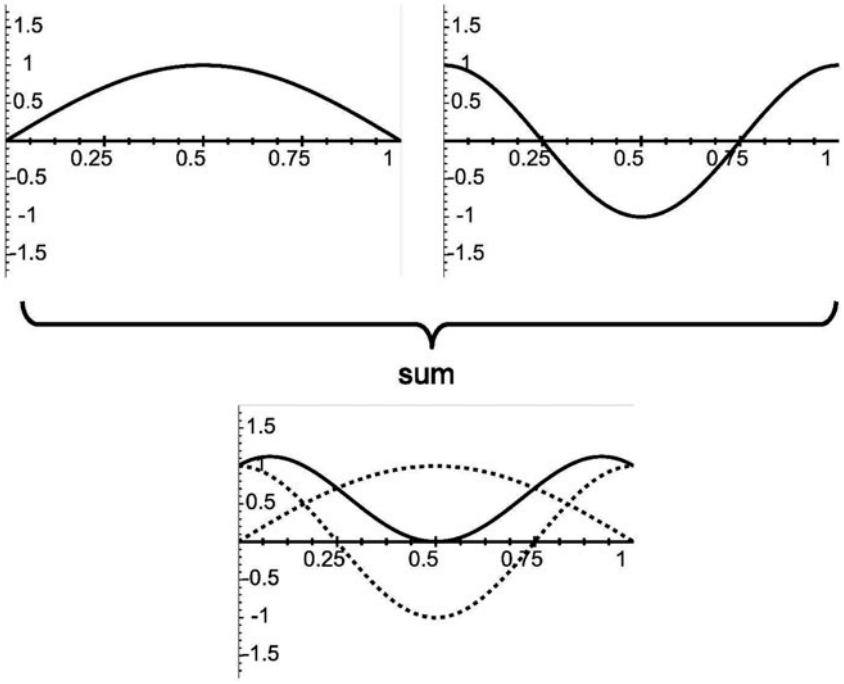


Fig. 2.26 Summing two sine waves

lies in a mathematical theorem due to **Jean Baptiste Joseph Fourier**, a French mathematician who lived from 1768 to 1830. Here is **Fourier's Theorem** in words:

> Any pattern can be expressed mathematically as a sum of sine waves, the amplitudes of each sine wave in the sum being unique.

This theorem is analogous to one of the most important theorems in number theory that any number can be expressed as a product of prime numbers. (For example, $60 = 2 \times 2 \times 3 \times 5$. The representation of the number by such a product is unique. In particular, the number of times that any prime number appears in the product is unique.) And so it is with the amplitudes of the various sine waves in the sum which represents the pattern.

Given a particular pattern, there are mathematical as well as electronic means for obtaining the unique mixture of sine waves associated with the pattern, a process known as **Fourier analysis**. Each individual sine wave is referred to as a **Fourier component**. To specify a Fourier component, we need to know three factors: its **frequency**, its **amplitude**, and its **relative phase**.

The set of frequencies in the mixture of sine waves is called the **Fourier spectrum** or simply the **frequency spectrum**. The amplitudes of the sine waves in the sum are called **Fourier amplitudes**. The "sum" is obtained by a straightforward graphical sum of the curves representing the waves. For example, in Fig. 2.26 we exhibit the sum of two Fourier components A and B, with the resultant SUM.

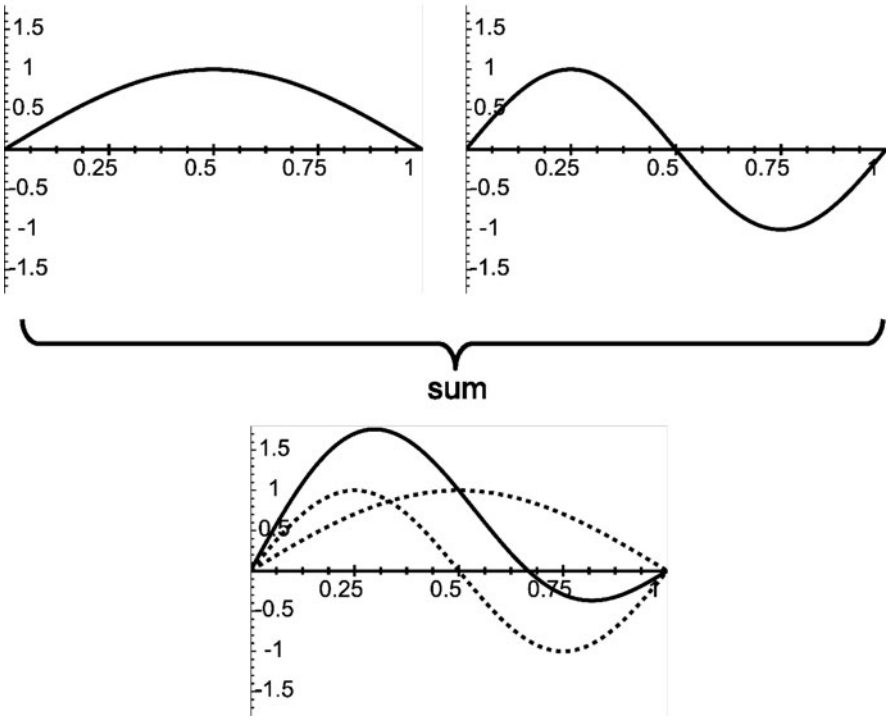


Fig. 2.27 Summing two sine waves with different phase relation from above

The **relative phase** refers to the relative positions of the waves. To appreciate the significance and importance of the relative phase, in Fig. 2.27 we exhibit the sum of the same two Fourier components as Fig. 2.26, except that component B has been shifted by a quarter of a wavelength so as to produce component C.

The reverse process of adding the given mixture of sine waves corresponding to a specific pattern so as to produce that pattern is called **Fourier synthesis**.

Generally, the frequency spectrum will include all frequencies, from zero to infinity. This is not the case for a finite vibrating string. Here, Fourier's theorem leads to the result that any pattern of vibration is a sum of the modes of vibration of the string, with a unique set of amplitudes for each mode in the sum. Hence, in this case the Fourier spectrum is a harmonic series.

There is a **corollary** to Fourier's theorem that is central to understanding the basis for obtaining a sense of pitch from musical instruments:

The Fourier spectrum of any periodic wave – and hence any sound wave that has a well-defined single pitch – must be a harmonic series with a fundamental frequency equal to the frequency of the periodic wave.

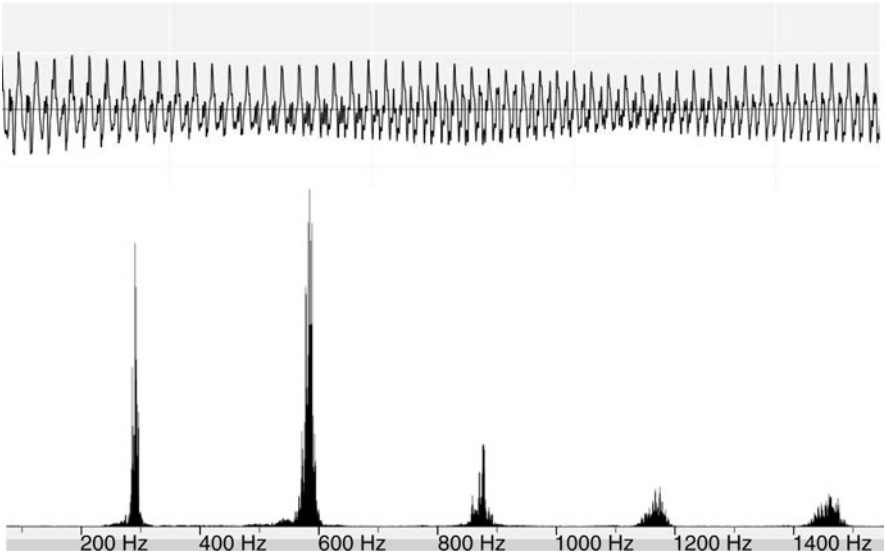


Fig. 2.28 VIOLIN wave and frequency spectrum

As a consequence, the frequency spectrum for the modes of vibration of a musical instrument that produces a well-defined pitch must be a harmonic series. In contrast, the frequency spectrum of a gong, a tuning fork, or a drum is not a harmonic series. Unless one excites one mode alone of these instruments, the sound produced will be perceived to have more than one sense of pitch. This fact is exhibited in the spectra of some musical instruments. In Fig. 2.28, we see the wave and spectrum of a short segment of sound from a violin. We can see the peaks at the harmonics with a fundamental frequency of about 280 Hz. Note the variation in the envelope of the wave, corresponding to varying loudness. More interesting are the numerous spikes surrounding the main peak. These partially reflect the **vibrato**, which is briefly discussed below. In contrast, we see in Fig. 2.29 the wave and spectrum of a segment of sound from a flute. The absence of much contribution from harmonics above the second is evident.[10]

NOTE: We must remember that the spectrum for any given instrument varies, depending upon how a note is played by the musician.

[10]The waves and spectra were produced using mp3s of instrumental sounds downloaded into the program **AmadeusPro**.

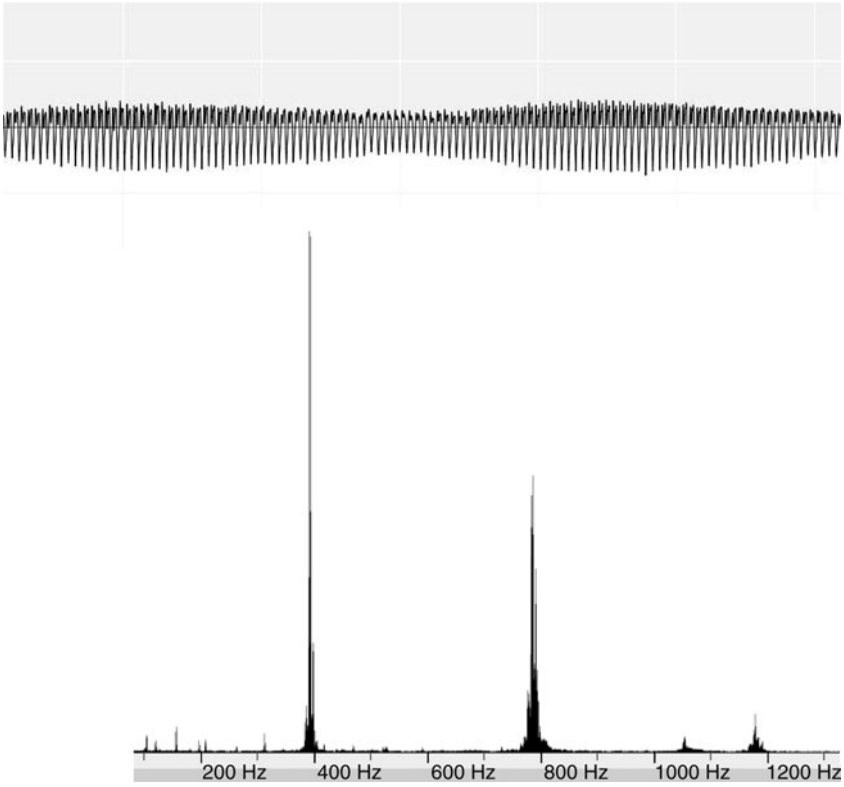


Fig. 2.29 FLUTE wave and frequency spectrum

Note

The sound of a violin exhibits characteristics that are analogous to the two most important modes of communicating audio signals with radio – AM (**amplitude modulation**) and FM (**frequency modulation**).[11]

Consider the central fundamental frequency of the above violin wave – 380 Hz. Let us suppose that the envelope of the wave oscillates at a frequency of 2 Hz. This is the **amplitude modulation frequency**. An AM radio wave of WEEI in Boston, MA, has a frequency of the **carrier wave** of 550 kHz. This frequency corresponds to the 380 Hz of the violin. If the station wants to transmit an audio signal of 400 Hz, the amplitude modulation frequency is 400 Hz.

[11] See the extremely informative applet on this website (2-15-2011): http://engweb.info/courses/wdt/lecture07/wdt07-am-fm.html#FM_Applet You can vary the modulation frequency as well as the amplitude of modulation and observe the changing wave form as well the resulting frequency spectrum.

> Now let us turn to the frequency modulation, which is strongly produced by **vibrato**[12]: The fundamental frequency of the violin is determined by the position of a finger on the violin string that restricts the length that is free to vibrate. If the violinist rocks the finger back and forth on the string at a frequency of 5 Hz, the length that is free to vibrate will oscillate at this frequency and the sound will be frequency modulated at this frequency.[13] In the case of FM radio, an **FM radio wave** from WCRB-FM in Boston would have a **carrier frequency** of 89.7 MHz; the audio signal of 400 Hz would be the **FM modulation frequency**.

NOTE: The general term for the frequency of a mode of vibration of a system is the **partial**. The first partial is always equivalent to the fundamental frequency. For a vibrating string without stiffness, the set of partials forms a harmonic series.

In order to illustrate the results of a Fourier analysis of a vibrating system, consider the vibration of a string that is plucked at its midpoint, as shown previously in Fig. 2.14. It can be shown using mathematical analysis that the pattern of vibration is a sum of all the odd modes of vibration of the string. Suppose that at some instant the amplitude is $\pi^2/8$ at its midpoint. The shape of the string is triangular. The amplitudes A_1, A_2, A_3, ...of the sine waves that reproduce this pattern are given by:

$$\text{Fundamental: } A_1 = 1$$

$$\text{3rd harmonic: } A_3 = -\frac{1}{9}$$

$$\text{5th harmonic: } A_5 = \frac{1}{25}$$

$$\text{General odd harmonic: } A_n = \frac{(-1)^{(n-1)/2}}{n^2}, \quad \text{where } n = 1, 3, 5, \ldots$$

We illustrate this result in Fig. 2.30, where we show how adding sine waves produces a triangular wave pattern. The black curve is the desired triangular wave. The blue curve is the first harmonic – $n = 1$ or the function $\sin(\pi t)$. The red curve

[12]We realize that vibrato is what gives the violin its sweet tone. However, vibrato is probably a very important factor in a number of other specific ways: For example, see Sect. 10.7 for a discussion of the fusion of harmonics and Sect. 11.7, wherein we discuss the important role that vibrato certainly plays in allowing us not to be affected by the impossibility of performing combinations of musical pure tones that are consistently consonant.

[13]It can be shown that the resulting frequency spectrum consists of a central peak at 380 Hz along with side peaks at frequencies, $380 \pm 5 = 375$ and 385, $380 \pm 10 = 370$ and 390, and $380 \pm 15 = 365$ and 395, ... The weight of these side frequencies falls off as we move to greater distances from the fundamental and depends upon the amplitude of the rocking motion.

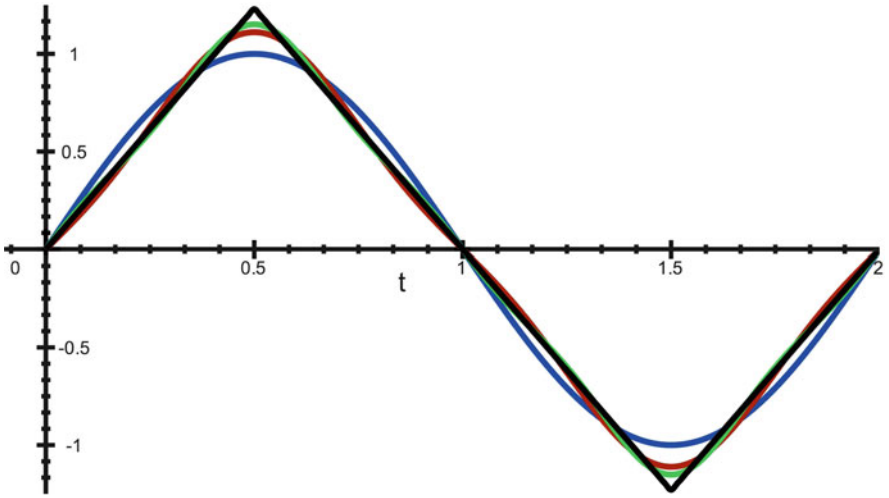


Fig. 2.30 Sum of Fourier components to produce a triangular wave

is the sum of the first and third harmonics – $n = 1$ and $n = 3$ – that is, the function $\sin(\pi t) - (1/9)\sin(3\pi t)$. We see that these two terms alone are within about 10% of reproducing the triangular wave. The green curve is the sum of the first, third, and fifth harmonic.

2.15.1 Frequency of a Wave with Missing Fundamental

It is probably obvious that a wave that includes the fundamental in its spectrum has the frequency of the fundamental. For example, the frequency of a mixture 100, 200, and 500 Hz is 100 Hz. However, consider the mixture 200 and 500 Hz. What is the frequency in this case?

Suppose we focus our attention on the displacement of a specific point along a vibrating string. According to Fourier's theorem, its wave pattern in time is a sum of sine waves, all of which are members of the harmonic series of the string's mode spectrum. It can be shown that the pattern is always **periodic, with a frequency equal to the largest common denominator (LCD) of all the frequencies in the Fourier spectrum**. For this last example, the LCD is 100 Hz, which is the frequency of wave, even though the fundamental 100 Hz is missing. We can appreciate this result as follows: During one cycle of oscillation over the period of $1/100 = 0.01$ s there will be exactly two cycles of the 200 Hz component and five cycles of the 500 Hz component.

2.16 Periodic Waves and Timbre

We can now appreciate a major factor that distinguishes the timbre of one musical instrument from another: Two instruments that are producing the same *steady* musical note are producing periodic patterns having the same frequency. It is this frequency that determines the *pitch* of the note. However, the two sets of relative amplitudes of the Fourier components are different. This difference is one of the important factors that distinguishes the timbres of musical instruments.

There are two other factors that contribute to our ability to distinguish one instrument from another when the notes are not steady but have a beginning and end; they are the **attack** and the **decay** parts of the note, which are depicted in Fig. 2.31. The variation of the amplitude – defining with the growth and final decay – is called the **envelope**. It is given by the pair of dashed curves in the figure.

Experiments have shown that in the absence of differing envelopes, it is often difficult to distinguish the sounds of different instruments.

2.17 An Application of Fourier's Theorem to Resonance Between Strings

When a string is disturbed, generally a mixture of modes is excited. The Fourier amplitudes depend upon the manner in which the string is excited. This fact has important ramifications with regards to resonance between strings: Consider two strings, one tuned to 440 Hz and the second to 660 Hz. (These two frequencies correspond to the **fundamental** frequencies of the respective strings.) The frequency spectra are:

440 Hz string: 440, 880, 1,320, 1,760, ...
660 Hz string: 660, 1,320, 1,980, 2,640, ...

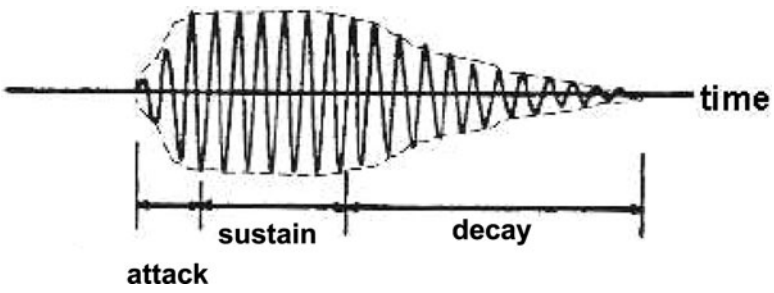


Fig. 2.31 The attack, decay, and envelope of a wave

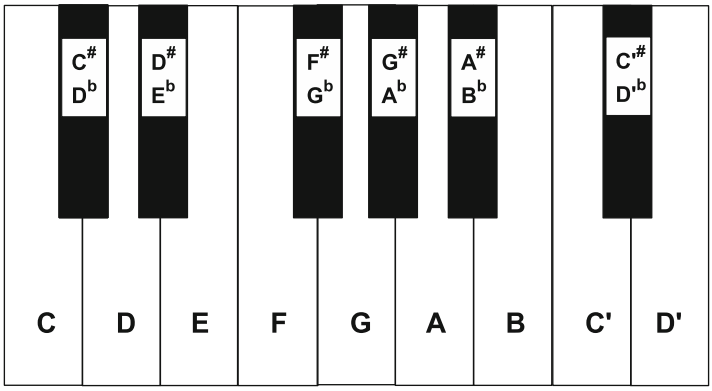


Fig. 2.32 Piano keyboard

We see that the third harmonic of the 440 Hz string and the second harmonic of the 660 Hz string have the same frequency. Thus, a general excitation of the 440 Hz string can strongly excite the second harmonic of the 660 Hz string or, a general excitation of the 660 Hz string can excite the third harmonic of the 440 Hz string.

Can you find a **second** matching pair of frequencies for the above strings? Discuss resonance between a 440 Hz string and a second string tuned to its octave at 880 Hz.

Resonance among the strings of a stringed instrument enriches tone quality. It therefore provides us with a partial explanation as to why good intonation of a string player – that is, playing notes "in tune" – improves tone quality. Conversely, by becoming more aware of the resonant response among strings, a string player can improve intonation.

Home Exercise with a Piano

If you have a piano available, you can observe the resonances discussed above as follows. Let us refer to the **piano keyboard** depicted in Fig. 2.32:

You will note that there is a pattern of the keys that repeats itself. Each cycle of keys is called an octave, with the white keys labeled from A through G. Focus on the key labeled C. This C is called *"middle-C"*. The A above middle-C is the key that is first tuned by a piano tuner, presently usually at a frequency of 440 Hz. We will call this key $A-440$. The A above it, being one octave above, is tuned at double this frequency. The E above $A-440$ is tuned at a frequency that is close to 660 Hz. (See Chapter 11, TUNING, INTONATION, AND TEMPERAMENT: CHOOSING FREQUENCIES FOR MUSICAL NOTES for more details on the choice of frequencies.)

Now, hold the $A-880$ key down so as to free the string from a damper which prevents it from vibrating. Next, give the $A-440$ key a sharp, "staccato" blow, so

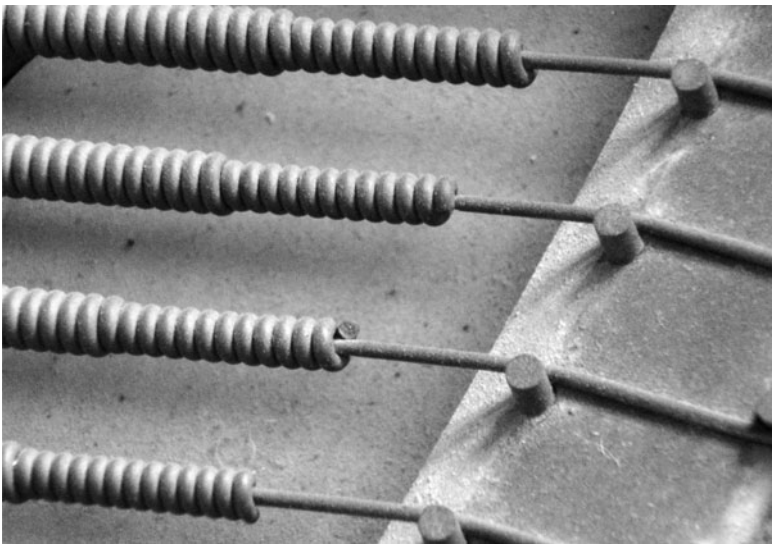


Fig. 2.33 Wound piano strings

that the $A-440$ tone will sound long enough as to excite the $A-880$ string, but short enough so that you can eventually hear the sound of the $A-880$ string. To confirm that you are hearing a sound produced by the $A-880$ string, release the $A-880$ key so as to dampen that string's vibration.

Repeat the above by exchanging the roles of the two strings. Next, repeat all of the above with a second pair of strings – say, the $A-440$ and $E-660$ strings.

> ## Note
>
> Let us now recall that stiffness causes the frequency spectrum of a string not to be a harmonic series. The effect is strongest for the thick strings of a piano at the low end, where increased thickness is necessary to produce the low frequencies. As a result, stiffness reduces the degree of resonance among these strings. In order to reduce the effect of stiffness, these strings are constructed out of a central core of steel that is surrounded by a coil, as seen in Fig. 2.33.

In an effort to reduce the mismatch of common harmonics, pianos are **stretch tuned** – a feature discussed further in a problem of Chap. 11. Furthermore, it is interesting to note that while we tend to regard resonance as a desirable characteristic, the reduced resonance in a piano is often regarded as an attractive feature of the sound of a piano.

2.18 A Standing Wave as a Sum of Traveling Waves

A standing wave is not a traveling sine wave since it is moving neither to the right nor to the left. According to Fourier's theorem, a standing wave must be a sum of sine waves. In fact, it is a sum of two sine waves having the same wavelength and amplitude but traveling in opposite directions. This fact is depicted in Fig. 2.34. Diagrams (a)–(e) one-eighth of a cycle apart. Each diagram depicts the position of the two component sine waves and their sum. The figure reflects a special property of sine waves: **The sum of two sine waves having the same wavelength is a sine wave having the same common wavelength. The amplitude of the resultant sine wave depends upon the amplitudes of the components and their relative phase.** In our case, the components have exactly the same amplitude. In (a), the components are in phase and the resultant wave has an amplitude that is double that of the components. In (c), the components are out of phase so that the components cancel each other. Note that while the displacement vanishes everywhere in (c), the string does have an instantaneous velocity. This situation can be compared to the SHO whose mass is passing through the equilibrium position.

We can now understand how we are able to set up a standard wave along a string of finite extent: With our hand, we propagate a sine wave down the string. The reflected wave is a sine wave traveling in the opposite direction, which when added to the original wave forms a standing wave! (Of course, the observed standing wave is only a portion of an infinite standing wave.)

2.19 Terms

- Amplification
- Amplitude analyzer
- Antinode
- Attenuation
- Bending force
- Centi- 10^{-2}
- Chladni plate
- Cycle
- Damping
- Direction of propagation
- Dispersion
- Dispersive
- Displacement
- Dissipation
- Equilibrium state
- Excitation
- Force constant (or 'spring constant') **k**
- Fourier Analysis
- Fourier component
- Fourier spectrum
- Fourier Synthesis
- Fourier theorem
- Frequency **f**
- Fundamental mode
- Fusion of harmonics
- Giga- 10^{9}
- Gong sound
- Harmonic
- Harmonic series
- Hertz (Hz) (a unit of frequency)
- Integral multiples

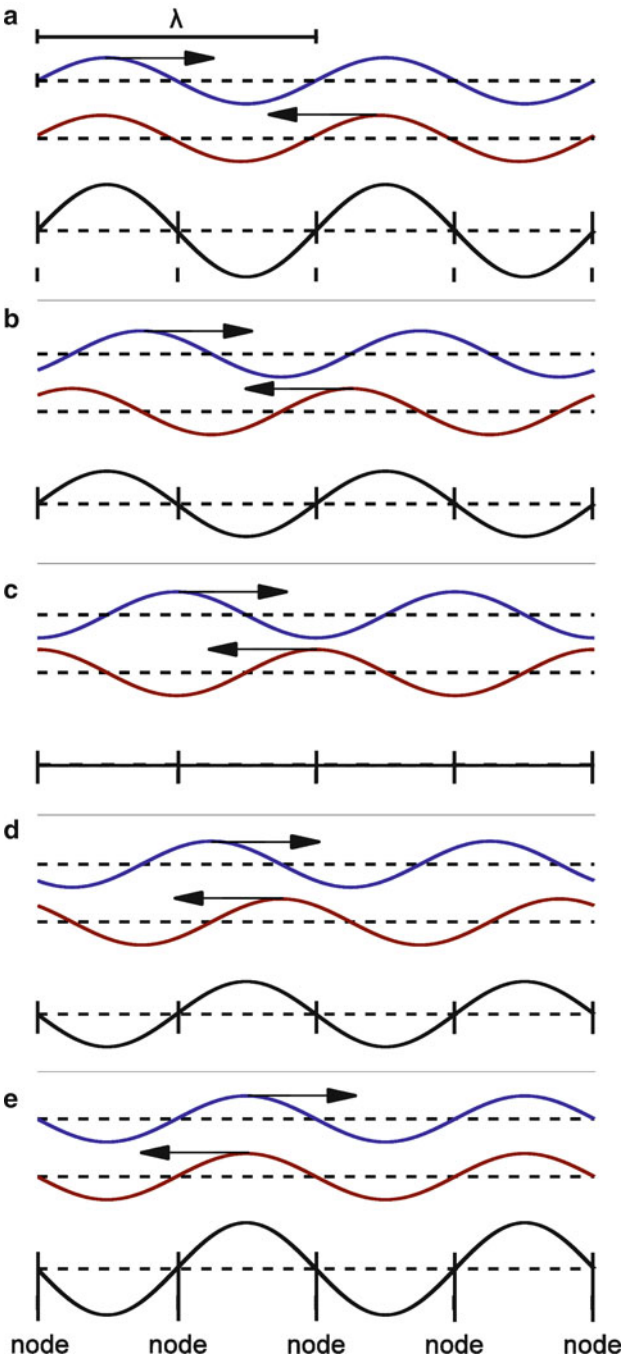


Fig. 2.34 Standing wave from two traveling waves

- Kilo- 10^3
- Largest common denominator
- Linear mass density μ
- Longitudinal wave
- Medium for wave propagation
- Mega- 10^6
- Micro- 10^{-6}
- Milli- 10^{-3}
- Nano- 10^{-9}
- Newton (N) (a unit of force)
- Nodal line
- Node
- Octave of notes oscillation
- Overtone
- Period **T**
- Periodic wave (in time or space)
- Phase relation
- Pitch
- Pluck
- Pulse
- Resonance
- Restoring force
- Simple harmonic oscillator (SHO)
- Sinusoidal
- Sonometer
- Spectrum
- Standing wave
- Stiffness
- Stretch tuning
- Stroboscope
- Tension $\mathcal{T}$
- Timbre or tone quality
- Transverse wave
- Travelling wave
- Tuning fork
- Wave propagation
- Wave velocity **v**

2.20 Important Equations

$$x = vt. \tag{2.47}$$

$$f = \frac{1}{T}. \tag{2.48}$$

$$f_1 = \frac{v}{2\ell}, f_2 = 2\frac{v}{2\ell} = \frac{v}{\ell}, f_3 = 3\frac{v}{2\ell}, \ldots. \tag{2.49}$$

$$\lambda f = v. \tag{2.50}$$

Linear mass density:

$$\mu = \frac{m}{\ell}. \tag{2.51}$$

$$v = \sqrt{\frac{\mathcal{T}}{\mu}}. \tag{2.52}$$

Hooke's Law:

$$F = ky. \tag{2.53}$$

Frequency of a Simple Harmonic Oscillator:

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}. \tag{2.54}$$

General form of the wave velocity:

$$v = \sqrt{\frac{\text{Restoring force}}{\text{Mass density}}}. \tag{2.55}$$

2.21 Problems for Chap. 2

1. Suppose you see a flash of lightning and then hear the sound of its thunder 5 s later. Assuming a speed of light equals 3×10^8 m/s and a speed of sound equal to 340 m/s:
   (a) Estimate the distance between you and the lightning flash.
   (b) How long did it take for the light of the lightning flash to travel from the lightning to you?
2. Consider a violin string of length 31.6 cm. Waves travel on this string with a velocity of 277 m/s.
   (a) What is the largest period a wave can have if it is to be accommodated by the string as a standing wave?
   (b) Waves of other periods can also exist as standing waves on the string. What are some of these periods
3. Suppose that there are two strings, with one string excited at the fundamental and the other excited at the second harmonic. Both are vibrating at 800 Hz with the same internodal distance of 0.45 m.
   (a) Draw a diagram depicting the two strings at the same scale.
   (b) Show that the wave velocity v is the same for the two strings and calculate its value.
4. Let us assume that a musical instrument has strings of length 76 cm. The player can press the string against the "fingerboard" so as to reduce the length of string that is free to vibrate. Suppose that when the string is fingered a distance of 8 cm from one end, leaving a length of 68 cm free to vibrate, the string vibrates with a frequency of 300 Hz.
   Where would the finger have to be placed to obtain a frequency of 311 Hz?
5. How many antinodes does a string vibrating in its fifth *overtone* have?
6. What is the fundamental frequency of a string that has five antinodes when vibrating at 450 Hz?
7. **Ernst Chladni**, who lived from 1756 to 1827, studied the vibrations of a metal plate by sprinkling sand on its surface and exciting one of its modes of vibration. A mode is distinguished by having **nodal lines**, along which the displacement of the metal plate vanishes. The frequency spectrum is *not* a harmonic series. Generally, vibrations of the plate are composed of

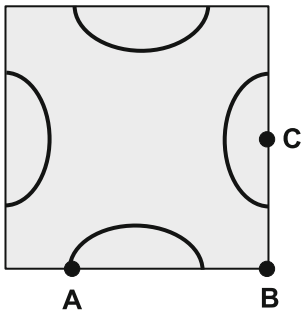


Fig. 2.35 Chladni plate

superpositions of many modes. Excitation of a single mode is facilitated by bowing the plate with a violin bow at some position along the edge and holding the plate at another position along the edge. If the plate is held horizontal, the sand particles dance around as the surface vibrates, being tossed into the air wherever there are vibrations and ultimately settling close to nodal lines. (Where you bow cannot be a node. Why so?) Such plates – in the context of its modes – are called **Chladni plates**.

A pattern of sand on a square Chladni plate is shown in the Fig. 2.35. This pattern could result from

(a) Bowing the plate at A and holding it at B
(b) Bowing the plate at C and holding it at B
(c) Bowing the plate at A and holding it at C
(d) Bowing the plate at C and holding it at A
(e) Bowing the plate at B and holding it at C

8. A pendulum swings back and forth at 20 Hz. Find its period and frequency.
9. (a) What characteristic of the relation between the displacement of an SHO and an applied force is central to its behavior and distinguishes it from other oscillators?
   (b) Specify **at least two** characteristics of the oscillation of a SHO that make it **unique**.
10. (a) Suppose an SHO has a spring constant of 32 N/m and mass of 500 g. Find its vibration frequency.
    (b) How must the mass be changed so as to double the frequency; to halve the frequency?
    (c) How must the **spring constant** be changed so as to double the frequency; to **halve** the frequency?
11. A simple harmonic oscillator (SHO) has a period of 0.002273 s. What is the frequency of the oscillator?
12. A telephone wire electrician needs to determine the tension in a 16 m segment of wire that is suspended between two poles. The wire is known to have a linear

mass density of 0.2 kg/m. He plucks the wire at one end of the segment and finds that the pulse returns in 8 s.

(a) Find the wave velocity.
(b) Find the tension.

13. A piano wire of length 2 m has a mass of 8 g and is kept under a tension of 160 N.

    (a) Find the wave velocity along the wire.
    (b) To double the wave velocity:

        i. The tension can be changed to _____.
        ii. The mass can be changed to _____.

14. (a) A tightrope walker tends to avoid walking at a pace equal to a multiple of the fundamental frequency of the tightrope. **Explain why.**
    Describe what would happen if he were to do so.
    Now suppose that the tightrope is 25 m long, has a mass per unit length of 0.2 kg/m, and has a tension of 2,000 N.
    (b) Calculate the speed of wave propagation along the rope.
    (c) Calculate the rope's fundamental frequency and its two **lowest** overtone frequencies.
15. (a) What is stiffness in a string?
    (b) Does it exist in the absence of tension?
    (c) How does stiffness affect the frequency spectrum of a vibrating string?
    (d) Because of their longer strings, grand pianos need less stretch tuning than upright pianos. Give **two** reasons why this is so.
16. Express the wavelength of a standing wave in terms of the distance between nodes. Now do so in terms of the distance between antinodes.
17. (a) Find the wavelength of a sound wave in water (with a wave velocity of 1,400 m/s) that has a frequency of 10 kHz.
    (b) Find the frequency of a light wave in vacuum that has a wavelength of 0.5 μm ($=5 \times 10^{-7}$ m).
18. (a) Suppose that a guitar string is plucked at its midpoint. Which Fourier components **CANNOT** be excited?
    (b) Repeat the previous question when the string is plucked at a point 1/3 from one end.
19. Explain how a vibrating 440 Hz string can cause a 550 Hz string to vibrate.
20. Find the frequency and period of a periodic wave whose **only** Fourier components are equal to the following:

    (a) 500 Hz, 1,000 Hz.
    (b) 500 Hz, 800 Hz, and 1,000 Hz.

21. What is special about the Fourier frequency spectrum of a periodic wave?

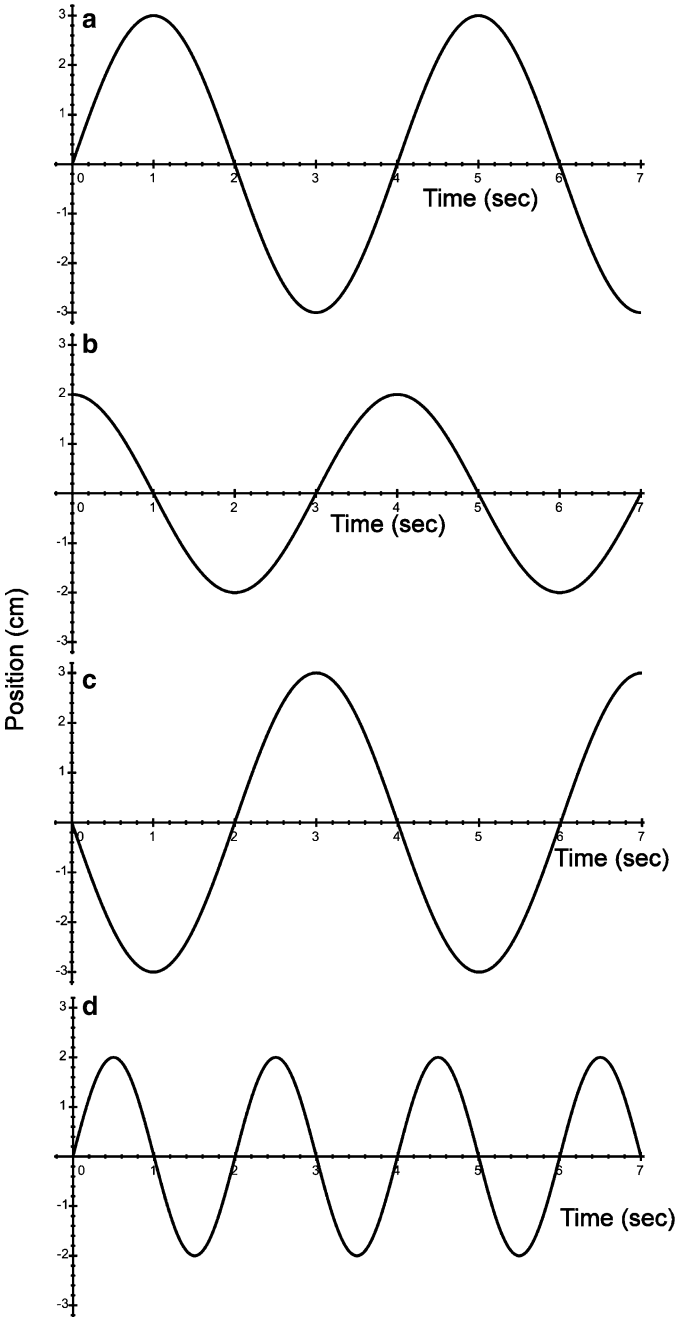


Fig. 2.36 Four different sinusoidal waves

22. Using (2.26), (2.27), and (2.43), derive equation (2.45).
23. In Fig. 2.36 are depicted the displacement vs. time of four wave patterns, labeled A–D:

    (a) What are the respective amplitudes of patterns A and B?
    (b) Which two waves differ **only** in phase?
    (c) What is the frequency of pattern A?

Chapter 3
The Vibrating Air Column

While the vibrating strings of guitars and violins are plainly visible, the sound that they produce in air is invisible. We often associate sound with air because we are used to hearing sounds that reach our ears from the air. We also learn that in the absence of air, sound cannot propagate – movies with sound propagating in outer space, not withstanding. The fact that air is so transparent is not the issue here. Sound travels through liquids such as water and solids such as steel, as well as other gases such as air; nevertheless, we cannot see sound propagating through liquids or solids either. So, what is sound? That is the first subject of this chapter. Once we understand the nature of sound, we will go on to study the modes of vibration of air that is contained in pipes, that is, air columns. These are the basic components for all wind instruments, such as the recorder, flute, and trumpet.

3.1 The Air of Our Atmosphere

Many people think that we cannot see air, not realizing or forgetting that the sky is blue because the air of the atmosphere scatters sunlight. (See Chapter 8, PROPAGATION PHENOMENA, for more details on the origin of the blueness of our sky.) Yet we know that air reveals its presence in the force of blowing wind. To gain a better understanding of the physics and power of invisible and tenuous air, we will begin our study by characterizing air under normal conditions at sea level – in particular, at "standard temperature and pressure" (which is abbreviated as "STP"), that is a temperature of 0°C and a pressure of "one atmosphere".

- Air has weight, represented by a **mass density** ρ of $1.3\,\mathrm{kg/m^3}$. Thus, a room having dimensions of $5 \times 5 \times 2.5$ m, corresponding to a volume of $62.5\,\mathrm{m^3}$, has a mass of $62.5 \times 1.3 = 81$ kg, corresponding to a weight of $81\,\mathrm{kg} \times 2.2\,\mathrm{lbs/kg} = 180$ lbs!
- Air consists of a mixture of various molecules, mostly nitrogen (∼79%), oxygen (∼20%), water vapor (of varied percentage in relation to the relative humidity),

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_3,

and carbon dioxide. These nonspherical molecules have a mean diameter of about 6 Ångstroms, abbreviated as 6 $\mathring{A}$. (1 $\mathring{A}$ $= 10^{-8}$ cm or, equivalently, 1 cm = 100 million $\mathring{A}$.) In one cubic centimeter (cc) there are about 2.7×10^{19} molecules. We say that the number of molecules per unit volume (also referred to as the **number density**) is about 2.7×10^{19} per cc (written as 2.7×10^{19}/cc). In contrast, water has about 10^{23} molecules/cc. (For comparison sake, the number of stars in the entire observable universe is estimated at 10^{22}.)

- The average distance between a molecule of air and a nearest neighbor molecule is about 34 $\mathring{A}$. Its molecular diameter ($\sim$3.5 $\mathring{A}$) is much smaller by a factor of about ten. As a result, it can be shown that on the average a molecule has to travel a relatively great distance, $\sim$670 $\mathring{A}$, about a 100 times its own diameter, before it collides with another molecule. This distance is referred to as the **mean free path** . Thus, most of air is empty![1]

  It is difficult to imagine the minuscule dimensions of molecules. Therefore, to get a sense of the proportions involved, suppose that a molecule has a diameter on the order of a football player – say a sphere one yard in diameter. Now imagine a checker board laid out without boundaries, with each side being 10 yards; there would be an infinite number of squares. Next, place one football player randomly within each square. On the average, the nearest neighbor to a football player will then be 10 yards, just ten times the diameter of one player. Finally, imagine that one football player starts running in a straight line and in a random direction. How far would the player have to run on the average till he runs into another player? The answer is 100 yards – which is the square of their nearest neighbor distance apart divided by the diameter of a football player.[2] For a gas in a three-dimensional volume, with all molecules moving about randomly, the mean free path is about (distance between nearest neighbors)^3/(diameter)^2.[3]
- Gas molecules are in constant random motion, with an average speed of about 300 m/s ($\sim$1,100 km/h $\sim$1,000 ft/s). This speed happens to be close to the speed of sound in air.
- If most of the space in air is devoid of molecules, how then can the air sustain a sound wave that requires the propagation of density variations through the air? What is the level of interaction between molecules of air. One measure is the rate of collisions between molecules: In 1 cm^3 of air there are on the order of 10^{28} collisions between molecules per second.

---

[1] Another way to appreciate this observation is to note that if we were to take a volume of air and compress it so that all the molecules are just touching each other, the volume would be reduced by a factor of $10^3 = 1{,}000$.

[2] The following website ([12-26-2010] http://comp.uark.edu/~jgeabana/mol_dyn/) has an animation that shows a collection of particles moving in a square chamber. You can choose the number and size of the particles. You can also run the animation slow enough to be able to follow a single molecule to see how far it travels before colliding with another molecule.

[3] The exact expression for the mean free path is $(\sqrt{2}\,\pi \times \text{number density} \times \text{diameter}^2)^{-1}$. Also, note that if we were to distribute football players in a three-dimensional array of boxes, each side being 10 yards. the mean free path turns out to be about 1,000 yards or a bit over one-half of a mile!

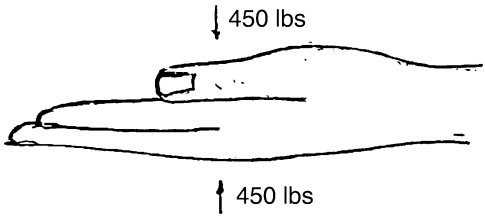


Fig. 3.1 Force by air on hand

- We barely notice the presence of air but easily note that we need its oxygen to survive. Yet air is not inert even when the wind is not blowing. There is **air pressure**: A value of **one atmosphere** pressure corresponds to a pressure of 10^5 N/m^2 $\sim$15 lbs/sq-in; that is, every square inch of flat surface experiences a force of 15 lbs, **whatever its orientation**, be it horizontal, vertical, or otherwise. Mathematically, we write

$$p = \frac{F}{A}. \tag{3.1}$$

  Alternatively, we write:

$$\text{Force} = \text{Pressure} \times \text{Area} \qquad \text{or} \qquad F = pA. \tag{3.2}$$

  Thus, for example, consider the palm of my hand, which has an area of $\sim$30 sq-in. See Fig. 3.1. When held horizontal, the top of my palm experiences a force of $30 \times 15 = 450$ lbs downward, while the bottom experiences a force of 450 lbs upward. These two forces cancel each other, so that I do not have to put any effort into preventing my hand from being moved by the air.

  My abdomen has dimensions on the order of $15''$ by $12''$, corresponding to an area of 180 sq-in. The force of the air on my abdomen is therefore about $180 \times 15 = 2{,}700$ lbs! Why don't I feel this tremendous force? Why doesn't my hand get crushed? The reason is that there is pressure within the tissues of my hand, and within my nerve cells in particular, that prevents any crushing from taking place. Furthermore, my nerves function in such a way that this normal background pressure of one atmosphere does not produce any sensation. Imagine if it were otherwise!! We will see that there is a background noise impinging upon our ears; but we are simply insensitive to it.
- What is the source of the force associated with air pressure? It is the rapid rate at which molecules of air collide with a surface, as depicted in Fig. 3.2. In fact, there are about 10^{23} collisions per second on each square centimeter of area. To help you sense the nature of this force, imagine rain drops from a torrential rainstorm striking your hand at a rapid rate. You would feel the force they exert; but that force would normally be much less than 15 lbs/sq-in: While the drops have much greater mass than does a molecule, the collision rate of the rain drops would be much less than that for the air molecules (Fig. 3.2).

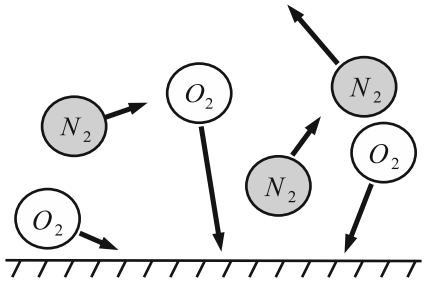


Fig. 3.2 Molecules colliding with wall

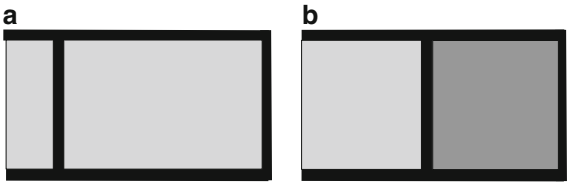


Fig. 3.3 (**a**) A chamber of air fitted with a piston. (**b**) The piston compresses the air in the chamber

- At a given temperature, the pressure is proportional to the density. In Fig. 3.3, we depict a chamber of air fitted with a piston. Outside the chamber is air at one atmosphere pressure. In Fig. 3.3a, the air in the chamber has a pressure of one atmosphere.

  The force on the piston from the outside air to the left of the piston then balances the force of the air inside the chamber. In Fig. 3.3b, the air in the chamber has a pressure of 1.3 atm. The density of air in this chamber is 30% greater than the density of air in the chamber of Fig. 3.3a. (Note the difference in density of shading.) In this second case, the force on the piston from the inside is greater than the force from the outside and the piston would tend to move to the left, corresponding to a reduction in the air density and hence the pressure within the chamber. The **difference** in pressure acts as a **restoring force**.[4]

3.1.1 Generating a Sound Pulse

Consider again the chamber as in Fig. 3.3a. What would happen if the piston was moved **suddenly** to the right and held fixed in position? (See Fig. 3.4.)

The air in front of the piston will be compressed locally, thus increasing the pressure locally. The nonuniformity in pressure will lead to a tendency for the

[4]Here is a beautiful animation that displays the collisions of molecules on a piston. (12-26-2010, http://wilsonspirit.com/) You can vary the number of molecules in the chamber. You can move the piston so as to change the volume so as to change the collision rate with the piston as well as the pressure. Finally, you can vary the temperature so that the speed of the molecules varies. A shortcoming of the animation is that it omits intermolecular collisions.

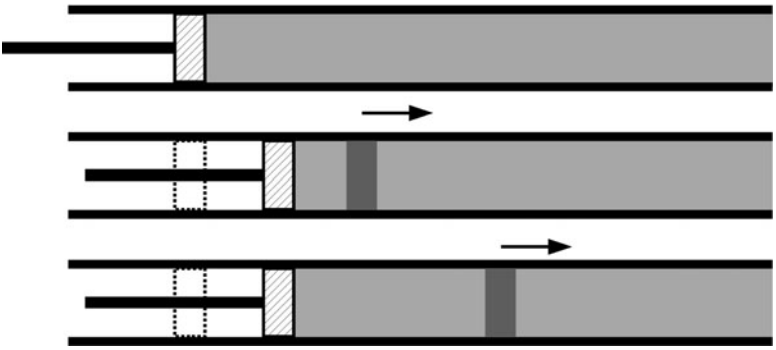


Fig. 3.4 A sudden move of the piston creates a pulse

compressed region to expand outwardly to the right. The net result will be that a compression pulse will move to the right as depicted in Fig. 3.4 to the right. This pulse will strike the opposite end of the chamber and be reflected back, followed by repeated cycles back and forth. In time, because of attenuation (see Sect. 4.8), the pulse will spread out in width and diminish in amplitude, eventually disappearing. The final state will be a chamber with air at a uniform density and pressure, corresponding to the current reduced volume of the chamber.

3.1.2 Digression on Pushing a Block of Wood

When you start pushing a block of material at one of its ends, you might assume that all of the material begins to move at once. The fact is that it takes time for the entire piece of material to respond to your push. Your initial push produces a compression pulse that travels through the block at the speed of sound. Eventually, this pulse is spread out, so that the block moves as a whole. If the block is made of a hard wood such as oak and the pulse is traveling in a direction parallel to the fibers, the speed of sound and hence the pulse is ~4,000 m/s. If the block is 10 cm long, the pulse takes $0.10/4{,}100 = 0.00003$ s $= 30$ microseconds (30 μs) to travel the length of the block.

3.2 The Nature of Sound Waves in Air

We can now understand what the essence of a sound wave is: **A sound wave represents the propagation in a material of a nonuniformity in the density**. Nonuniform pressure provides the restoring force toward the equilibrium state of uniform density. As an example, suppose that the above piston undergoes sinusoidal motion – that is, simple harmonic motion (SHM) – at a frequency f. In Fig. 3.5, we

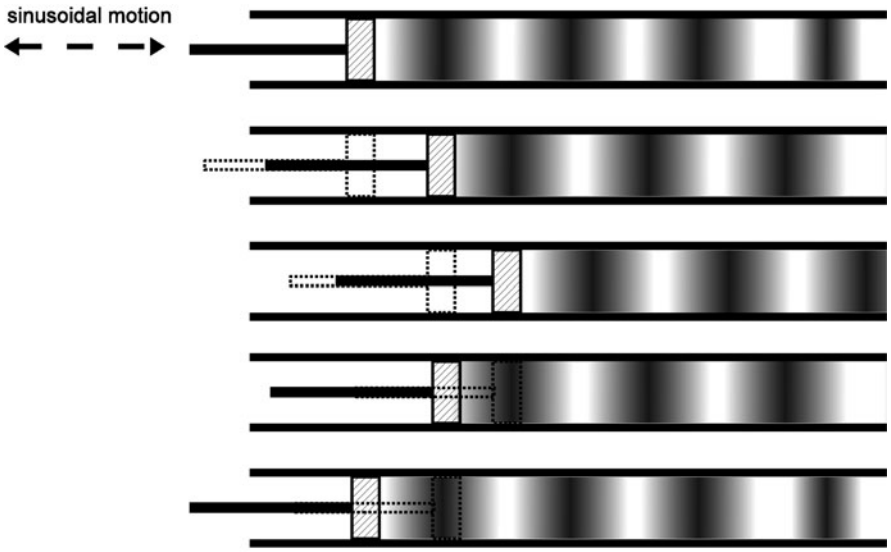


Fig. 3.5 A sinusoidal wave in air produced by a piston oscillating sinusoidally

depict the resulting sinusoidal sound wave that propagates down the chamber. This motion can be produced by attaching a vibrating tuning fork to the piston. Regions having a high density and pressure are called **condensations**, while regions with low density and pressure are called **rarefactions**.

Note that the action of the piston is similar to that of the cone of a loudspeaker, which vibrates in some pattern time. If the cone vibrates sinusoidally, it produces a sinusoidal wave in space. Generally, the pattern in time produces a sound wave with a corresponding pattern in space.

Since the motion of air in a sound wave is in the same direction as the direction of propagation of the wave. The wave is said to be **longitudinal**, in contrast to the transverse waves along a stretched string.

Indicated too in the figure are the wavelength λ and the wave forms at various fractions of the period $T = 1/f$. After a time interval of one period, the wave progresses a distance of one wavelength.

Equations (2.23) and (2.24) of Chap. 2 hold:

$$v = \frac{\lambda}{T} = \lambda f. \tag{3.3}$$

The normal person can hear pure sinusoidal sounds having a frequency ranging from about 20 to 20,000 Hz.[5]

[5] As a person gets older and/or subjects himself/herself to loud sounds such as rock music, the upper limit goes down. In the author's testing of students from 1973 to 2000, the limit for most students has dropped from about 22,000 to about 18,000 Hz.

Let us calculate the corresponding range of wavelengths of audible sound traveling in air. We will use the velocity of sound at a temperature of 15°C – 340 m/s. For the minimum frequency of 20 Hz, we obtain

$$\lambda = \frac{v}{f} = \frac{340}{20} = 17\,\text{m}. \tag{3.4}$$

The wavelength corresponding to a frequency of 20,000 Hz is just $20{,}000/20 = 1{,}000$ times smaller, that is 17 mm =1.7 cm.

3.3 Characterizing a Sound Wave

A sound wave can be characterized by the **change in pressure** from the ambient equilibrium pressure, which is usually about one atmosphere. The amplitude of the variation of the change in pressure is called the **sound pressure**. (See Fig. 3.6.) The softest sound that can be heard has a sound pressure of only one-ten-billionth (10^{-10}) of an atmosphere! A sound can produce pain if the sound pressure is one-ten-thousandth (10^{-4}) of an atmosphere. Thus, the ratio of the largest sound pressure tolerable to the smallest discernable is $10^{-4}/10^{-10} = 10^6$, or a million to one! Alternatively, a sound wave can be characterized by the change in mass density from the equilibrium density. We then refer to the **sound density** of the wave.

One can also characterize a sound wave by the **displacement** of the air, although this is not done in practice. The corresponding range of displacements is incredibly small – ranging from 0.1 Å (about 1/30th the size of an atom (!)), to 1/100th of a millimeter (about twice the length of a bacterium).

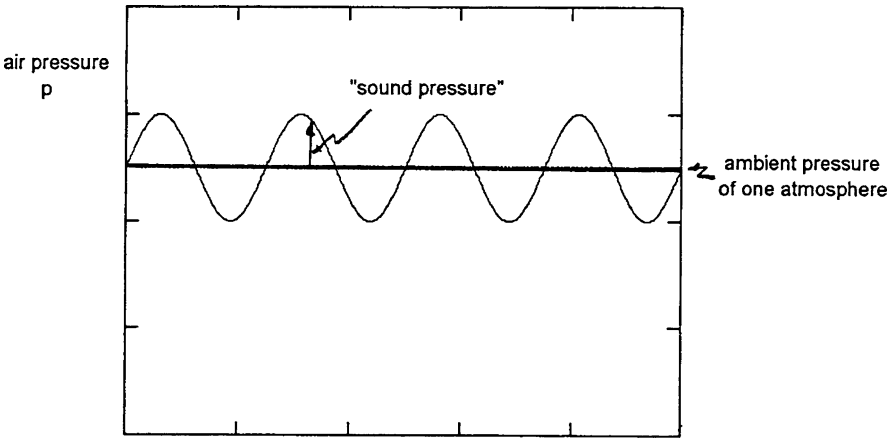


Fig. 3.6 Sound pressure

> **Note**
>
> Sound waves in liquids and solids are essentially the same as sound waves in a gas such as air. The major difference lies in the density of material: Solids and liquids are typically about 1,000 times more dense than air at STP. Nevertheless, we will see that the velocity of sound is about the same order of magnitude for all materials.

3.4 Visualizing a Sound Wave

Suppose that we have a sound wave and we want to be able to visualize its corresponding wave pattern. A neat way to do so is summarized in Fig. 3.7.

The figure begins on the left with a source of sound – the **signal source** – that is produced by a generator of an electric signal that corresponds to the sound wave. One possible signal source is a microphone into which we might sing. Another can be a **wave generator**, which is an electronic device that produces periodic electric signals having a frequency that we can control. The periodic pattern of the signal is usually a sine wave. (Other patterns include a **sawtooth wave** and a **rectangular wave**. The names have obvious meanings.) Next, the electric signal from the source is fed into both an **amplifier** and an **oscilloscope**. The purpose of the amplifier is to increase the amplitude of the signal manyfold, so that the resulting signal can drive the loudspeaker. The oscilloscope gives us a visual image on the screen of the pattern of the periodic electric signal. This visual image ideally represents the wave pattern of the sound wave that we hear produced by the loudspeaker.

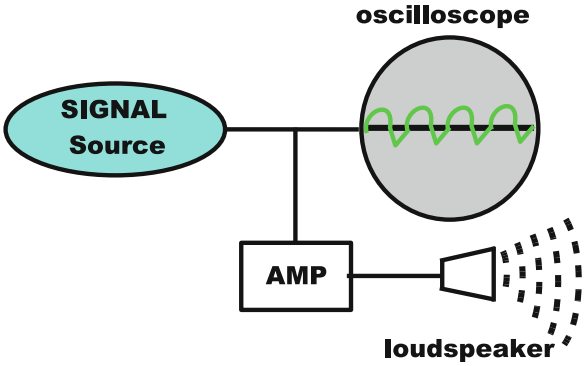


Fig. 3.7 Studying a sound wave

3.5 The Velocity of Sound

What determines the velocity of sound? Recall that generally, the wave velocity of all types of waves depends upon the square root of the ratio of an effective restoring force and an effective mass. In the case of a sound wave, that effective mass is the mass density ρ, which is expressible in kg/m^3. (For example, the mass density of water is 1 g/cc = 1,000 kg/m^3, while the mass density of air at STP is 0.3 kg/m^3.) The restoring force for a sound wave is known as the **bulk modulus**, which has the symbol B: This parameter tells us the relative **decrease** in volume of a sample in response to an **increase** in the pressure. In some situations, the temperature remains fixed during the process; in others, the sample is insulated and the temperature will usually rise. For small relative changes in volume, the relative change in volume $\Delta V/V$ is generally proportional to the change in pressure Δp. (This is the analog of **Hooke's Law**: displacement of spring $\propto$ force.) We write

$$\frac{\Delta V}{V} = -\frac{\Delta p}{B}. \tag{3.5}$$

The **minus** sign is inserted because an **increase** in the pressure leads to a **decrease** in the volume. The ratio $\Delta V/V$ is referred to as the **relative change in the volume**. By multiplying by 100, we obtain the percent change.

As the volume increases at fixed amount of matter, the mass density decreases. Correspondingly the relative change in mass density ρ, given by $\Delta\rho/\rho$, is

$$\frac{\Delta \rho}{\rho} = +\frac{\Delta p}{B}. \tag{3.6}$$

We see that the $(-)$ sign of equation (3.6) is simply replaced by a $(+)$ sign.

> **Note**
>
> In the case of air, the bulk modulus is about equal to the pressure itself, so that $\Delta V/V \sim -\Delta p/p$. Thus, a 1% increase in pressure leads to about a 1% decrease in volume. On the other hand, the bulk modulus of liquids and solids is much greater than that for gases: It requires a much greater increase in pressure to produce a given reduction in the volume of a liquid or solid; air is much more compressible. For example, the value of the bulk modulus for water is 20,000 atm, while for steel it is 1,400,000 atm! Thus, to produce a 1% decrease in the volume of a sample of water requires an increase in pressure of $20{,}000 \times 0.01 = 200$ atm, while for a sample of steel a pressure of $1{,}400{,}000 \times 0.01 = 14{,}000$ atm is required!

Down at the deepest depths of ocean, the pressure is about 1,000 atm, so that the density of the water is increased by a factor of $\Delta p/B = 1{,}000/20{,}000 = 1/20$, corresponding to 5%.

One can show that generally the velocity of sound is given by

$$v = \sqrt{\frac{B}{\rho}}. \tag{3.7}$$

In using this equation, it is essential to express both B and D in terms of a consistent set of units, such as N/m^2 and kg/m^3, respectively.

Recall from Chap. 2 that the wave velocity for a stretched string is the square root of the ratio of a force parameter (that restores the string to its straight line shape) and a mass parameter. Here, the bulk modulus B is the associated with the increase in pressure when the air is compressed; it reflects the effective restoring force. The mass density ρ is the mass parameter.

Let us then calculate the speed of sound in water. We recall that the mass density of water is 1 g/cm^3=1,000 kg/m^3 and that 1 atm=10^5 N/m^2, so that $B = 20{,}000 \times 10^5$ N/m^2. Thus,

$$v = \sqrt{\frac{20{,}000 \times 10^5}{1{,}000}} = 1{,}400\,\text{m/s}. \tag{3.8}$$

You can carry out a similar calculation for air ($B = 1.4$ atm, $\rho = 1.3$ kg/m^3) and for steel ($B = 1{,}400{,}000$ atm and $\rho = 7{,}900$ kg/m^3).

3.5.1 Temperature Dependence of Speed of Sound in Air

The speed of sound in a gas increases with increasing temperature. At 0°C, the speed is 332 m/s. To obtain the approximate speed for a temperature between ~ -50°C and $\sim +50$°C, simply add 0.61 multiplied by the temperature in °C. Thus,

$$v\left(\frac{\text{m}}{\text{s}}\right) = 332 + 0.61 \times \text{Temperature (°C)}. \tag{3.9}$$

For example, at a temperature of 20°C, the speed of sound is

$$v = 332 + 0.61 \times 20 = 332 + 12 = 344\,\text{m/s}.$$

The relative change is $12/332 = 0.04$, or 4%. Consider the effect on a pitch pipe, which is set to produce a single musical note. We will see in the next section that the frequency of the sound produced by a pitch pipe is proportional to the wave velocity. Thus, the frequency will increase by close to 4%. In Chap. 11, we will see that this change in frequency corresponds to about 3/5 of a **semitone** interval. Thus, a musical note of A will change to a note close to $A\#$. For this reason, wind instrumentalists must tune their instrument as their breath heats up the air within – if at all possible.

3.6 Standing Waves in an Air Column

Consider a long pipe, that is, one whose diameter is much less than its length ℓ, and which is open at both ends to the outside air. The pipe is thus filled with air, so that we have a column of air. It is possible to excite modes of vibration, standing waves, in a column of air.[6] These standing waves are the source of sound as a result of the variations of the pressure that the air column exerts on the air outside the column. A standing wave consists of a sinusoidal pattern of the sound pressure, whose amplitude oscillates sinusoidally in time, as in the case of a standing wave along a stretched string. In essence, air rushes in and out of the two ends of the pipe. See Fig. 3.8.

The air **displacement** has an **antinode** at an **open end** of a pipe. A small amount of air rushes in and out of the pipe. In addition, because of the huge volume of outside air, the outside air acts like a cushion that prevents the pressure at an opening from being different from that of the outside air. The **sound pressure** has a **node** at an open end of the pipe.

In the case of the *fundamental* mode, there is no motion of air at the center. That is, the displacement has a **node** at the center. Correspondingly, the density and, hence, pressure are maxima there. In other words, the **sound pressure** and **density** have an antinode at the center. The variation of the sound pressure with position in the pipe is thus the same as that for the displacement of a vibrating string with fixed ends, as seen in Fig. 3.9.

Analysis of the standing wave vibrations of a pipe leads to the same formula for the fundamental frequency as that for the string, namely

$$f_1 = \frac{v}{2\ell}, \tag{3.10}$$

where v is the wave velocity, here the speed of sound in air.

Looking beyond the fundamental mode, the mode frequency spectrum is a harmonic series, as with the string (see (2.26)):

$$f_1 = \frac{v}{2\ell}, \qquad f_2 = 2\frac{v}{2\ell} = \frac{v}{\ell}, \qquad f_3 = 3\frac{v}{2\ell}, \qquad f_4 = 4\frac{v}{2\ell} = \frac{2v}{\ell}, \qquad \ldots \tag{3.11}$$

Fig. 3.8 Motion of air in a pipe for the fundamental

[6] We have restricted our attention to pipes with a relatively small diameter only because in this case the frequency spectrum of the modes is a harmonic series. The larger the ratio of the diameter to the length, the greater the deviation of the spectrum from a harmonic series.

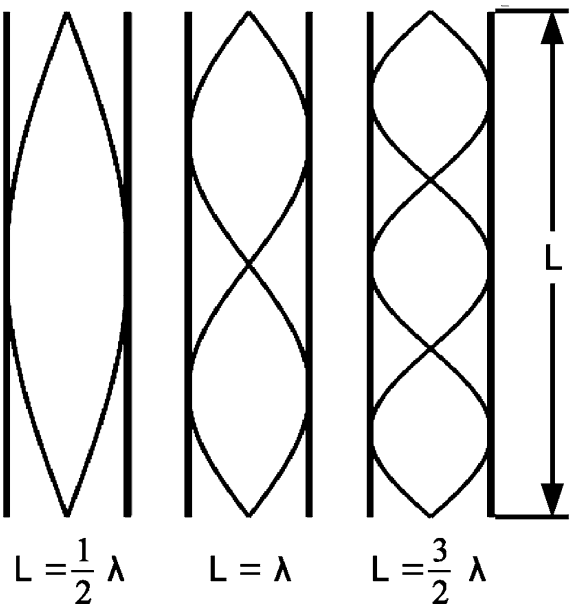


Fig. 3.9 Wave patterns of the sound pressure and sound density for the first three harmonics of an open pipe

or

$$f_n = n\frac{v}{2\ell}, \qquad n = 1, 2, 3, 4, \ldots \tag{3.12}$$

Sample Problem 3-1

Consider a person playing a flute that is open at both ends, having a length of 60 cm, with air at a temperature of 20°C. Find the fundamental frequency.

Solution

Recall that we found that the speed of sound air at 20°C is 344 m/s. Hence

$$f_1 = \frac{v}{2\ell} = \frac{344}{2(0.60)} = 287\,\text{Hz}.$$

Now suppose that the temperature increases to 30°C by virtue of a person's warm breath. The 10° increase in temperature leads to an increase of the speed of sound by $0.6(10) = 6$ m/s, that is, to 350 m/s. The new fundamental frequency is then

$$f_1 = 350/2(0.60) = 292\,\text{Hz}.$$

Alternatively, the calculation can be carried out in this illuminating way: The change in the frequency is given by Δf and can be expressed as

$$\Delta f = \frac{\Delta v}{2\ell}, \tag{3.13}$$

where Δv is the change in the speed of sound. Then

$$\frac{\Delta f_1}{f_1} = \frac{\Delta v/2\ell}{v/2\ell} = \frac{\Delta v}{v}. \tag{3.14}$$

This equation tells us that the **relative change in the frequency** is equal to the **relative change in the speed of sound**. Correspondingly the percent changes are equal. In our problem, the relative change in speed is $6/344 = 0.017$, so that the relative change in the frequency is also 0.017. The original frequency is 287 Hz. Hence the change in frequency is $0.017(287) = 4.9$ Hz and the new frequency is $287 + 4.9 = 292$ Hz.

What is the benefit of going through this alternative approach? It is the following: The relative change in frequency, here 0.017, holds for any pipe, **whatever its length**, as well as for any particular mode. Thus, we need not know the original length of the pipe to determine the effect of a temperature change. Furthermore, as we will see in Chap. 11, the relative change in frequency is directly related to the change in pitch. For example, a change of 6% corresponds to a half tone.

An aside: An increase in temperature will also lead to an increase in the length of the pipe. The actual increase depends upon the material out of which the pipe is made. That increase is minuscule. For steel, it amounts to a relative change of only one part in 100,000 *per* °C, or one part in 10,000 (0.01%) for the temperature increase of 10° in our above problem. This **increase** in length would by itself lead to a **decrease** in the frequency because the frequency is **inversely** proportional to the length. However, that decrease is imperceptibly small.

Sample Problem 3-2

What should the length of a pipe be to obtain a fundamental frequency of 20 Hz, with air at 20°C?

Solution

We first note that

$$f_1 = \frac{v}{2\ell} \qquad \longrightarrow \qquad \ell = \frac{v}{2f_1}. \tag{3.15}$$

Thus, given a speed of 344 m/s, we obtain

$$\ell = 344/2(20) = 8.6\,\text{m} \sim 28\,\text{ft}!$$

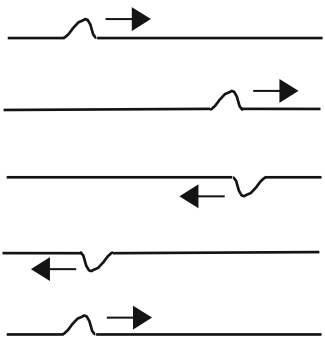


Fig. 3.10 A pulse along a string fixed at both ends

3.6.1 Standing Waves in a Closed Pipe

One could use a 28 ft organ pipe to produce a 20 Hz note. However, often such a length is quite unwieldy. How can one avoid it? The solution lies in using a pipe which is **open** at one end but **closed** at the other. We will refer to such a pipe as a **closed pipe**, even though one end of such a pipe is open. We will refer to a pipe that is open at both ends as an **open pipe**.

At the closed end of a *closed pipe*, the air cannot move. Therefore, the displacement of air flow has a **node**. On the other hand, the density has a maximum variation and therefore has an **antinode**. Analysis leads to the result that the fundamental frequency of a **closed pipe** is half that for an open pipe of the same length.

Below we present a simple way of understanding this result. It is based on the way in which a pulse of sound is reflected off an open end and a closed end of a pipe. The fundamental frequency is the same as the frequency of a pulse that moves back and forth down the length of a pipe.

We start by considering a pulse along a taut string, that is fixed at both ends. See Fig. 3.10. The figure depicts the displacement of the string at various times. We note that the direction of the displacement is reversed upon reflection off an end. A cycle requires but two traversals of the pulse along the length of the string.

Examine the figure very carefully. You will note that the reflected pulse has a pattern that differs from the incident pulse in two ways: First, the pulse is flipped upside down. Second, the steeper side of the incident pulse is on the right, while the steeper side of the reflected pulse is on the left. Thus, the front of the incident pulse remains the front of the reflected pulse.

The distance traveled by the pulse in one cycle is 2ℓ. The corresponding frequency is then

$$f = \frac{v}{2\ell}. \tag{3.16}$$

Now we return to the behavior of a pulse in a pipe. When a condensation travels down a pipe and reaches an open end, the air rushes out and the reflected wave is a rarefaction. If the condensation reaches a closed end, the reflected wave is a

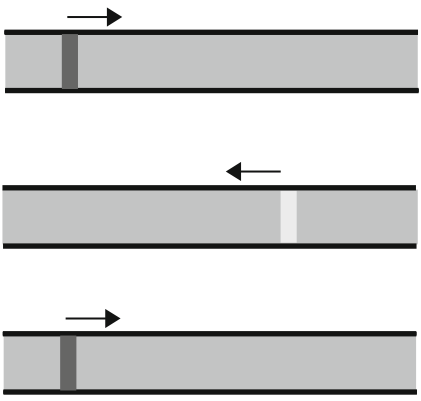


Fig. 3.11 A pulse along a pipe open at both ends

condensation. On the other hand, when a rarefaction travels down a pipe and reaches an open end, the air rushes in, so that the reflected wave becomes a compression. If the rarefaction reaches a closed end, the reflected wave is a rarefaction. (See Fig. 3.11.) Thus, the spectra of a taut string fixed at both ends and an open pipe are both harmonic series.

The distance traveled by the pulse in one cycle is 2ℓ so that the frequency is

$$f = \frac{v}{2\ell}. \tag{3.17}$$

On the other hand, when a sound pressure pulse is reflected off a closed end, its character is not changed. A condensation is reflected as a condensation and a rarefaction is reflected as a rarefaction. The result is that a cycle requires four traversals and the fundamental frequency is half that of the open pipe. See Fig. 3.12.

Correspondingly, for a given frequency, we would need only half the length of pipe, or 14 ft for our problem.

The distance traveled by the pulse in one cycle is 4ℓ so that the frequency is

$$f = \frac{v}{4\ell}. \tag{3.18}$$

The frequency spectrum of a closed pipe can be shown to consist of all the **odd harmonics** of the fundamental frequency, that is

$$f_1 = \frac{v}{4\ell}, \qquad f_2 = 3\frac{v}{4\ell}, \qquad f_3 = 5\frac{v}{4\ell}, \qquad f_4 = 7\frac{v}{4\ell}, \qquad \ldots \tag{3.19}$$

Thus, for a 14-ft pipe and sound velocity of 345 m/s, the fundamental frequency will be 20 Hz. The overtones will be

$$3 \times 20 = 60\,\text{Hz}, \qquad 5 \times 20 = 100\,\text{Hz}, \qquad 7 \times 20 = 140\,\text{Hz}, \ldots$$

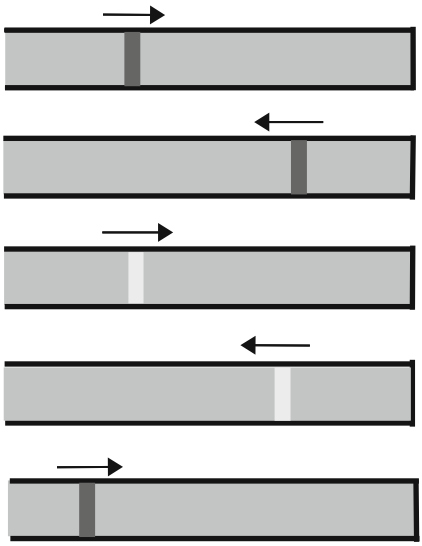


Fig. 3.12 A pulse along a pipe open at one end but closed at the other

Table 3.1 Summary of pipe properties and their behavior

Parameter	Open end	Closed end
Sound pressure	Node	Antinode
Sound density	Node	Antinode
Displacement	Antinode	Node

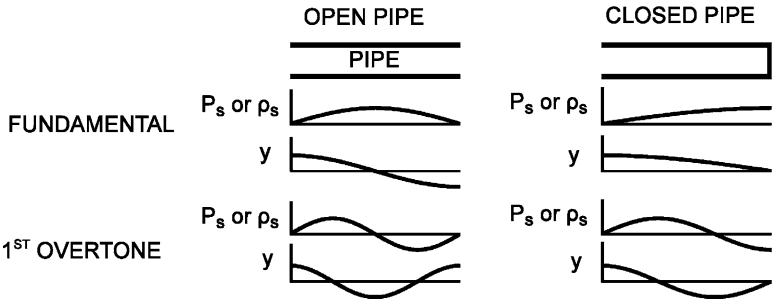


Fig. 3.13 Wave patterns of the sound pressure, sound density, and displacement for the fundamental and the first overtone for both the open pipe and the closed pipe

In Table 3.1, we summarize the conditions that hold at the boundaries of a vibrating air column for the three parameters: *sound pressure*, *sound density*, and *displacement* of fluid.

In Fig. 3.13, we display two graphs for each of the following: the fundamental and the first overtone, for both an open pipe and a closed pipe. One graph displays the variation of sound pressure p_s and change in density ρ_s due to the sound wave; the other graph displays the displacement y of the gas.

3.6.2 End Correction for Modes in a Pipe

The formulas that we have exhibited for the frequencies of modes in open and closed pipes are in fact approximations that hold to the extent that the length ℓ of a pipe is much longer than its radius R, that is, when $\ell \gg R$. A better approximation is to replace the length ℓ in the formulas so as to include so-called end corrections: For each open end, we add $0.3R$ to ℓ.

For a closed pipe, which has one open end, the frequencies are

$$f_1 = \frac{v}{4(\ell + 0.3R)}, \qquad f_2 = 3\frac{v}{4(\ell + 0.3R)}, \qquad \ldots. \tag{3.20}$$

However, for an open pipe, which has two open ends, the frequencies are

$$f_1 = \frac{v}{2(\ell + 0.6R)}, \qquad f_2 = 2\frac{v}{2(\ell + 0.6R)}, \qquad \ldots. \tag{3.21}$$

3.7 Magic in a Cup of Cocoa**

Suppose you make yourself a cup of cocoa, mixing the cocoa powder in with hot water. You stir well and then tap the top of the cup. You will hear a sound with a well-defined pitch. Now you tap the top repeatedly. You will find that as time progresses, the pitch of the sound rises steadily so much so that the final maximum pitch corresponds to more than double the original frequency. If you stir the cocoa you can repeat the above procedure with the same consequences. The question is why does this happen?

We can quickly guess that the sound corresponds to a mode of vibration of sound within the volume of cocoa. This can be checked by noting that the empty cup produces no sound in the range of the sound above. The frequency depends upon the shape of the volume – which does not change – and the velocity of sound. Therefore, we surmise that the effect of stirring the cocoa lowers the velocity of sound. Two possibilities present themselves: Either the stirring mixes up microscopic particles of undissolved cocoa or the stirring produces microscopic bubbles of air within.

Let us focus first on the cocoa particles as a possible explanation. We recall that the sound velocity depends upon the ratio of the bulk modulus and the mass density. This ratio must therefore change by a factor of 4 to account for a doubling of the frequency! The particles would increase the average density of liquid, thus lowering the sound velocity; that is a good sign. However, the concentration of particles is small. (Remember that you typically put into the cup about two tablespoons in 8 oz of water and the cocoa powder is fluffy so that it takes up much less than two tablespoons.) Therefore, the effect of increased density is much too small to account

for the observations. Moreover, the relatively higher bulk modulus compared to that of water, would *increase* the sound velocity. We therefore conclude that the stirring up of cocoa particles is not the explanation.

We next focus on the possibility that stirring produces microscopic air bubbles, which eventually rise and leave the cup. Their presence would *lower* the average density and therefore raise the frequency – but only by a small amount because of their low concentration. However, they have a much lower bulk modulus, by a factor of about 100,000! Furthermore, when a sound wave is compressing the cocoa over a typically volume of about half a wavelength, the fractional *change* in volume is extremely small. As a consequence, all of this reduction in volume can be taken up by the small volume of bubbles even if the bubbles take up a small fraction of the original volume. The effective bulk modulus of the cocoa is then dominated by the bubbles. It can be shown that all we need is a volume of bubbles equal to one part in 1,000 to change the frequency by a factor of 2!

With all this descriptive and semi-quantitative analysis, how can we be confident that we have arrived at the correct explanation? The answer is that a detailed theory leads to quantitative agreement with experimental observations. This is the final test in our hypothesis.

3.8 Terms

- Adiabatic change
- Ångstrom
- Atmospheric pressure
- Boundary condition
- Bulk modulus
- End correction
- Isothermal change
- Open-closed pipe
- Open-open pipe
- Pressure (force per area)
- Sound density change
- Sound pressure
- Standard temperature and pressure (STP)

3.9 Important Equations

Pressure defined:

$$p = \frac{F}{A}. \tag{3.22}$$

Fundamental relation between wavelength, frequency, and wave velocity:

$$\lambda = \frac{v}{f} \tag{3.23}$$

Bulk modulus defined:

$$\frac{\Delta V}{V} = -\frac{\Delta p}{B}. \tag{3.24}$$

Speed of sound in an isotropic material:

$$v = \sqrt{\frac{B}{\rho}}. \tag{3.25}$$

Speed of sound in air:

$$v\left(\frac{\text{m}}{\text{s}}\right) = 332 + 0.61 \times \text{temperature } (^\circ\text{C}). \tag{3.26}$$

Frequency spectrum for a pipe that is open at both ends:

$$f_1 = \frac{v}{2\ell}, f_2 = 2\frac{v}{2\ell} = \frac{v}{\ell}, f_3 = 3\frac{v}{2\ell}, \ldots. \tag{3.27}$$

or

$$f_n = \frac{nv}{2\ell}, \qquad n = 1, 2, 3, 4, \ldots. \tag{3.28}$$

Frequency spectrum for a pipe that is closed at one end:

$$f_1 = 1\frac{v}{4\ell}, \qquad f_2 = 3\frac{v}{4\ell}, \qquad f_3 = 5\frac{v}{4\ell}, \qquad f_4 = 7\frac{v}{4\ell}, \qquad \ldots. \tag{3.29}$$

3.10 Problems for Chap. 3

1. Find the force acting on an area of dimensions $3'' \times 4''$ at the bottom of the deepest part of the Pacific Ocean, at 10,000 m(= 10 km), where the water pressure is 1,000 atm.
2. What is the "sound pressure"?
3. What is the effective restoring force with respect to the velocity of a sound wave?
4. Find the mass and weight (in lbs.) of the volume of air in a room with the dimensions $4 \times 5 \times 6$ m.
5. The ear is most sensitive to sinusoidal sounds ("pure tones") having a frequency of about 3,000 Hz. Calculate the wavelength of a sound wave with this frequency if the wave is traveling in air with a sound velocity of 340 m/s. Calculate the wavelength if the sound wave is traveling in water with a speed of 1,400 m/s.
6. Calculate the sound velocity in steel, whose bulk modulus is 1.4×10^6 atm and mass density is 7,900 kg/m^3.

7. An open cylindrical pipe is 0.05 m long. If you ignore the end correction factor, what would be the frequency of the fundamental and the first overtone? Assume a temperature of 15°C. Repeat the above for a closed (half-open, half-closed) pipe.
8. A pipe is sounded at room temperature (20°C). It is observed that in the range 1,000–2,000 Hz, the pipe can be made to oscillate only at the frequencies 1,000, 1,400, and 1,800 Hz.

   (a) Is the pipe an open or a closed pipe?
   (b) Based on your answer, what is the length of the pipe?

9. **The Helmholtz Resonator**
   Have you ever blown over the top of an open soda bottle? The sound is like a hoarse human voice. The bottle is acting like a **Helmholtz resonator**, named after **Hermann von Helmholtz**[7] who first studied this device. A real bottle with air within has many modes of vibration associated with the motion of air – standing waves that are analogous to waves in a pipe. On the other hand, the so-called Helmholtz resonator refers to a very specific mode wherein the air that is concentrated in the mouth of the bottle moves into and out of the bottle. It is analogous to the so-called air resonance of a violin that involves air moving into and out of the interior of the body of a violin through its **f-holes**.

   We see in Figs. 3.15 and 3.16 an actual Helmholtz resonator followed by a schematic of a simple model of a Helmholtz resonator.

   The actual shape of the bottle is not relevant; all that matters is that there is a narrow mouth whose volume is much less than the volume V of the bulk of the bottle. The mass of air that oscillates is highlighted in a dark gray shade. The air in the narrow mouth of the bottle acts as the mass of the oscillator, while the bulk of the air in the bottle acts as the spring constant. Note that the body of air within a stringed instrument such as a violin acts like a Helmholtz resonator. The corresponding mode of vibration is called the **main air resonance** of the stringed instrument.

   Three geometrical parameters determine the frequency: the volume V of the body of air; the area A of the mouth; and the length ℓ. The fourth parameter is the speed of sound in air, v. If the radius of the mouth is much less than the length ℓ of the mouth, the frequency of a Helmholtz resonator is given by

$$f = \frac{v}{2\pi}\sqrt{\frac{A}{V\ell}}. \tag{3.30}$$

[7] We will see Helmholtz's name later in the text in connection with his study of hearing. He is the inventor of the ophthalmoscope that is used to examine the interior of the eye. He is also famous for his major contribution in the development of the **Principle of Conservation of Energy**, to be discussed in Chap. 4.

Fig. 3.14 Hermann von Helmholtz (photo credit: http://en.wikipedia.org/wiki/Hermann_von_Helmholtz)

Fig. 3.15 An actual Helmholtz resonator (photo credit: http://en.wikipedia.org/wiki/ Hermann_von_Helmholtz)

For a derivation of this result see below.

(a) Suppose that a bottle has a volume of one liter $= 1000\,\text{cm}^3$, a mouth area of $2.4\,\text{cm}^2$, and a mouth length of 3.0 cm. Given a speed of sound 340 m/s, find the frequency.

(b) The **mouth** for a violin are its f-holes. Suppose that their combined area is $5.0\,\text{cm}^2$ and that the thickness of body ℓ of the wood is 3.0 mm. Finally, suppose that the volume of the body is $1{,}800\,\text{cm}^3$. Given a speed of sound 340 m/s, find the frequency. (Note that the Helmholtz formula assumes rigid walls; the flexibility of the walls of a violin reduces the actual frequency from the value calculated from the formula.)

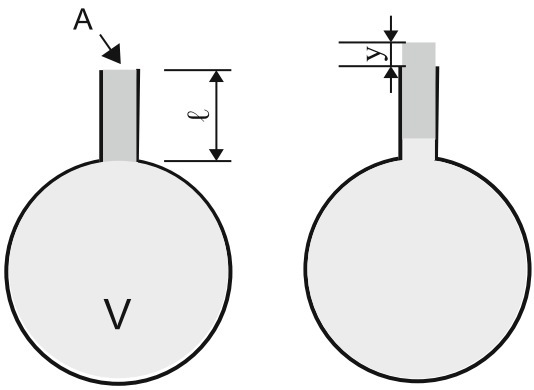


Fig. 3.16 A schematic of a model of a Helmholtz resonator

3.10.1 Derivation of the Helmholtz Formula

The mass m that oscillates is given by

$$m = \rho A \ell, \tag{3.31}$$

where ρ is the mass density of the air. We show in the figure this mass having moved upward by a small amount y.

In equilibrium, the pressure inside the bottle is equal to the pressure outside the bottle. The force downward on this mass is balanced by the force upward by the air inside the bottle. As a result of the motion of the mass upward, the pressure in the bottle is reduced, resulting in a net force downward – or a restoring force for the displaced mass. Since this force will be found to be proportional to the displacement, the system obeys Hooke's Law and the mass oscillates like an SHO. Further details follow below.

The volume of the air in the bottle properly increases by $V = Ay$. This increased volume leads to a decreased pressure, given by (3.5):

$$\Delta p = -B\frac{\Delta V}{V} = -B\sqrt{\frac{Ay}{V}}. \tag{3.32}$$

As a consequence, there is a net force on the mass m given by

$$F = \Delta p A = -\frac{BA^2}{V}y. \tag{3.33}$$

Comparison with Hooke's Law, $F = -ky$, where k is the spring constant, shows that the oscillator of the Helmholtz resonator has a spring constant given by

$$k = \frac{BA^2}{V}. \tag{3.34}$$

Using the equation for the frequency of an SHO

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}. \tag{3.35}$$

we obtain a frequency

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{BA^2/V}{\rho A\ell}} = \frac{1}{2\pi}\sqrt{\frac{B}{\rho}}\sqrt{\frac{A}{V\ell}}. \tag{3.36}$$

Equation (3.7), $v = \sqrt{B/\rho}$, leads to

$$f = \frac{v}{2\pi}\sqrt{\frac{A}{V\ell}}. \tag{3.37}$$

Chapter 4
Energy

We are all aware that electricity is needed to operate an audio amplifier. We casually say that we cannot get something for nothing. We pay the electric company an amount that is based on the number of kilowatt-hours of electricity used. In fundamental physics terms, electricity is a form of **energy** that is needed to power and operate an amplifier. The expenses of the electric company include the production of electrical energy from other forms of energy and the transmission of this form of energy from the electrical generator plants to your home.

Amplifiers are rated by the number of watts output per channel. The **Watt** is a unit of **power**, which expresses the rate of energy exchange. A certain fraction of this output of energy is associated with the sound waves emitted by the loudspeakers of the audio system. We can make similar observations regarding light bulbs, which are powered by electricity.

The output from a loudspeaker is not the only factor that determines how loud a sound we hear. Our distance from the loudspeaker matters too. The closer we are to a speaker, the louder the sound. This statement is merely a qualitative one: **Loudness** is a **subjective** parameter; not only does it depend upon the individual, but also it cannot be given a numerical value. Nevertheless, there exists an **objective** physical parameter, **intensity**, that can be used as a reproducible reference standard and is directly associated with loudness. Intensity reflects the **concentration of power over space**.

In describing vibrations of strings and pipes, we assumed for simplicity that once excited, vibrations would last forever. We recognize that in real systems, vibrations will die out unless they are sustained by excitation from without. This phenomenon of **attenuation** is fortunate: for example, in its absence our ears would be overwhelmed by all the sounds that have been produced in the past! Nevertheless, it is reasonable to ask where all the "action" connected with sound has gone. It could not merely disappear.

We sometimes ask this question in the context of money flow. Money changes hands. Money is often converted from one currency to another. Were it not for the printing of money by governments (and counterfeiters) the amount of money would remain constant. Similarly, we observe changes of various sorts in physical

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_4,

systems. Is there something about these systems that is nevertheless constant and would allow us to make a check on our measurements so that we can discover possible errors in the measurements. The answer is yes. Generally, the physical quantity known as **energy** is a measuring stick for keeping our accounts straight, as we will see. In the context of sound that dies out, the energy associated with the sound is replaced by what is known as **thermal energy**. In this chapter, we will study the physical parameter called *energy*, along with its related parameters, *power*, *intensity*, and *attenuation*. These parameters are necessary for providing us with an objective means of characterizing sound and light, so that we can better understand and appreciate our subjective experience.

4.1 Forms of Energy and Energy Conservation

Over the past 15 years or so, there has been much talk of an energy crisis. People talk about the need to conserve energy. And yet, there is a fundamental principle of physics known as the **Principle of Conservation of Energy**, that is, energy is automatically conserved, whatever we do! How can we reconcile these two contradictory claims?

The answer is simple: there are many different **forms** of energy. One can assign a numerical value to the amount of energy present among the various forms. The Conservation of Energy Principle states that whereas the amounts of each form of energy can change, the sum total of all forms of energy is constant. In other words, any amount of a given form of energy that is lost is replaced by a net increase in amounts of the other forms. Two systems can exchange energy; in the exchange process, the forms of energy may change. Nevertheless, whatever energy one system gains the other system must lose. The call to conserve energy is really a call to conserve those forms which are important to us.

We will be using the **Joule** as the preferred unit of energy in order to make direct use of the fundamental equations of physics. The "Joule" is named after **James Prescott Joule**, who identified heat as a process whereby thermal energy is transferred between one material and another. There are a number of units of energy in common use, depending upon the context. Some are listed in Table 4.1 along with their conversion to Joules. (Recall our comments earlier in the text on the need to use a consistent set of units.)

Table 4.1 Units of energy

Units of energy	Number of joules
Joule (J)	1
Kilowatt-hour (kWh)	$3.6 \times 10^6 = 3{,}600{,}000$
Food calorie = kilocalorie (kcal or Cal)	4,200
British thermal unit (BTU)	4.0
Erg (erg)	1.0×10^{-7}
Electron-volt (eV)	1.6×10^{-19}

4.1.1 Fundamental Forms of Energy

1. **Kinetic Energy (KE)** – energy associated with the motion of a massive object

$$KE = \frac{1}{2}mv^2, \tag{4.1}$$

Here m is the mass of the object and v is the speed.

Sample Problem 4-1

A 10-**tonne** truck (1 tonne = 1,000 kg ∼ 2,200 lbs) is moving at a speed of 10 m/s (1 m/s = 3.6 km/h ∼ 2.2 mph). What is its KE?

Solution

$$KE = \frac{1}{2}(10 \times 1{,}000)(10)^2 = 100{,}000\,\text{J} = 100{,}000\,\text{J}.$$

Sample Problem 4-2

What if the speed of the truck is doubled to 20 m/s?

Solution

Because KE is proportional to the square of the speed, the KE is quadrupled to

$$\frac{1}{2}(1{,}000)(20)^2 = 400{,}000\,\text{J}.$$

Sample Problem 4-3

Find the KE of a 100,000 tonne oil tanker that is moving at a speed of 2 m/s.

Solution

$$KE = \frac{1}{2}(10{,}000 \times 1{,}000)(2)^2 = 20{,}000{,}000\,\text{J}.$$

2. **Potential Energy (PE)** – energy associated with changing configurations of objects. This definition is bound to be obscure. Examples will help.

   (a) *Potential energy of a stretched or compressed spring*
   The amount is given by

$$PE = \frac{1}{2}ky^2, \tag{4.2}$$

   where k is the spring constant and y is the displacement. Note that the PE is positive whether the spring is stretched ($y > 0$) or compressed ($y < 0$).

Sample Problem 4-4

Find the PE of a spring which has a spring constant 25 N/m and a displacement of 2 cm.

Solution

$$PE = \frac{1}{2}(25)(0.02)^2 = 0.005\,\text{J}.$$

Sample Problem 4-5

Find the PE of the spring of the previous problem if the displacement is doubled to 4 cm.

Solution

Because the PE is proportional to the square of the displacement, the PE is quadrupled to $4 \times 0.005 = 0.020$ J.

   (b) **Gravitational Potential Energy** Suppose I lift an object and then release it from rest, allowing it to fall. It will accelerate, picking up kinetic energy. We regard the increased elevation as increasing the potential energy of the object in connection with the force of gravity of the earth. This potential energy is replaced by kinetic energy as the object falls.

   A change in elevation results in a change in PE, which we symbolize by ΔPE. Theory leads to the relation

$$\Delta PE = \text{Weight} \times (\text{Change in elevation}). \tag{4.3}$$

   If we let w be the weight and h the change in elevation we have

$$\Delta PE = wh. \tag{4.4}$$

   For example, if an object of weight 200 $Newtons$ ($\sim$44 lbs) is raised by 3 m, its PE increases by $200 \times 3 = 600$ J.

Fig. 4.1 Climbing a mountain (photo credit: http://vceoes.wikispaces.com/file/view/rock%2520climbing.jpg/41719235/rock%2520climbing.jpg)

Now suppose that a person weighing 150 lbs climbs a mountain, with an increase in elevation of 4,000 m. See Fig. 4.1 to give you a feeling for this activity.

What is his increase in *PE*? The stress on his body would lead us to believe that the PE change must be enormous. To the contrary, we will find that the total energy expended in climbing is many times greater than the change in PE. The process is incredibly inefficient.

To solve this problem, we need to relate a weight in *lbs* to a weight in Newtons. Our starting point will be the fact that a mass of 1 kg is equivalent to a weight of 2.2 lbs. Why are mass and weight related? What is the difference between the two?

Mass is a measure of the quantity of matter. It is independent of where an object is located. On the other hand, the **weight** of an object is the force

of gravity by the earth on the object. You may have heard reference to the weight of an object when situated on the surface of the moon or another planet. We would then be referring to the force of gravity of these bodies on the object. For example, a 60-pound object on earth weighs 10 lbs on the moon. Thus, the weight of an object depends upon the body exerting the force of gravity. **Isaac Newton** showed that the weight of an object is proportional to its mass m. The proportionality constant is the acceleration that an object would experience in falling in the absence of air resistance. Such a fall is referred to as "free fall." The acceleration under free fall is referred to as the "gravitational acceleration constant" and is given the symbol g. Its value is

$$g = 9.8\frac{\text{m}}{\text{s}^2} = 32\ \frac{\text{ft}}{\text{s}^2}. \tag{4.5}$$

Then we write

$$\text{weight} = w = mg. \tag{4.6}$$

The change in PE due a change in elevation is then given by

$$\Delta PE = mgh. \tag{4.7}$$

Sample Problem 4-6

Find the change in PE if a 10 kg object is raised by an elevation of 2 m.

Solution

$$\Delta PE = mgh = 9.8(10)(2) = 196\,\text{J}.$$

Back to the Mountain Climbling Problem

We need to express the weight 150 lbs in Newtons. To obtain the conversion, we can use the two corresponding values $m = 1$ kg and $w = 2.2$ lbs. A mass of 1 kg has a weight of

$$w = mg = 1 \times 9.8 = 9.8\,\text{Newtons} = 2.2\,\text{lbs}.$$

Therefore

$$1\ \text{Newton} = 22/9.8 = 0.22\,\text{lbs}.$$

or

$$1\,\text{lb} = 4.5\,\text{N}.$$

Therefore

$$150\,\text{lbs} = (150/0.22)\,\text{N} = 4.5 \times 150 = 680\,\text{N}.$$

The change in PE of the mountain climber is then

$$\Delta PE = wh = 680 \times 4{,}000 = 2{,}700{,}000\,\text{J}.$$

Thus, we have come up with the amount of PE needed for the climb but most readers have little sense as to what this number 2,700,000 J means. We need to express the PE in units that are more familiar. We will do so later on in the chapter when we discuss the food calorie.

(c) Later on in this text, we will discuss the force between electric charges. There is an **electrostatic potential energy** associated with this force that depends upon the distance between the charges.

3. **Electromagnetic Radiation** – This form of energy will be discussed further in Chapter 5, ELECTRICITY, MAGNETISM, AND ELECTROMAGNETIC WAVES. For the time being, we should note that it is this form of energy that the Sun transmits to us on earth and that is our utterly basic source of energy.

4.1.2 "Derived" Forms of Energy

The reader might wonder why electrical energy or nuclear energy was not included in the list of fundamental forms of energy. The reason is that these terms refer to energy that constitutes a mixture of what physicists regard as fundamental forms of energy. We list some examples below:

- **Chemical Energy** – energy associated with the KE and PE that is directly connected with the interaction of atoms in their binding together to form molecules.
- **Electrical Energy** – energy associated with the KE of electrical charge in electric currents or chemical energy stored in electric batteries. By means of chemical reactions, the chemical energy in a battery can be harnessed so as to produce electrical energy.
- **Nuclear Energy** – energy associated with the KE and PE of protons, neutrons, and lesser known particles, that is directly connected to their binding together to form the nuclei of atoms.
- **Thermal Energy** – energy associated with the random positions and velocities of atoms and molecules in a macroscopic body. Both kinetic energy and potential energy contribute to thermal energy.

Objects of all sizes can have kinetic energy and/or potential energy. Objects that are visible are referred to as **macroscopic bodies**. Taken together, the kinetic energy and potential energy of a macroscopic body is referred to as **Mechanical Energy**. That is,

$$\text{Mechanical Energy} = \text{Kinetic Energy} + \text{Potential Energy}$$

A macroscopic body consists of a huge number of atoms that move at velocities with great speeds and in random directions. In addition, the atoms interact with each other. As a consequence, there is kinetic energy and potential energy at the microscopic (here, atomic) level. This energy is called **thermal energy**. An increase in thermal energy is accompanied by an increase in the temperature. The common term for thermal energy is **heat**.

Consider a standing wave in a pitch pipe. As time progresses, the sound emitted from the open ends carries away energy. At the same time, the standing wave within the pipe loses amplitude and wave energy due to attenuation – the wave energy is replaced by thermal energy.

4.1.3 The Energy of Cheerios

Cheerios cereal has 110 Cal per *oz*. This means that when one oz. of Cheerios is digested, the chemical changes provide the body with 110 Cal of energy in a form that can be used by muscles to provide the body with mechanical energy. In fact, most of the chemical energy is replaced by thermal energy.

The chemical changes that take place in association with this process occur at the molecular level. Studies reveal that the typical change in the energy of a molecule is on the order of a few electron volts (eV). From this fact it is possible to estimate the number of molecules of Cheerios in an ounce of Cheerios. See Problem 4.2, at the end of the chapter.

Sample Problem 4-7

Remember the mountain climber whose *PE* rises by 2,700,000 J in climbing an elevation of 4,000 m. What is the weight of Cheerios that she has to eat to supply this *PE*, assuming 100% efficiency?

Solution

$$\text{Weight of Cheerios} = \frac{2{,}700{,}000\,\text{J}}{(110\,\text{Cal/oz})(4{,}200\,\text{J/Cal})} = 5.8\,\text{oz}. \quad (4.8)$$

Only a few bowls of Cheerios are necessary! It is amazing how little Cheerios would be needed if the conversion from chemical to *PE* were 100% efficient. We see that mountain climbing is incredibly low in efficiency. Most of the additional food one must consume is wasted and goes into thermal energy.

Diet-conscious people have a tendency to read carefully the number of calories per gram for every food they encounter. The fact is that the variation is not very

great once we take into account the fractions that are fat, protein, or carbohydrate. This is so because there is very little variation in the Cal/g for pure samples of these foods:

- Pure oils: 9 Cal/g$\sim$270 Cal/oz
- Pure protein or carbohydrate: 4 Cal/g$\sim$120 Cal/oz

Thus, digesting a gallon ($= 128$ fluid oz ~ 128 oz weight) of oil provides $128 \times 270 = 35{,}000$ Cal of thermal energy plus mechanical energy. Since digestion produces essentially the same chemical changes in oil as does **burning** – which represents a chemical reaction of a substance with oxygen (called **oxidation**) – burning one gallon of oil produces 35,000 Cal of thermal energy, which can be used to produce mechanical energy and/or electrical energy.

Question: How many gallons of oil would provide the energy needed to make that mountain climb?

A heat engine is a device that "extracts" mechanical energy or electrical energy from thermal energy. The **Second Law of Thermodynamics** informs us that a complete conversion from thermal energy to mechanical energy and/or electrical energy is impossible. Current heat engines do not even provide us with the maximum that this law provides. To get an idea about what is currently realized, consider that typical electric power plant is about 40% efficient. Suppose it were to burn oil to run the heat engine. Note that from one gallon of oil we obtain $35{,}000\,\text{Cal} = 1.5 \times 10^8\,\text{J} = 40\,\text{kWh}$ of thermal energy. This means that burning one gallon of oil can produce about $0.40 \times 40\,\text{kWh} = 16\,\text{kWh}$ of electrical energy.

4.2 The Principle of Conservation of Energy, Work, and Heat

Suppose that the above 10-kg object is released from rest at its elevation of 2 m from ground level. According to the Principle of Conservation of Energy, the object will accelerate downward, acquiring KE as it loses PE. Its loss in PE is compensated for by its increase in KE. When it reaches ground level, its KE must equal 196 J. We can therefore calculate its speed just before it hits the ground as follows:

We set

$$KE = \frac{1}{2}mv^2 = 196\,\text{J}.$$

Then

$$v^2 = 2KE/m = 2\left(\frac{196}{10}\right) = 39.2.$$

Thus

$$v = \sqrt{39.2} = 6.3\,\text{m/s}.$$

It is because one can retrieve KE in this way that the word "potential" is used to refer to the latent nature of PE.

We have indicated that the changes in the chemical energy of the food, known as metabolism, are the origin of the increase in PE associated with a person's raising an object. *We* say that the person does **work** in raising the object. In the framework of physics, ***work*** *is a process whereby energy is transferred from one body to another* through the application of a force and the motion of the body. The changes in chemical energy allow the person to exert that force and move the body.

On the other hand, when the temperature of a person's body rises above the temperature of the surroundings, thermal energy will be transferred from the person to the surroundings. This **heat transfer** is a second *process whereby energy is transferred from one body to another*.

4.3 Energy of Vibrating Systems

4.3.1 The Simple Harmonic Oscillator

A vibrating SHO has both KE and PE. Generally, its total vibrational energy E is given by

$$E = KE + PE = \frac{1}{2}mv^2 + \frac{1}{2}ky^2. \tag{4.9}$$

As the SHO vibrates, the displacement and speed are constantly changing, so that the respective amounts of KE and PE are constantly changing. Nevertheless, the total energy is constant. This result is a simple confirmation of the Conservation of Energy Principle. Let us apply the equation to an SHO which is vibrating with an amplitude A and a period T. In Fig. 4.2, we display the displacement as well as the velocity, as they vary with time over one cycle.

The displacement y varies from $-A$ to $+A$. The velocity ranges from $-v_m$ to $+v_m$. We will refer to v_m as the **velocity amplitude**; it is the maximum speed.

In Fig. 4.3, we display the PE and KE vs. time over a cycle. Note that when the *SHO is in its* ***equilibrium*** *configuration, the* ***kinetic energy*** *is a* ***maximum*** *while the* ***potential energy*** *is* ***zero***. On the other hand, when the SHO is *farthest* from its equilibrium configuration, the kinetic energy is ***zero*** while potential energy is a ***maximum***. Also, we see that the PE and KE both vary from zero to E, in just such a way that their sum is the constant E.

Now we will see how we can relate the displacement amplitude A to the velocity amplitude v_m. Initially, the object is at rest, so that there is no KE. All the energy resides in PE:

$$E = PE = \frac{1}{2}kA^2. \tag{4.10}$$

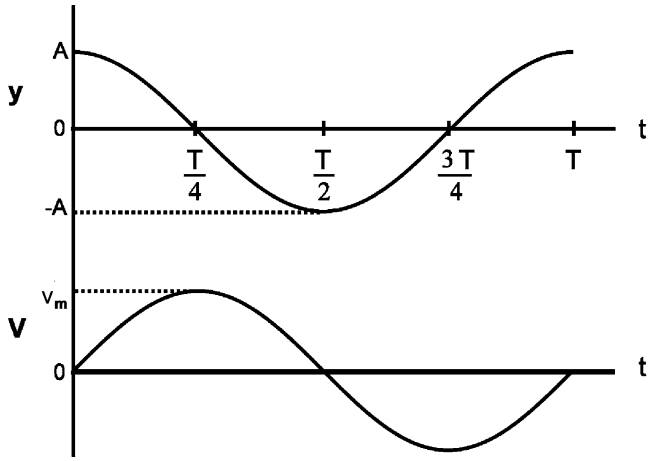


Fig. 4.2 Displacement and velocity of an SHO vs. time

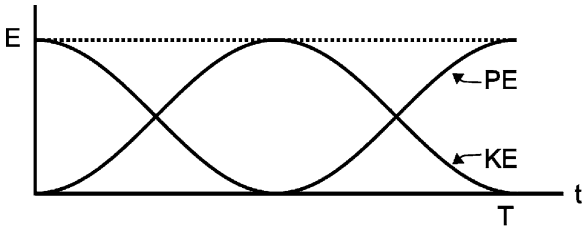


Fig. 4.3 Kinetic energy and potential energy of an SHO vs. time

On the other hand, when the object is at the equilibrium position ($y = 0$), one quarter of a cycle later, there is no *PE*. All the energy resides in *KE*:

$$E = KE = \frac{1}{2} m {v_m}^2. \tag{4.11}$$

The two expressions for the total energy E, (4.6) and (4.7), are equal, so that

$$E = \frac{1}{2} m {v_m}^2 = \frac{1}{2} k A^2$$

$$m {v_m}^2 = k A^2. \tag{4.12}$$

Thus,

$$v_m = \sqrt{\frac{k}{m}} A = 2\pi f A = 2\pi \frac{A}{T}. \tag{4.13}$$

Now recall that in Sect. 2.8 we *estimated* the velocity amplitude as $4A/T$; this is the average speed during a cycle. The ratio of the two expressions is $2\pi/4 \sim 1.6$. We expect the maximum speed to exceed the average speed!

Sample Problem 4-8

Find the maximum speed (velocity amplitude) and total energy of an SHO which has a spring constant of 100 N/m, a mass of 250 g, and an amplitude (displacement amplitude) of 20 cm.

Solution

$$v_m = \sqrt{\frac{k}{m}}A = \sqrt{\frac{100}{0.25}}(0.2) = 4.0\,\text{m/s}$$

$$E = \frac{1}{2}kA^2 = \frac{1}{2}(100)(0.2)^2 = 2\,\text{J}.$$

4.3.2 Energy in a Vibrating String

Let us consider all of the processes occurring when a **string is plucked**. In doing so, we will keep track of the various forms of energy involved:

- In plucking a string you do **work** in order to pull the string aside. That work involves a transfer of energy from your body (stored in the form of the food that you have eaten) to *PE* of the string.
- You then release the string. As it whips back toward the equilibrium, straight configuration, it picks up speed, hence *KE*, and necessarily loses *PE*. As the string vibrates, the amount of *KE* and *PE* oscillates, with the sum being a constant, as long as we can neglect attenuation.
- As the string moves through the air, it does work on the air in setting the air in motion. Sound waves are produced! The sound waves carry away energy in the form of *KE* and *PE*. Thus, the energy of the vibrating string must decrease. Its vibration is said to **attenuate**. Attenuation takes place also because of internal friction forces within the string, as its shape keeps on changing due to the vibration. This attenuation gives rise directly to an increase in the **thermal energy** of the string. The sound waves that are produced also attenuate, being replaced by increased **thermal energy** of the air.
- At any stage, the sum total of all forms of energy – your body's chemical and thermal energy, the string's vibrational and thermal energy, the sound wave energy and thermal energy of the air – is a constant.

4.3.3 *Energy in a Sound Wave*

A sound wave has the same two forms of energy as an SHO or a vibrating string: kinetic energy – associated with the motion of the medium, such as air – and potential energy – associated with condensation or rarefaction of the medium. To understand the origin of its potential energy, it is useful to note that it takes work to compress a volume of gas, so that its potential energy is thereby increased. Correspondingly, a compressed gas has the potential to do work on its surroundings. Conversely, in rarefying a volume of gas, the potential energy of the gas is *decreased*. Work would have to be done on the gas to compress it, so as to remove the rarefaction.

For all three of the above systems, as time moves on, the two forms of energy, potential and kinetic, vary the same way, as depicted in Fig. 4.3.

4.4 Power

Consider a traveling wave which is infinite in extent. There is no interest in the total energy in the wave since the total energy is infinite. We can instead deal with the amount of energy in a given length of traveling wave. The *energy per unit length* of a traveling wave is proportional to the square of the amplitude,[1] as in the case of an SHO. Furthermore, the same expression holds for the energy per unit length of a *standing wave* of a vibrating string or vibrating pipe.

Generally, energy is proportional to the square of the amplitude:

$$\text{Energy} \propto (\text{Amplitude})^2 \tag{4.14}$$

It is often more interesting to focus on the **rate at which energy is carried by the wave past a given point along the string**. This rate is equal to the rate at which energy would be given to the string by your hand (as in Fig. 2.5) in sending the traveling wave down the string, that is, the POWER input to the string. Why? Because whatever energy you pump into the string with your hand must pass the given point along the string.

We define "**power**" as

$$\text{Power} = \frac{\text{Energy}}{\text{Time interval}} \qquad \text{or} \qquad P = \frac{E}{\Delta t}. \tag{4.15}$$

[1] We can estimate the energy per unit length as follows: a unit length has a mass $m = \mu(1) = \mu$. Its average speed should be a bit less than the maximum velocity v_m. Thus, its *KE* should be a bit less than $(1/2)m{v_m}^2 = (1/2)\mu{v_m}^2$. This turns out to be the exact answer; it includes the *PE* too. Of course, $KE = PE =$ constant.

Recall that $v_m = 2\pi fA$ (see (4.8)). Thus, the energy per unit length is proportional to the square of the amplitude A, as in the case of the energy of a standing wave.

What energy E is used in this expression for power? We list some examples below:

- Energy transferred in the case of work
- Energy delivered in the form of electrical energy
- Energy passing a given point along the string
- Energy lost due to **dissipation**, which involves the production of thermal energy

As we noted at the beginning of this chapter, a basic unit of power is the watt, named after **James Watt**, who contributed important improvements in the design of the steam engine:

$$\text{One Watt} = \text{One Joule per second}$$

$$1\,\text{W} = 1\,\text{J/s}. \tag{4.16}$$

Another common unit of power is the **horsepower** (hp):

$$1\,\text{hp} \simeq 746\,\text{W}. \tag{4.17}$$

It is important for us to be clear as to what "power" represents in physics. When we talk about how "powerful" a person is, we are most often referring loosely to the force that person can exert; for example, how heavy an object that the person can carry. "Force" is different from "power." The rate at which the person can increase the potential energy by lifting them would express the person's power in physics terms.

Example 4-1

The human heart does quite a bit of work over the course of a 75-year lifetime in pushing blood through a person's body – enough to raise a typical battleship 10 m upwards, if the energy could be harnessed! The energy transferred to the blood amounts to about 20 billion Joules. However, this work is done over such a long period of time that the equivalent power is quite small:

Since one year $= 3 \times 10^7$ s,

$$75\,\text{years} = 75 \times 3 \times 10^7\ \text{s} = 2 \times 10^9\,\text{s}.$$

Thus, the power of the human heart is given by

$$\frac{E}{\Delta t} = \frac{20 \times 10^9}{2 \times 10^9} = 10\,\text{W}.$$

Example 4-2

A 100-W light bulb consumes 100 J/s of electrical power. The electrical energy is converted to light energy, as well as thermal energy. In fact, only a few percent of the total is light energy.

Example 4-3

Only a small fraction of the electrical power fed into a loudspeaker is converted to sound wave power. Most of the electrical energy is lost to thermal energy. According to Cambridge Sound Waves Inc., their loudspeaker had an **efficiency** of a mere 0.4%. That is, 1 W of electrical power produces only 0.004 W of sound power.

Example 4-4

The horsepower unit of power is based on a study of the performance of a real horse. In fact, horses can generate *many* horsepower. It is not difficult for a person to generate one *hp* of work. This would be accomplished by a person weighing 155 lbs by running up a hill and increasing his/her elevation at a rate of 3.6 ft/s. It is understood that the entire 1 hp is going into an increase in gravitational potential energy.

Power of Various Sources of Sound

In Table 4.2, we present the sound power produced by a number of musical instruments. Except for the cases so indicated, the maximum power is given. The range of powers is quite extraordinary, the ratio of the largest to the smallest value being $70/0.0000038 \sim 2 \times 10^7$, or 20 million to one! It is no small wonder that our auditory system can be responsive to such a large range.

4.5 Intensity

Loudness and **brightness** are subjective, psychological experiences. They cannot be quantified and are not scientific parameters. They both *reflect* the physical parameter called **intensity** that is applied to both sound and light. For a given frequency spectrum, we can generally expect loudness and brightness to increase

Table 4.2 Power of sound sources

Source of sound	Power (W)
Orchestra of 75 instruments	70
Bass drum	25
Trombone	6
Piano	0.4
Average sound power of an orchestra of 75	0.09
Flute	0.06
Clarinet	0.05
French horn	0.05
Average speech	0.000024
Softest violin passage	0.0000038

with increasing intensity. **Intensity** characterizes how concentrated the flow of energy is in space. Specifically, intensity is the rate at which energy passes through a unit area. We write

$$\text{Intensity} = \frac{\text{Power}}{\text{Area}}$$

$$I = \frac{P}{A} \tag{4.18}$$

which has the common units W/m^2 and W/cm^2.

The problems below should make the definition of intensity clear.

Sample Problem 4-9

Find the intensity of my voice, with a power of 10^{-3} W, which is traveling down a pipe of radius 2 cm.

Solution

$$A = \pi r^2 = \pi(2 \times 10^{-2})^2 = 1.3 \times 10^{-3}\,\text{m}^2$$

$$I = \frac{P}{A} = \frac{10^{-3}}{1.3 \times 10^{-3}} = 0.8\,\text{W/m}^2.$$

Sample Problem 4-10

Find the intensity of a straight laser beam, with a power of 20 mW $= 20 \times 10^{-3}$ W and a beam diameter of 1 mm ($= 10^{-3}$ m).

Solution

We have

$$A = \pi(\text{diameter})^2/4 = \pi(10^{-3})^2/4 = 7.9 \times 10^{-7}\,\text{m}^2,$$

so that

$$I = \frac{P}{A} = \frac{20 \times 10^{-3}}{7.9 \times 10^{-7}} = 2.5 \times 10^4\,\text{W/m}^2.$$

Fact: The intensity of sunlight just above our atmosphere is known as the **solar constant** and is given by 1,400 W/m^2.

4.6 Intensity of a Point Source

Let us turn our attention back to the *sun*. We can measure the solar constant by placing a light detector in a satellite just above our atmosphere. An interesting question is: what is the rate at which light energy is emitted by the sun, that is, the **luminosity** of the sun? The *solar constant* is related to the luminosity and the distance from the earth to the sun. We expect the intensity of sunlight to increase with increasing luminosity and to decrease with increasing distance from the sun. We now discuss the precise relationship between the three parameters.

To an excellent approximation, the sun emits light in all directions with the same intensity. We say that it is an **isotropic source**.[2] An isotropic source has the following important characteristics:

1. The intensity outside the source does not depend upon the *radius* of the source. In fact, because the radius can shrink to any size whatever without a change in intensity, an isotropic source is often referred to as a **point source**.
2. The intensity of a point source depends only upon the distance d from the center of the source and its power P, and not upon the direction from the source. Specifically, the intensity is inversely proportional to the square of the distance d.

$$I = \frac{P}{4\pi d^2}. \tag{4.19}$$

[2] The mere existence of sunspots tells us that the emission of light from the SUN cannot be perfectly isotropic.

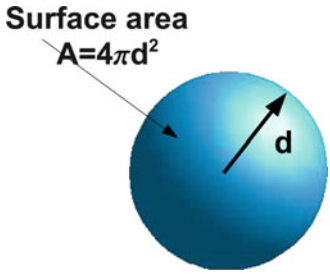


Fig. 4.4 Point source

We now provide a proof of (4.19): suppose that the source has a power P. Now consider a spherical surface of radius d, centered at the point source. The rate P at which energy leaves the source equals the rate at which energy flows through the sphere. Finally, the energy flows radially through this surface.

In Fig. 4.4, we have a point source at the center of a spherical surface. The power P is the rate at which energy is emitted from the source. The rate at which energy crosses the sphere must be P also since we have a steady state situation. The area A of the spherical surface is $4\pi d^2$. Equation (4.19) follows by direct substitution from (4.18).

Sample Problem 4-11

Estimate the light intensity of a 100-W light bulb having an efficiency of 3% at a distance of 1 m from the bulb by treating it as a point source.

Solution

light power = $P = 0.03 \times 100 = 3\,\text{W}$ with $d = 1\,\text{m}$,

$$I = \frac{P}{4\pi d^2} = \frac{3}{4\pi(1)^2} = 0.24\,\text{W/m}^2.$$

Sample Problem 4-12

The sun is 150 million km away from the earth. Find the light power emitted by the sun.

Solution

Recall that the sun's light intensity at the earth is $1{,}400\,\text{W/m}^2$. Thus,

$$P = 4\pi d^2 \times I = 4\pi(150 \times 10^6 \times 10^3)^2 \times 1{,}400 = 4.0 \times 10^{26}\,\text{W}.$$

Sample Problem 4-13

Suppose a loudspeaker emits sound power at 0.004 W, isotropically in the forward direction alone. Estimate the sound intensity at a distance of 2 m.

Solution

As an approximation, we can replace the area $4\pi d^2$ by $2\pi d^2$. Then

$$I = \frac{P}{2\pi d^2} = \frac{0.004}{2\pi(2)^2} = 1.6 \times 10^{-4}\,\mathrm{W/m^2}.$$

4.7 Sound Level and the Decibel System

The range of intensities of sound that a person would perceive as sound – that is, would detect and yet feel no pain – is enormous: from $\sim 10^{-12}\,\mathrm{W/m^2}$ to $\sim 1\,\mathrm{W/m^2}$. We will refer to this range of intensities as the "range of audible sound." The ratio of the highest to lowest intensities is 10^{12} = a trillion to one. Furthermore, the loudest sound does not feel a trillion times louder than the quietest perceivable sound. Also, doubling the sound intensity produces a change in loudness which most people would describe as being much less than a doubling in the loudness. Sound intensity thus does not give us a qualitative sense of loudness.[3]

Because of the above facts, an alternative way of expressing sound intensity was devised, called the **sound level**, or sometimes **sound pressure level**, which is abbreviated "*SL*." The unit for sound level is the **decibel**, abbreviated **dB**, in honor of **Alexander Graham Bell**, who is credited with inventing the telephone. In order to appreciate this system, we need to first review the important properties of logarithms.

4.7.1 Logarithms

The **logarithm** is a mapping of numbers in that the logarithm replaces any number by another number. (In mathematics, a mapping is often referred to as a **function**.)

[3]The range of sensitivity and tolerance of vision to light intensities is also understood to be about twelve orders of magnitude, from $\sim 10^{-10}\,\mathrm{W/m^2}$ to $\sim 100\,\mathrm{W/m^2}$.

The function depends upon the choice of **base**, which we choose to be ten. Some simple examples follow:

$$1/1{,}000 \rightarrow \log(1/1{,}000) = \log 10^{-3} = -3$$
$$1/100 \rightarrow \log(1/100) = \log 10^{-2} = -2$$
$$1/10 \rightarrow \log(1/10) = \log 10^{-1} = -1$$
$$1 \rightarrow \log 1 = 0$$
$$10 \rightarrow \log 10 = 1$$
$$100 \rightarrow \log 100 = 2$$
$$1{,}000 \rightarrow \log 1{,}000 = 3.$$

Generally, $10^n \rightarrow \log 10^n = n$. So much for the logarithm of powers of ten.

For *integers* between 1 and 10 we have:

$$\log 2 = 0.30$$
$$\log 3 = 0.48$$
$$\log 4 = 0.60$$
$$\log 5 = 0.70$$
$$\log 6 = 0.78$$
$$\log 7 = 0.85$$
$$\log 8 = 0.90$$
$$\log 9 = 0.95.$$

The logarithm of a number between one and ten can be calculated mathematically and is available in tables and from pocket calculators.

General mathematical properties of logarithms:

$$\log xy = \log x + \log y$$
$$\log x/y = \log x - \log y \quad \text{Note: } \log(1/y) = \log 1 - \log y = 0 - \log y = -\log y$$
$$\log x^y = y \log x$$
$$\text{If } z = \log x, \qquad x = 10^z.$$

Examples of the logarithm of some numbers that illustrate the use of these mathematical properties:

$$\log 40 = \log(4 \times 10) = \log 4 + \log 10 = 0.60 + 1 = 1.60$$
$$\log(3 \times 10^{11}) = \log 3 + \log(10^{11}) = 0.48 + 11 = 11.48.$$

4.7.2 Sound Level

We will now define a quantity that is an objective complement of the sound intensity. We define the **sound level**, SL, as follows:

$$SL = 10 \log \frac{I}{10^{-12}}. \tag{4.20}$$

Let us see how we calculate the sound level.

Example 4-5

$I = I_0 = 10^{-12}\,\mathrm{W/m^2}$:

$$SL = 10 \log \frac{I}{I_0} = 10 \log 1 = 0\,\mathrm{dB}.$$

The intensity $10^{-12}\,\mathrm{W/m^2}$ is called the **reference level**, being the lowest intensity that can be heard and corresponding to a SL of 0 dB.

Example 4-6

$I = 0.1 I_0 = 10^{-13}\,\mathrm{W/m^2}$, so that $I/I_0 = 0.1$:

$$SL = 10 \log(0.1 I_0) = 10 \log(0.1) = 10(-1) = -10\,\mathrm{dB}.$$

A sound level can be negative!

Example 4-7

$I = 1\,\mathrm{W/m^2}$,

$$SL = 10 \log \left(1/10^{-12}\right) = 10 \log \left(10^{12}\right) = 10 \times 12 = 120\,\mathrm{dB}.$$

We note that the range of audible sound is from $\sim$0 to $\sim$120 dB. This is not to imply that there are not people who can hear a sound that has a negative sound level.

4.7.3 *From Sound Level to Intensity*

We know how to calculate the sound level from the intensity. Suppose that we know the sound level and we want to calculate the corresponding intensity. We can invert equation (4.20) so that:

$$I = 10^{-12+SL/10}. \tag{4.21}$$

Sample Problem 4-14

Suppose that the sound level is, say, 85 dB. What is the corresponding intensity?

Solution

$$\begin{aligned} I &= 10^{-12+SL/10} = 10^{-12+85/10} = 10^{-12+8.5} = 10^{-3.5} \\ &= 3.2 \times 10^{-4}\,\mathrm{W/m^2}. \end{aligned} \tag{4.22}$$

We next describe how changes in intensities are related to **changes** in sound level. [4]

Sample Problem 4-15

A sound intensity is doubled, from I to $2I$. Find the change in SL.

Solution

The change in SL is given by

$$\begin{aligned} SL \equiv \Delta SL &= 10\log\left(\frac{2I}{I_0}\right) - 10\log\left(\frac{I}{I_0}\right) \\ &= 10[\log(2I) - \log I_0] - 10[\log I - \log I_0] \\ &= 10\log\left(\frac{2I}{I}\right) = 10\log 2 = 10(0.30) = 3\,\mathrm{dB}. \end{aligned}$$

[4] An excellent website for appreciating *changes* in sound levels is http://www.phys.unsw.edu.au/~jw/dB.html.

Note

The **change** in SL is independent of the initial intensity I; it depends only upon the ratio of intensities. Generally,

$$\begin{aligned}\Delta SL &\equiv SL_2 - SL_1 \\ &= 10\log\frac{I_2}{10^{-12}} - 10\log\frac{I_1}{10^{-12}} = 10\log\frac{I_2}{I_1}. \end{aligned} \tag{4.23}$$

Equation (4.23) can be "inverted": Suppose that we know the change in sound level ΔSL and want to know the corresponding ratio of intensities. We have

$$\frac{I_2}{I_1} = 10^{(\Delta SL/10)}. \tag{4.24}$$

Sample Problem 4-16

The sound level is increased by 10 dB. By what factor has the intensity increased?

Solution

Let I_1 = initial intensity and I_2 = final intensity. We seek I_2/I_1, given that $\Delta SL = 10$ dB. According to (4.23),

$$10\log\left(\frac{I_2}{I_1}\right) = \Delta SL = 10,$$

so that

$$\log\left(\frac{I_2}{I_1}\right) = 1.$$

Thus, $I_2/I_1 = 10$.
If $\Delta SL = 25$ dB,

$$\frac{I_2}{I_1} = 10^{\Delta SL/10} = 10^{25/10} = 10^{2.5} = 316.$$

We close this section by pointing out that we have assumed that the wave has a specific frequency, and hence wavelength. What happens if we have a more complex sound wave. An example is a wave that is produced by more than one source. The source might produce waves that have a definite fixed phase relation. In this case, we say the wave is **coherent**. Otherwise the wave is said to be **incoherent**. We will discuss these situations in Chap. 7.

4.8 Attenuation

Up to this point, we have assumed for simplicity that once a wave is established it will last forever. In fact, we know from experience that waves die out spontaneously. The technical term for this process is **attenuation**, or **damping**. In the case of a string, attenuation is mainly due to the force of the surrounding air on the string. In order to compensate for attenuation so as to keep a string vibrating, an external excitation force on the string must be maintained. Attenuation of sound is due to the very same intermolecular forces that sustain the wave itself. Energy is conserved in the process of attenuation through the production of thermal energy. This aspect of attenuation is referred to as **dissipation**. In this section we will discuss how we characterize attenuation numerically.

4.8.1 Attenuation in Time

We will first deal with *attenuation in time* of a mode of vibration. We will focus on the *amplitude* of the vibration. Its attenuation is exponential and is depicted in Fig. 4.5.

The black curves above and below the sinusoidal blue curve together comprise the **envelope**. The envelope above represents the attenuated amplitude and is shown as a specific quantitative curve in Fig. 4.6 below.

Let us define the **attenuation time** T as the **time it takes for the amplitude to be reduced in half**. In the figure, we have an initial amplitude of 8 units and we have an attenuation time of 2 s. We note that after 2 s, the initial amplitude has been cut in half – to $8/2 = 4$. After an additional 2 s, amounting to a total of 4 s elapsed, the amplitude is reduced by additional factor of 2 – to $4/2 = 2$, and so on.

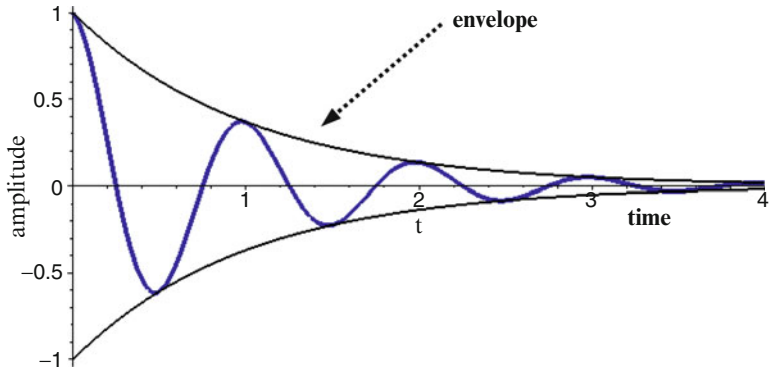


Fig. 4.5 Attenuated sine wave

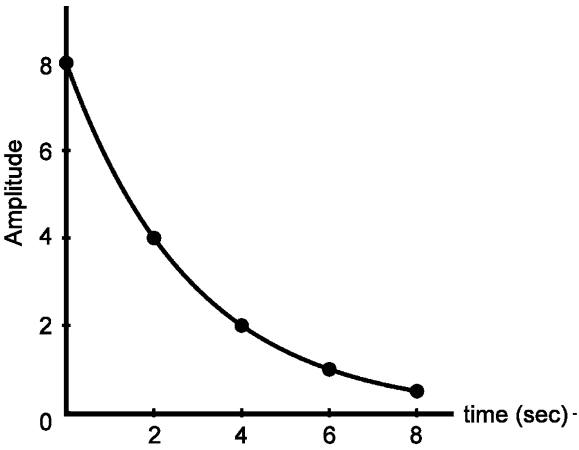


Fig. 4.6 Amplitude vs. time due to attenuation

Question: What will be the amplitude in the above example after 6 s? After 8 s? After 10 s?

The attenuation time is different for each mode. Generally, the higher the frequency is, the stronger will be the attenuation, so that the *shorter* will be the attenuation time. This fact explains why when a string is excited in a haphazard manner and the vibrations are allowed to attenuate, eventually one sees a pattern of vibration very close to that of the fundamental mode, which has the lowest frequency. Similarly, when striking a tuning fork, one often hears the fundamental frequency, which is usually the desired frequency, masked by an overtone frequency. Eventually, only the fundamental frequency is heard, because the fundamental mode attenuates slowest.

We can make this point clearer from the following mathematical digression. Suppose that only the first and second modes are excited, with initial amplitudes $A_1 = 16$ units and $A_2 = 64$ units, respectively. Suppose further that the respective attenuation times are $T_1 = 2$ s and $T_2 = 1$ s. In the table below, we provide the two amplitudes, A_1 and A_2, at a number of times after the initial time.

Time period (s)	A_1 – number of units	A_2 – number of units
0	16	64
2	$16/2 = 8$	$64/2^2 = 64/4 = 16$
4	$16/2^2 = 16/4 = 4$	$64/2^4 = 64/16 = 4$
6	$16/2^3 = 16/8 = 2$	$64/2^6 = 64/64 = 1$
10	$16/2^5 = 16/32 = 1/2$	$64/2^{10} = 64/1{,}024 = 1/16$

We see that while the ratio of the initial amplitudes A_1/A_2 was 1 : 4, after 10 s it is 8 : 1. The fundamental dominates.

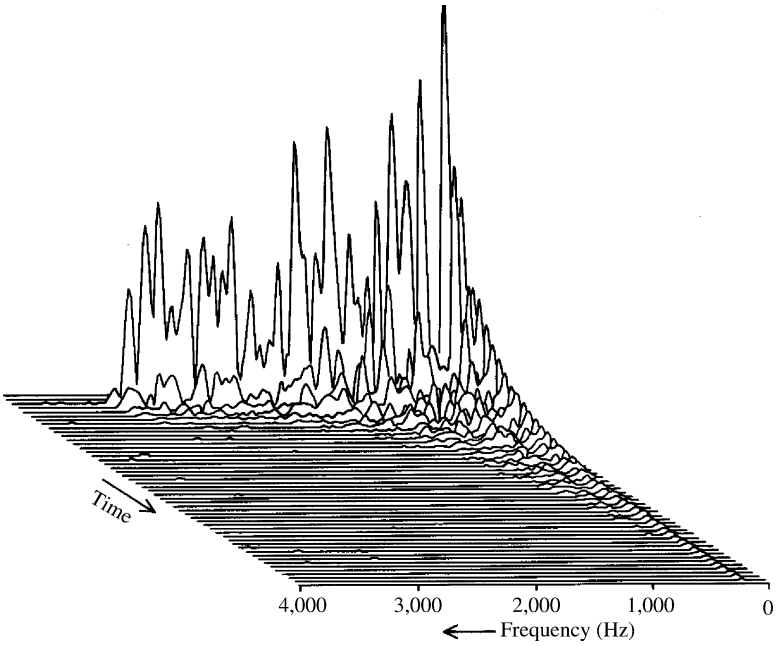


Fig. 4.7 The change in the frequency spectrum of a tom–tom with time (photo source: *The Science of Musical Sound*, by John R. Pierce (Scientific American Books, New York, 1983))

The behavior summarized above can be seen clearly in Fig. 4.7, wherein we see a depiction of how the spectrum of the sound of a tom–tom changes with time after the tom–tom is struck.

What we see is a set of many sound spectra taken one after another. Each spectrum runs from right to left. The earliest spectrum, at the rear of the set, has a series of many peaks, each peak reflecting the excitation of a particular mode of the tom–tom. As time progresses, each peak decreases in amplitude. However, we see that the higher the frequency of a mode, the faster the mode dies out. Toward the end, only the peak of the fundamental is observable, albeit with quite a small amplitude.

Note

We should be aware that it takes some time interval over which the sound is sampled to obtain a single spectrum. The time interval must be well chosen: if it is not much longer than the period of the fundamental, the peak of the fundamental would be washed out and not be clearly discernable. If the interval is comparable to and longer than the attenuation times, the above spectra would not accurately reflect the attenuation at a given point in time.

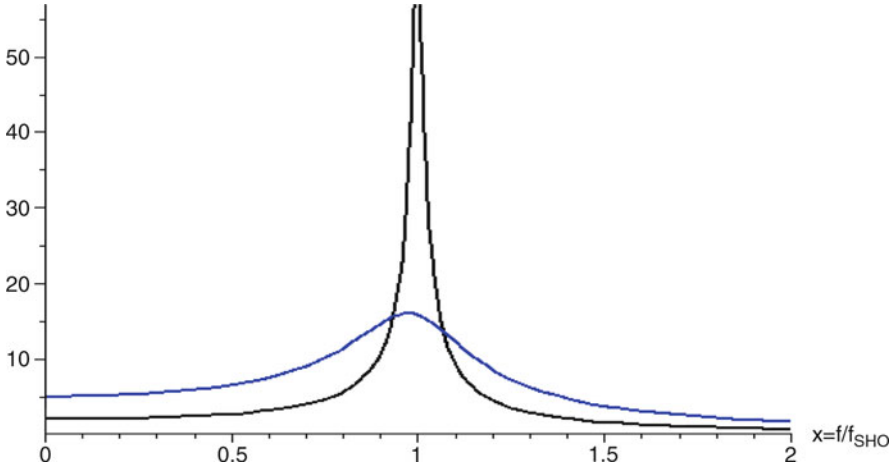


Fig. 4.8 Response of an SHO vs. frequency – *black* low attenuation; *blue* high attenuation

> The ear must also be carrying out an analysis that takes into account this dilemma: to produce a sense of pitch, a pure tone must last long enough to contain many oscillations. Its duration must be much longer than the period. To hear a sense of pitch, the ear must be sampling the sound over such a long duration of time. (That is one reason why it is difficult to have a sense of pitch of a percussive sound having a very low frequency and therefore a very long period.) On the other hand, since we do sense changing pitches, the sampling time must be shorter than the interval of time over which the pitch is changing in order to discern that changing pitch.

4.8.2 Resonance in the Presence of Attenuation

In Chap. 2 we noted that we have resonance when we have system A disturbed by another system B in a periodic way at a frequency equal to the frequency of one of the modes of system A. It might have seemed that resonance requires an exact equality of frequencies. This is not the case. Any amount of attenuation reduces the required degree of equality. In Fig. 4.8, we see the response of an SHO to an external force of frequency f. We see that the resonance at $f = f_{\text{SHO}}$, corresponding to $x = 1$, is sharper for lower attenuation. In this case, if you want a significant response, the frequency f must be relatively very close to the frequency of the SHO. At a higher attenuation, you can get a greater response even with a great mismatch. *It can be shown that the width of the peak is on the order of the inverse of the attenuation time.*

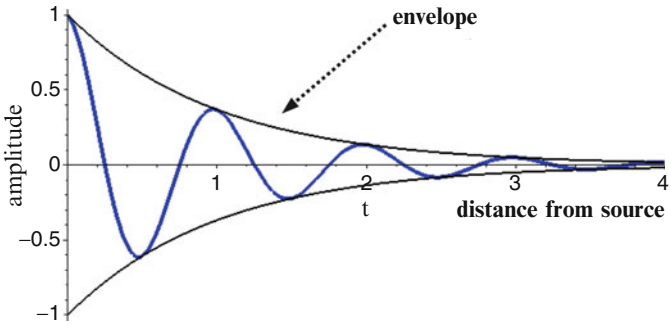


Fig. 4.9 Amplitude vs. distance from source due to attenuation

There are situations when you want to drive an SHO and avoid the resonance so that you can have a uniform response with respect to a specific range of frequencies. This is the case in a loudspeaker which has an SHO connected to the cone that vibrates so as to produce sound waves. Note that with modest attenuation, as we see in the blue curve, the response is constant for low frequencies. Thus, for a loudspeaker, you would want the frequency of the SHO to be considerably above the range of audible sound 20 kHz and an attenuation time that is a bit greater than the frequency so that you can avoid the resonance but leave a significant response in the audible range.

4.8.3 Attenuation of Travelling Waves: Attenuation in Space

Suppose a steady source produces a sinusoidal wave travelling in *one direction*[5], such as a wave along a string or a sound wave down a very long pipe. In the absence of attenuation, the amplitude of the wave will be constant along the length of the wave. However, because of attenuation, the waveform will have an ever decreasing amplitude, as shown in Fig. 4.9.

The **envelope** of the waveform describes an attenuation of the amplitude *in space*. Numerically, the envelope is characterized by the **attenuation length**, which is the *distance* along the wave over which the amplitude decreases by a factor of 2. The stronger the attenuation, the *shorter* the attenuation length.

[5]It is **essential** to keep in mind that even in the absence of attenuation, the intensity of a wave that is emitted by a *point source* will decrease according to the "inverse square law" of (4.19). Attenuation will produce an additional contribution to the decrease in the intensity with increasing distance from the point source.

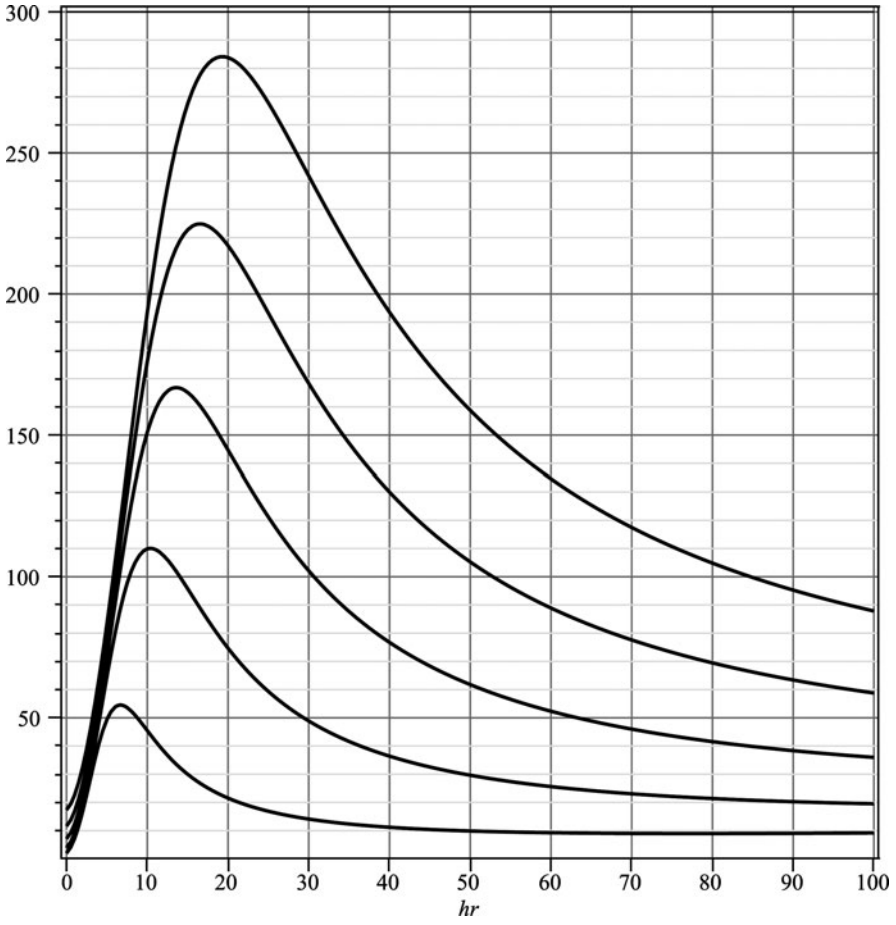


Fig. 4.10 Attenuation vs. relative humidity and frequency

Typically, the **attenuation constant** α_L is used to characterize the attenuation of sound through a medium. It is inversely proportional to the attenuation length.[6] In Fig. 4.10,[7] we see graphs of the attenuation constant of sound in air vs. **relative humidity** for various frequencies of sinusoidal sound waves.

In Fig. 4.10, you will find the degree of attenuation of sound in air. Each curve is associated with a different frequency. The relative humidity is indicated on the

[6] We have $\alpha_L = \ln 2/(\text{attenuation length})$.

[7] The graph was produced using the standard **ISO 9613-1:1993**.
See http://www.iso.org/iso/catalogue_detail.htm?csnumber=17426. I obtained the formula from the website (2-2-2011): http://www.sengpielaudio.com/AirdampingFormula.htm.
I am grateful to Eberhard Sengpiel for his help.

horizontal axis. The attenuation constant α_L is specified on the left vertical axis. On the right vertical axis, you will find indicated a quantity referred to as the **attenuation**, which we indicate by the symbol α. This measure of attenuation describes the attenuation of the sound level as opposed to the amplitude. (It is clear from the figure that this unit of attenuation is proportional to the attenuation constant.[8]) The attenuation α refers to the drop in sound level in dB per kilometer. Therefore, if the sound travels through a distance x, the change in sound level is given by

$$\Delta SL = -\alpha x. \tag{4.25}$$

How do we use this equation? Suppose that the attenuation $\alpha = 20\,\text{dB/km}$ and that the sound travels through a distance of 5 km. Then the change in sound level will be

$$\Delta SL = -\alpha x = -20(5) = -100\,\text{dB}. \tag{4.26}$$

The peaks in the individual curves indicate that for a given frequency, adding moisture to the dry air first leads to an increase in the degree of attenuation (hence, a decrease in the attenuation length) and that further increases in moisture lead to a decrease in the degree of attenuation. Also, for a given relative humidity, the **degree of attenuation increases with increasing frequency**. (Recall the increase in the degree of attenuation of modes with increasing mode frequency.)

The problems below will explain how we use the graph.

Sample Problem 4-17

Suppose we have a sound wave of frequency 5,000 Hz (= 5 kHz) and the relative humidity is 20%. Suppose further that the initial intensity I_1 is $10\,\text{W/m}^2$. Find the intensity I_2 after a distance of 1 m, after 10 m, and after 100 m.

Solution

According to Fig. 4.10, the curve for a frequency of 5 kHz gives an attenuation of about 100 dB/km. One meter equals 0.001 km, so that the reduction in *SL* is

$$\Delta SL = -100 \times 0.001 = -0.1\,\text{dB}. \tag{4.27}$$

[8] Analysis leads to: [attenuation constant in m^{-1}] = [0.00005 ln(10)] × [attenuation in dB/km] = 0.00012[attenuation in dB/km].

Thus,

$$\Delta SL \equiv 10\log(I_2/I_1) = -0.1. \tag{4.28}$$

According to (4.24),

$$\frac{I_2}{I_1} = 10^{-\Delta SL/10} = 10^{-0.01} = 0.98, \tag{4.29}$$

so that

$$I_2 = 0.98 I_1 = 0.98 \times 10 = 9.8\,\mathrm{W/m^2}. \tag{4.30}$$

For a distance of 10 m = 0.01 km, we have

$$\Delta SL = -100 \times 0.01 = -1\,\mathrm{dB}. \tag{4.31}$$

so that

$$\Delta SL \equiv 10\log(I_2/I_1) = -1, \tag{4.32}$$

and

$$\log(I_2/I_1) = -0.1. \tag{4.33}$$

Thus,

$$\frac{I_2}{I_1} = 10^{-0.1} = 0.79, \tag{4.34}$$

and

$$I_2 = 0.79 I_1 = 0.79 \times 10 = 7.9\,\mathrm{W/m^2}. \tag{4.35}$$

After 100 m = 0.1 km, we have

$$\Delta SL = -100 \times 0.1 = -10\,\mathrm{dB}. \tag{4.36}$$

$$\Delta SL \equiv 10\log(I_2/I_1) = -10. \tag{4.37}$$

or

$$\log(I_2/I_1) = -1. \tag{4.38}$$

Thus,

$$I_2/I_1 = 10^{-1} = 0.1 \tag{4.39}$$

and

$$I_2 = 0.1 I_1 = 0.1 \times 10 = 1\,\mathrm{W/m^2}. \tag{4.40}$$

Sample Problem 4-18

Suppose that the attenuation is 10 dB/km. Find the distance x travelled by the wave such that the intensity drops by a factor of 2.

Solution

We set

$$10\log(I_2/I_1) = -(10\,\text{dB/km})x\,(\text{in km}),$$

where $I_2/I_1 = 1/2$. Since $\log(1/2) = -\log 2$, we have $-10\log 2 = -10x$ or

$$x = \log 2 = 0.3\,\text{km} = 300\,\text{m}.$$

Note

The attenuation of light in common glass is not negligible, amounting to on the order of 10,000 dB/km or 10 dB/m. Thus, in passing through a 1 m thickness of such glass, the light intensity will be reduced by a factor of 10. Fiber optics communication has depended upon the development, since around the year 1970, of glasses having an extraordinarily low attenuation of about 0.1 to 1 dB/km! The transition from wired connection to optical fiber connection resulted in an increase in the speed of communication by a factor of about ten.

Note that a 1-dB drop corresponds to a drop in intensity by a factor of $10^{0.1} = 1.3$.

4.9 Reverberation Time

If you produce a sound in a room, the sound you hear might resound like an echo. Such is often the case for a large empty room. On the other hand, if the room is small and/or is loaded with furniture, the sound dies out so fast that we tend to label the room as being **dead** to sound. Sound produced in an open field is quite dead. What is the process that determines the resulting character of the sound?

The sound produced by a source bounces off the surfaces of the room. Eventually the sound will die out – mostly because it passes on into these surfaces. How long a sound lasts is characterized by the **reverberation time** ($\equiv$ RT).[9] It is defined as the

[9]For more information on the reverberation time, see the website: http://www.yrbe.edu.on.ca/~mdhs/music/oac_proj97/music/reverb.html. You will be able to calculate the reverberation time of a room given the volume of the room and the area and absorption constant of each surface within the room.

time it takes for the sound intensity in the room to drop by a factor of one million. The corresponding **drop** in the sound level is

$$\Delta SL = 10\log(10^6) = 60\,\text{dB}.$$

Over the course of an interval of two reverberation times, the intensity will drop by a factor of (one million)^2 = one trillion.

Very long reverberation times produce an echo effect; very short reverberation times produce a dead space. In the design of auditoria or music rooms, one will design the space so as to produce a reverberation time that suits one's preference. Typically, a reverberation time on the order of one second is desired.

Sabine's Law: When a sound wave is incident upon a surface, a certain fraction of the intensity is reflected while the remainder is transmitted into the surface. If the surface is very thick, all of the transmitted sound will be absorbed. On the other hand, a wall might be so thin that some of the sound intensity passes through on into the air or other material on the other side of the wall. Often, this second fraction alone is referred to as the transmitted sound. For walls in a room, one often refers to the sum of both absorption and transmission as **absorption**.

Suppose now that a sound is produced in an empty room. The room has walls, floor, and ceiling. Each of these surfaces will reflect a certain fraction of sound. The fraction that is not reflected is called the **absorption coefficient** – with symbol α. (This symbol should not be confused with the symbol α_L used for the attenuation coefficient.) The absorption coefficient depends upon the frequency of the sound wave and the surface material.

Suppose that all the surfaces have the same absorption coefficient and that the total surface area is A. It can be shown that the reverberation time RT is approximately given by

$$\text{RT} = 55.2\frac{V}{v\alpha A}. \tag{4.41}$$

Here v is the sound velocity in air. If $v = 340\,\text{m/s}$,

$$\text{RT} = 0.16\frac{V}{\alpha A}, \tag{4.42}$$

where V is expressed in m^3, A is expressed in m^2, and the reverberation time is expressed in seconds. This equation is referred to as **Sabine's Law**. Note that the greater the volume of a room the longer the reverberation time. On the other hand, increasing the absorption coefficient or the surface areas decreases the reverberation time.

Usually a room has surfaces with different absorption coefficients. Then, with the surfaces having areas $A_1, A_2, A_3, \ldots$, and absorption coefficients $\alpha_1, \alpha_2, \alpha_3, \ldots$, the factor αA in (4.41) and (4.42) is replaced by the sum

$$\alpha A \equiv \alpha_1 A_1 + \alpha_2 A_2 + \alpha_3 A_3 + \cdots. \tag{4.43}$$

Table 4.3 Chart of absorption coefficients

Material	128 Hz	256 Hz	512 Hz	1,024 Hz	2,048 Hz	4,096 Hz
Draperies hung straight, in contact with wall, cotton fabric, 10 oz. per square yard	0.04	0.05	0.11	0.18	0.30	0.44
Rock wool (1 in.)	0.35	0.49	0.63	0.80	0.83	–
Carpet on concrete (0.4 in.)	0.09	0.08	0.21	0.26	0.27	0.37
Carpet, on $1/8''$ felt, on concrete ($0.4''$)	0.11	0.14	0.37	0.43	0.27	0.27
Concrete, unpainted	0.010	0.012	0.016	0.019	0.023	0.035
Wood sheeting, pine (0.8 in.)	0.10	0.11	0.10	0.08	0.08	0.11
Brick wall, painted	0.012	0.013	0.017	0.020	0.023	0.025
Plaster, lime on wood studs, rough finish ($1/2$ in.)	0.039	0.056	0.061	0.089	0.054	0.070

Acoustic tile, suspended from the ceiling, has an absorption coefficient close to unity:

	125 Hz	250 Hz	500 Hz	1,000 Hz	2,000 Hz	4,000 Hz
Suspended acoustic tile	0.76	0.93	0.83	0.99	0.99	0.94

Table 4.3 is a chart with absorption coefficients for a variety of materials and for various frequencies.

Source: http://www.sfu.ca/sonic-studio/handbook/Absorption_Coefficient.html

4.10 Terms

- Attenuation constant
- Attenuation length
- Attenuation time
- Brightness
- British Thermal Unit (BTU)
- Calorie or food calorie(Cal)
- Chemical energy
- Decibel scale
- Dissipation
- Electrical energy
- Electromagnetic energy
- Electron-volt (eV)
- Energy
- Envelope of an attenuated wave
- Exponential behavior
- Gravitational potential energy
- Heat transfer intensity
- Joule (J)
- Kilowatt-hour (kWh)
- Kinetic energy
- Loudness
- Nuclear energy
- Point source [=isotropic source]
- Potential energy

- Power
- Principle of conservation of energy
- Reference level reverberation time
- Sound level
- Stroboscope
- Thermal energy
- Weight
- Work
- Watt (W)

4.11 Important Equations

Kinetic energy:

$$KE = \frac{1}{2}mv^2. \tag{4.44}$$

Potential energy of a simple harmonic oscillator:

$$PE = \frac{1}{2}ky^2. \tag{4.45}$$

Change of gravitational potential energy with change of elevation:

$$\Delta PE = \text{weight} \times \text{change of elevation} = \Delta PE = wh = mgh. \tag{4.46}$$

Intensity defined:

$$\text{Intensity} = \frac{\text{Power}}{\text{Area}}$$

$$I = \frac{P}{A}. \tag{4.47}$$

Intensity of a point source:

$$I = \frac{P}{4\pi d^2}. \tag{4.48}$$

Sound level defined:

$$SL = 10 \log \frac{I}{10^{-12}}. \tag{4.49}$$

$$\Delta SL \equiv SL_2 - SL_1 = 10 \log \frac{I_2}{I_1}. \tag{4.50}$$

Change in sound level with change in intensity re-expressed:

$$\frac{I_2}{I_1} = 10^{(\Delta SL/10)}. \tag{4.51}$$

Change in sound level with distance due to attenuation:

$$\Delta SL = -\alpha x, \tag{4.52}$$

where α is in dB/km and x is in km.

4.12 Problems for Chap. 4

1. Estimate the number of Calories, Joules, and kiloWatt-hours (kW-h) in a gallon of oil.
2. One ounce of Cheerios provides about 100 Calories of food energy through chemical changes. On the average, one *molecule* of Cheerios provides about 5 eV of food energy. From this information, estimate the number of molecules in an ounce of Cheerios. (You will need to make use of the number of eV there are in a Calorie.)
3. How many *lbs*. of Cheerios would one have to eat to provide the mechanical energy output of about 25-billion J of a human heart in a 75-year lifetime?
4. What are the two basic modes of transfer of energy from one system to another?
5. When a person is said to be "powerful enough to lift a 200 lb. object," is one referring to the person's "power" according to the way that term is used in physics? **Explain**.
6. Calculate the power delivered by a jet engine that increases the kinetic energy of a jet plane from zero to 40 billion Joules in 5 min. (40 billion Joules of *KE* corresponds to a mass of 1,000 tonnes (1 tonne = 1,000 kg $\cong$ 2,200 lbs) moving at 300 m/s ($\cong$ 1,000 km/h).)
7. A loudspeaker delivers 0.002 W of sound power down a tube of diameter 4 cm. Calculate the sound intensity travelling down the tube.
8. A **stroboscope** is a device that produces a series of flashes of light having an extremely short duration in time (μs) but very large energy, with a variable frequency, that can range from about one flash per second to tens of thousands of flashes per second. Consider a particular stroboscope – model MVC-4100 – manufactured by the *Electromatic Equipment Corporation*. The corporation provides the following specifications for this model on its website: (http://www.inspectionstroboscope.com/prods/MVC-4000?PHPSESSID=7933e3da79f0401c1e54c9ea5d0e8367)

   Input energy per flash: 5.41 J; duration of a pulse of electrical: 30 μs (= 30 microseconds); radiometric light output (light energy per flash): 0.210 J.

   (a) Calculate the average electric power delivered in a single flash over the duration of a pulse of electric energy.
   (b) Calculate the efficiency in the production of light energy from electric energy.

9. Calculate the light intensity at the earth from a star that is 100 light years away (1 light year $= 9.5 \times 10^{15}$ m) and emits light with a power equal to that of our sun (4.0×10^{26} W).
10. (a) If the sound intensity from an isotropic source is 0.2 W/m^2 at a distance of 1 m, what will the sound intensity be at a distance of 2 m?
    (b) Calculate the sound level (in dB) corresponding to each of the two intensities of part (a). What is the corresponding change in sound level?
11. (a) What is the "solar constant"?
    (b) The atmosphere absorbs sunlight. Suppose that as a result during some period the sunlight intensity on the surface of the earth is one-half the solar constant. Suppose further that we have an array of solar cells on a roof with dimensions 10 m by 6 m facing directly toward the sun. Finally assume that for every 100 J of incident sunlight energy the cells produce 15 J of electrical energy. (Their efficiency is then 15%.) Calculate the electrical power output of the array.
12. Describe in detail, using all the physical principles we have covered so far, how the sound emitted by a tuning fork varies in time after it has been struck hard.
13. (a) What is the "attenuation time"?
    (b) What is the "attenuation length"?
14. Suppose that a ball is released from rest at a meter above a hard floor. It bounces back up and reaches a height of one-half meter. Now suppose that it keeps bouncing, with each bounce leading to a maximum height that is one-half the previous height.

    After how many bounces will the maximum height be one $\mathring{A}$ above the floor? (One $\mathring{A} = 10^{-10}$ m.)

    The significance of $1\mathring{A}$ is that this elevation is about equal to the size of a molecule, so that a bounce disappears within the typical motion of the molecules. For all intents and purposes, the ball has stopped bouncing.

    Hint: Let $n = number\ of\ bounces$. Find an expression for the height after n-bounces and take the log of that expression.
15. If the amplitude of vibration of an SHO decreases from 8 to 4 cm in 1 min, what will be its amplitude after 2 min? After 3 min?
16. The following is a fascinating application of attenuation time. We will consider a pipe that is excited in a particular mode. On the one hand, we recall that the assumption is that the ends are nodes for the sound pressure. In the section on standing waves, it was assumed that pulses are totally reflected from the ends. If this were exactly so, the pipe would not produce a sound wave outside the pipe, so that we would not hear the excited pipe!

    In fact, the ends are not nodes precisely and sound is emitted from an open end any time a pulse reaches that end. In order to produce a sound associated with a particular mode, we need a compromise, a balance between reflection and emission of sound into the air outside. The pertinent transmission coefficient T is defined as the fraction of sound energy that is emitted when a wave reaches an open end.

(a) Now imagine a wave moving back and forth in the pipe. With each incidence on an open end, a fraction T of sound intensity is lost. In order for the mode to be well defined and be heard with a clear frequency, it is necessary that there be many oscillations with negligible attenuation. What condition does this place on the transmission coefficient?
(b) Suppose that the wavelength is much larger than the diameter of the pipe. It can be shown that the transmission coefficient is approximately given by the ratio of the square of the circumference to the wavelength, that is

$$T = \left(\frac{2\pi R}{\lambda}\right)^2 . \tag{4.53}$$

Does the transmission coefficient increase or decrease with higher frequency?

Now consider a flute of radius one centimeter played at a frequency of 440 Hz. Calculate the corresponding transmission coefficient.

17. If the frequency of sound increases, the attenuation length will:

decrease / increase / remain the same.

18. If the humidity of air increases, the attenuation length will:

decrease / increase / remain the same / all three of the previous choices are possible.

19. What range of intensity is associated with a **negative** sound level?
20. Suppose the light intensity drops by a factor of 4 in passing through 1 m of a certain glass.
    (a) What is the corresponding drop in dB?
    (b) What drop would there be in passing through *four* meters of glass?
21. (a) Use Fig. 4.10 of the text to determine the attenuation of sound in dB/km of 2 kHz sound in air at 20% humidity.
    (b) If a sound wave of frequency 2 kHz is travelling in a straight line through the air, what would the change in sound level be and by what factor would the intensity drop in traversing a distance of 2 km?
22. Choose a room with which you are familiar. You will have to know a bit about the materials out of which the walls, floor, and ceiling are made. If you are not sure, make a guess. (Most walls in homes are made of gypsum. Ceilings are usually made of gypsum. Sometimes they are made of acoustic tile, which have a higher absorption coefficient.) Then go to one of the websites below and estimate the RT (reverberation time) for that room. Discuss your result in relation to the sound quality of the room.
    (a) A simple website to use to calculate the reverberation time (RT) for a room. But you obtain the RT for each of a set of frequencies, one at a time http://www.saecollege.de/reference_material/pages/Reverberation%20Time%20Calculator.htm

(b) This website calculator is a bit more cumbersome for inserting data but gives you RT for a number of frequencies after one click. http://www.atsacoustics.com/cgi-bin/cp-app.cgi

(c) This site has audio files so that you can hear the difference among various reverberation times: http://www.armstrong.com/reverb/main.jsp

23. Suppose that a room is a cube with sides of length 3 m. Compare the RT with a larger cubic room with sides of length 4 m, i.e. find the ratio of the RTs. Note that both the volume V and the total area A increase. Would you have expected RT to increase or decrease? Explain.

24. Suppose that the absorption coefficient were unity ($\alpha = 1$). Then any sound wave incident on a wall would be completely removed from the room. The reverberation time would have no meaning since any sound created in the room would be completely lost in a time at most equal to the time it takes for a sound wave to cross the room. The sound intensity would drop suddenly from its initial value to zero in this time. And yet, the Sabine formula, (4.41), leads to a finite reverberation time. Obviously, the formula is in error and can only be an approximation. It can be shown that Sabine's formula holds only when $\alpha \ll 1$.

A more accurate formula is identical to Sabine's formula except that α is replaced by $-\ln(1-\alpha)$. Here, **ln** is the natural log (or log to the base "e"). Thus, (4.42) is replaced by

$$RT = 0.16 \frac{V}{-\ln(1-\alpha) \cdot A}. \tag{4.54}$$

Note that if $\alpha = 1$*, this formula leads correctly to* $RT = 0$*, since* $\ln(0) = -\infty$*. Thus, according to this formula, it takes no time for the sound to die out.*

Question: Suppose that a sound wave were traveling across a room of the shape of a cube of side L. How long would the wave take to travel directly from one side to the other and back? Compare this time to the RT as expressed in (4.41). Remember that the formula includes the six sides of the cube. What is the ratio of the two expressions: how many round trips would take place in a RT?

In fact, the Sabine formula assumes many traversals of the sound wave before the sound intensity drops significantly. Such behavior requires an absorption coefficient that is much less than unity.

Chapter 5
Electricity, Magnetism, and Electromagnetic Waves

We have mentioned that **light** is an **electromagnetic wave** with a frequency that lies in a particular range: $\sim 4 \times 10^{14}$ to $\sim 7 \times 10^{14}$ Hz. But what is an electromagnetic wave? To answer this question, we will need to study **electricity** and **magnetism**. The principles of this branch of physics are the basis of operation of the various electronic instruments used in sound reproduction, such as radio transmitters and receivers, amplifiers, microphones, and speakers. There is an interesting further relevance: The **atom** often serves as a primary *source* of light and is the *receiver* of light in our eyes. How the atom performs these functions and how the atom is held together depend upon the laws of electricity and magnetism. We experience numerous manifestations of **electricity** which are not dependent upon technological developments: They were known by mankind before the age of science. Most powerful and majestic are thunder and lightning, which involve enormous currents of **electric charge** flowing from thousands of feet above, down to the earth's surface. **Static electricity** is another manifestation. It is said that **Thales of Miletus** was the first to note around 600 BC that amber rubbed with fur attracts straw. In the *magnetism arena*, many of us are familiar with "**lodestone**," which is the rock mineral "magnetite" in a **magnetized** state. Lodestone is capable of attracting iron and, for centuries, served as a **compass** for guiding sailors out at sea.

5.1 The Fundamental Forces of Nature

We note that electricity and magnetism are manifested in an obvious way by *forces*, the so-called electric force and magnetic force. These two **fundamental forces** in nature were originally incorrectly believed to be unrelated to one another. Subsequently, as we shall see later in this chapter, they were found to be so interrelated that they are together referred to as one force, the "**electromagnetic force**." Other fundamental forces include the "**gravitational force**," the "**nuclear force**," and the "**weak force**." (Interestingly, physicists have a goal to see whether *all* these forces are manifestations of only *one* all encompassing and more fundamental force.)

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_5,

Fig. 5.1 Isaac Newton – Portrait by Sir Godfrey Kneller in 1702 (source: http://en.wikipedia.org/wiki/File:Sir_Isaac_Newton_by_Sir_Godfrey_Kneller,_Bt.jpg)

Until this century, the force of gravity was the best understood of all forces, thanks to the "Universal Theory of Gravitation" that was expounded by **Isaac Newton** (1642–1727) (Fig. 5.1).

Below we highlight the main features of Newton's Theory of Gravitation.

1. Matter has a quantifiable attribute called **mass**. (Common units of mass are the **gram** (g) and the **kilogram** (kg).) Mass gives a body **inertia** (resistance to having a change in velocity) and is also the source of the gravitational force.
2. The gravitational force between two bodies can be determined in terms of the force between two **point masses**. These are idealized bodies which take up no space. Given the force between two point masses, one can determine the force between any two real bodies which take up space. The most important aspects of this force are that it is always attractive, with each body attracting the other toward its respective self with a force of the same magnitude, and that its magnitude decreases with increasing separation between the bodies.
3. While Isaac Newton successfully used his theory to account for the motion of the planets about the sun and the moon about the earth as well as the motion of projectiles such as bullets which are flying through the air near the earth's surface, he was deeply perplexed about how one object could exert a force on another with only empty space between them. We are referring here to what is called the **action-at-a-distance** enigma.

It is important to realize that this issue must be regarded as a philosophical one, being outside the realm of science. Ultimately, the goal and the measure of success of a scientific theory is found in terms of its ability to provide relationships among measured physical quantities, rather than its ability to explain why the phenomena take place. "Why" questions essentially seek underlying, more fundamental principles that can be applied to the specific phenomenon at hand. Thus, we might

ask why the sky is bright with the answer being that the molecules of the air scatter sunlight. But why do the atoms scatter light? Well, because they are comprised of electric charges, which experience a force from the light and so on. The sequence of questions must ultimately end when we reach the most fundamental level of theory that encompasses all phenomena below it. At that point, "why" questions cease to have meaning. Of course, physicists are never content to conclude that they have definitely reached the most fundamental level. They are always open to revelations of the new.

We will see in this chapter that the action-at-a-distance question did lead to the introduction of the concept of the "force-field," which, while it may not have satisfyingly addressed the action-at-a-distance question, has been extremely useful in the development of physics.

5.2 The Electric Force

In the late 1700s, experimental studies of the force between electrically charged bodies led to a theory for the electric force which is similar to the theory for gravitational forces. It is embodied in "**Coulomb's Law**." We summarize the theory below:

1. Bodies can have an attribute called **electric charge**.

   There are two types of charge – referred to as **positive charge** and **negative charge**. Letting q_1 and q_2, respectively, be the numerical values of the charges of two bodies, we find that the force is:

   - **Repulsive** if q_1 and q_2 have the same sign (hence either both positive or both negative).
   - **Attractive** if q_1 and q_2 have opposite signs.

   This summary is exhibited in Fig. 5.2:
2. The **action-at-a-distance** issue is manifest with electric forces too. How can two electric charges affect each other when they are not, so to speak, *touching*?

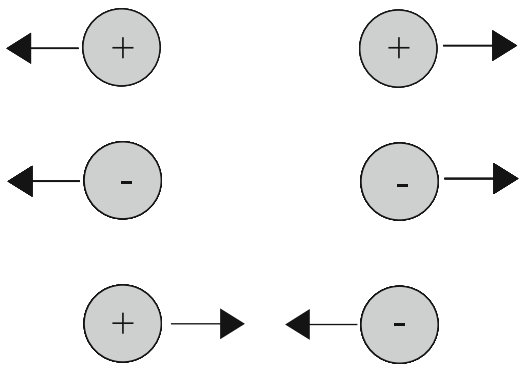


Fig. 5.2 Forces between electric charges – two positive, two negative and one positive with one negative

In fact, a study of the force on your body by another object along with the structure of atoms and the electric forces therein that are in fact responsible for this force reveals that the force on your body does not involve contact between the charges therein.

5.3 Electric Currents in Metal Wires

We begin this section by summarizing some important information about all matter on earth. Matter consists of atoms. These in turn consist of **nuclei** and **electrons**. The electron has a negative electric charge. The nucleus consists of a tightly bound collection of **protons** and **neutrons**. Protons have a positive charge. Neutrons, while being a charged structure, have no net electric charge – they are "electrically neutral." An **ion** is a neutral atom which has gained or lost electrons and thus is electrically charged.

Roughly, materials can be divided into two categories with respect to electrical behavior: The first category consists of the **insulators**. While they have electric charge, the charges are bound and cannot move freely within the material. In a sample of the second category – consisting of **conductors** – there are charges that are free to move throughout the material. **Metals** are a specific type of conductor. The atoms of a metal are **ionized**: Atoms have lost some electrons, which are then free to move about a solid array of positive ions. The former, the so-called free electrons (also called **conduction electrons**), move about randomly at speeds averaging about 1,000 km/s! An **electric current** in a metal wire represents a *net* flow of free electrons. We might refer to this flow on a large scale as a **macroflow** of free electrons. For comparison sake, the so-called flow of a river is a **macroflow** of water: Water molecules in a river are in random motion, with speeds on the order of a kilometer per second, even when the river as a whole is still. The flow of a river reflects an average motion in some specific direction that is added to the background random motion (Fig. 5.3).

One can compare the motion of the charges to that of a pinball in a pinball machine; the pinball moves helter–skelter, nevertheless making overall progress down the board.

The direction of an electric current is the same as the direction of motion of the positive charges, but opposite to the direction of the negative charges. If both signs of charge are present and moving, we must subtract the contribution of the negative charges from that of the positive charges. (See Fig. 5.4.) In the case of a

Fig. 5.3 Random motion of electrons in a wire

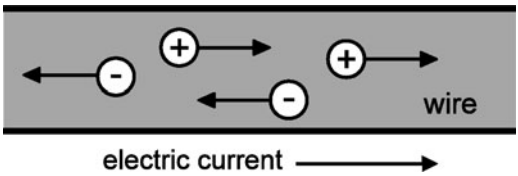


Fig. 5.4 Electric current in a wire

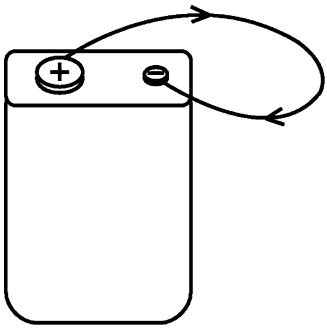


Fig. 5.5 Electrons flowing from the negative terminal to the positive terminal of a battery

current in a metal wire, there is no contribution from the positive ions. That is, only the conduction electrons contribute to the current and the current is opposite to the direction of the conduction electrons.

The common **electric battery** drives current from the **positive terminal** through the wire and back into the **negative terminal** of the battery, as shown in Fig. 5.5. Thus, electrons flow through the wire from the negative terminal to the positive terminal.

5.4 The Magnetic Force

Originally, magnetic forces were confined to magnetized bodies, called **magnets**. We note that a sample of magnetic material need not be magnetized. Magnetite is a naturally occurring magnetic material which when magnetized is called **lodestone**. A sample of iron can be magnetized by a piece of magnetite. The earth as a whole acts as if it has a huge mass of magnetized material. Typically, a pair of magnetized bodies exert forces on each other which tend to orient the bodies in a certain direction with respect to each other.

To the eye, a magnetized body appears **isotropic**. In order to appreciate what this implies, consider the rotation of a magnet having a spherically shaped body. If the magnet were rotated, you would not notice any change in its appearance. In fact, at an invisible level, the body is actually **anisotropic**. Its magnetic behavior can be characterized by picturing the body as having an axis with a **north pole** and a **south pole**, as seen below in the sample of magnetite. Thus, as far as its magnetic properties are concerned, one could easily detect that the magnet had been rotated. We indicate the north and south poles of a sample of magnetite and a **bar magnet** in the figure below. N represents the north pole, while S represents the south pole (Fig. 5.6).

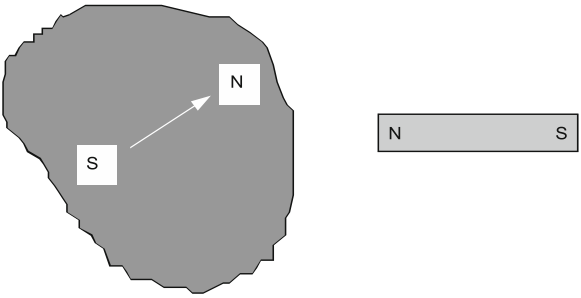


Fig. 5.6 Magnetite and a bar magnet

Fig. 5.7 The earth's magnetic south pole indicated by the small white circle in northern Canada

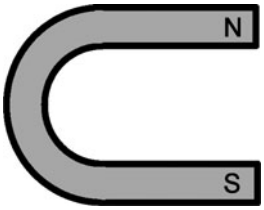


Fig. 5.8 Horseshoe magnet

The magnetic south pole of the earth is shown in Fig. 5.7 to be in *northern* Canada.[1]

Finally, a bar-like magnet can be bent into a **horseshoe magnet** so as to produce strong magnetic forces between its poles (see Fig. 5.8).

[1]The *north* magnetic pole is diametrically opposite, in the *southern* hemisphere, close to New Zealand. During this century, it has moved on average 10 km/year.
See the Wikipedia article (12-26-2010): http://en.wikipedia.org/wiki/Magnetic_declination.
Also: http://obsfur.geophysik.uni-muenchen.de/mag/news/e_nmpole.htm.

5.5 Magnetic Forces Characterized

1. As with electric forces, action-at-a-distance is manifest.
2. Like poles repel; unlike poles attract (Fig. 5.9).
3. Two freely suspended bar magnets tend to rotate so as to line up with parallel axes (Fig. 5.10).
4. A non-magnetized material can be **magnetized** through the presence of an already magnetized body. When the bar magnet is removed, the degree of magnetization may diminish, sometimes to an undetectable level (Fig. 5.11).

What happens to a bar magnet when it is cut in half? One might be inclined to guess that we end up with one half being a north pole and the other a south pole. To the contrary, we end up with two shorter bar magnets, each with a north pole and a south pole, as seen in the Fig. 5.12.

This result should be contrasted with what happens when a bar of electrically polarized material is cut in half. In Fig. 5.13, we see what happens to a bar of metal that is polarized by a neighboring point electric charge. Each half of the bar is charged, one positively, the other negatively. If the point charge is removed, the two halves remain charged. On the other hand, if a bar of insulating material is so polarized and halved, the two halves are not charged. In the presence of the point charge, each half will be polarized; while, if the point charge is not present when the bar is halved, neither half is charged or polarized.

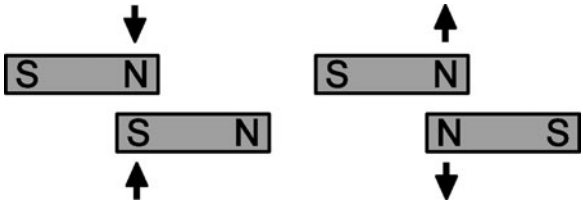


Fig. 5.9 Attracting and repelling poles

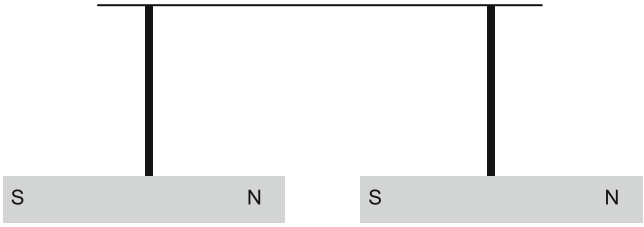


Fig. 5.10 Suspended magnets rotate so that axes align

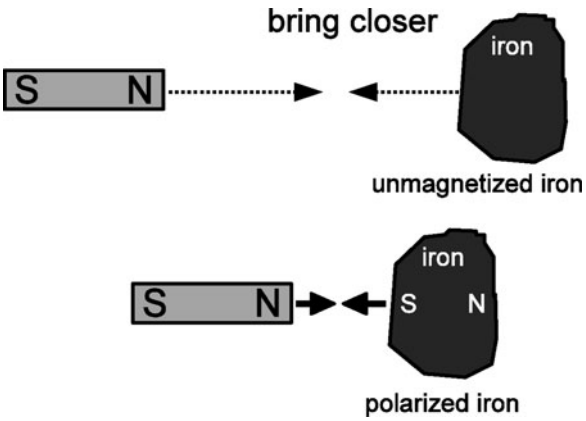


Fig. 5.11 Polarizing an unmagnetized piece of iron

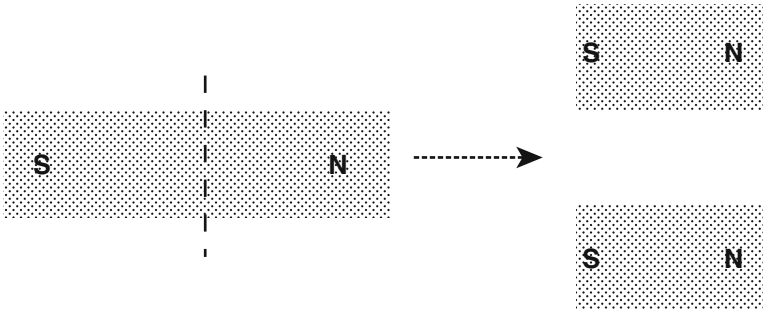


Fig. 5.12 Splitting a bar magnet creates two bar magnets

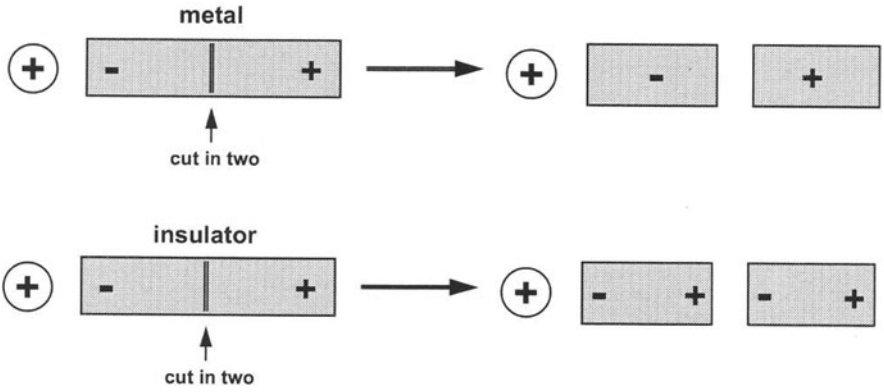


Fig. 5.13 Results of splitting polarized materials: metals vs. insulators

5.6 Is There a Connection Between Electricity and Magnetism?

In 1820, **Hans Oersted** (Fig. 5.14) is reported to have been lecturing his class on this issue with the following demonstration, exhibited in Fig. 5.15.

A compass needle was placed close to an electric wire that was connected to a switch and a battery. In advance of closing the switch and thus turning on a current through the wire, he told his class that it was obvious that there would be no effect on the wire. Alas, he was wrong! The needle tended to be oriented in a particular way, as depicted in Fig. 5.16.

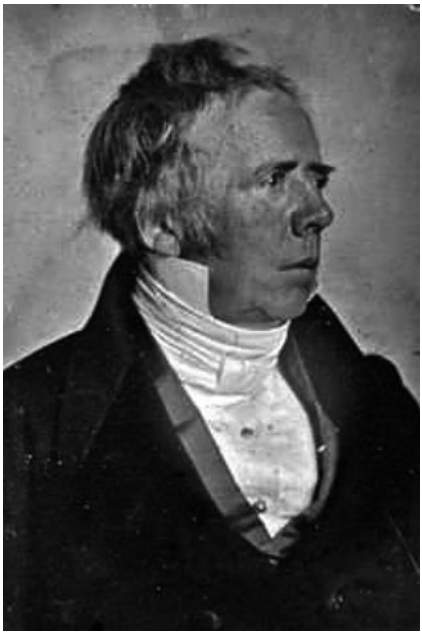


Fig. 5.14 Hans Oersted (source: http://en.wikipedia.org/wiki/Hans_Christian_Ørsted)

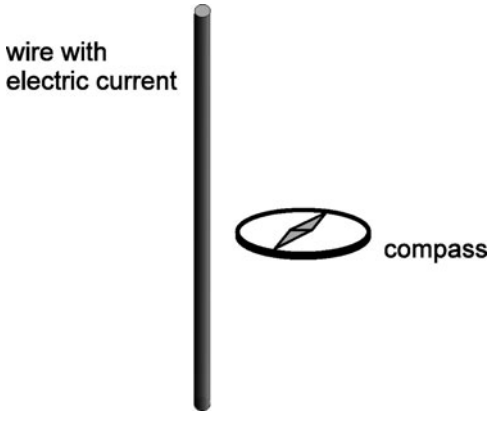


Fig. 5.15 Electric current exerting a force on a magnet

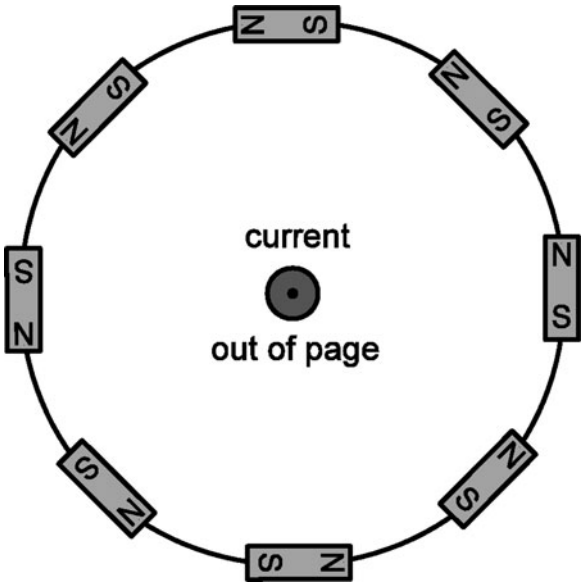


Fig. 5.16 Orientation of compasses around a current-carrying wire

We depict the wire oriented perpendicular to this page. A number of compass needles are suspended along a circle with the wire at its center. Current is in a direction out of the page. In equilibrium, the axis of each compass needle is tangent to the circle.

> **Note**
>
> We will use the symbol $\odot$ to indicate a direction *out of the page*. It represents the head of an arrow. We will use the symbol $\otimes$ to indicate a direction *into the page*. It represents the tail of an arrow.

We are inclined to conclude that in being able to orient a bar magnet, an electric current acts like a magnet.

The ability of a wire carrying an electric current to orient a magnet can be exhibited very nicely using iron filings. (These are elongated bits of iron, around a millimeter or less long.) In the figure below, iron filings have been distributed on the surface of a piece of cardboard. The cardboard is held horizontal (so that the filings will not fall off) and a wire carrying an electric current is passed through its center. The electric current magnetizes the filings, so that they provide us with a distribution of a huge number of bar magnets on the piece of cardboard. The alignment of the filings shows up clearly in the figure (Fig. 5.17).

The details of how specific shapes of wires carrying a current affect magnets was determined experimentally by André Ampère. See Fig. 5.18.

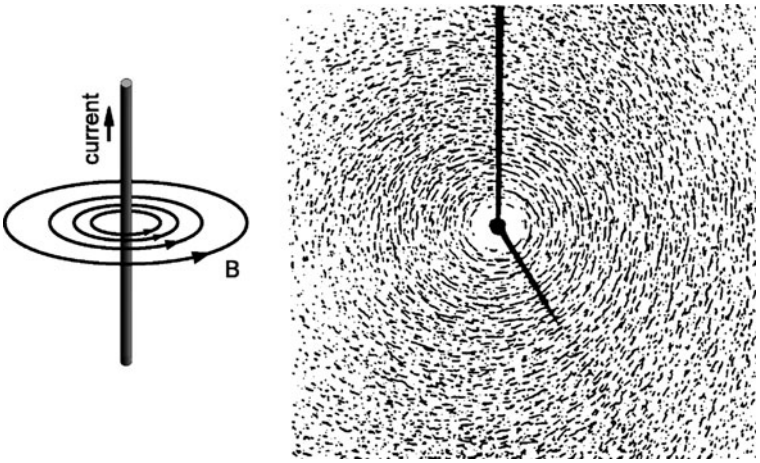


Fig. 5.17 Iron fillings aligning in the direction of the magnetic field

Fig. 5.18 André Marie Ampère (source: http://en.wikipedia.org/wiki/Andre|_Ampere)

Another significant discovery by Ampere was that two wires carrying a current exert a force on each other. No magnet is required. We can appreciate this result better by examining further Oersted's discovery that a current-carrying wire exerts a force on a magnet. We will see below that a magnet exerts a force on a current-carrying wire.

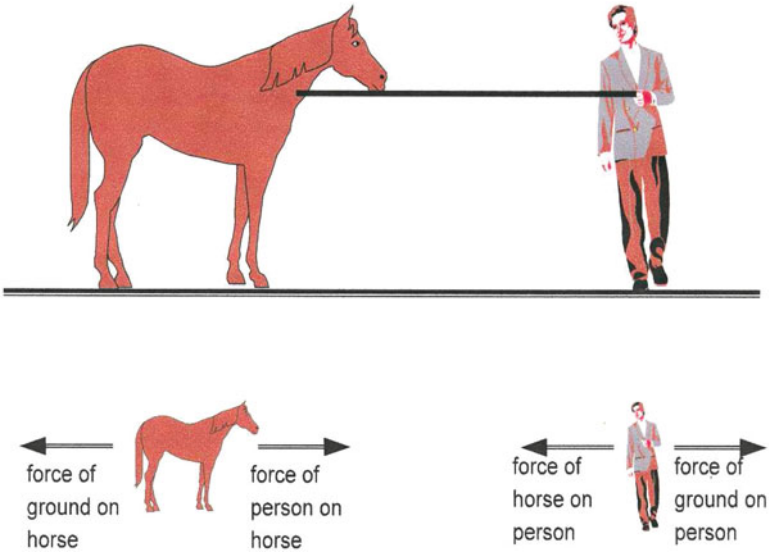


Fig. 5.19 Tug of war between a man and a horse. How does the horse win?

5.6.1 Action–Reaction Law and Force of Magnet on Current-Carrying Wire

We can learn something entirely new about magnets and electric currents using **Newton's Third Law** of Dynamics. It is often referred to as the **Law of Action and Reaction** and states that

a force of one body on a second body is always automatically accompanied by a force of the second body on the first.

This law is often difficult to comprehend; it is baffling and seems to contradict our intuition about forces. For example, if I push on a wall with a force of, say, 10 lbs, the wall must be concurrently pushing on me with a force of 10 lbs. Thus, as a further example, we arrive at the remarkable fact that it is impossible for a horse to push on me with a force that is greater than the force I exert on it! It is then reasonable to wonder how a live healthy horse can win a tug-of-war. To solve this dilemma, we turn to Fig. 5.19, wherein the forces on each of the participants is exhibited.

We note that both the horse and I experience forces by the ground: The ground pushes upward on the horse and on me, forces that overcome the downward force of gravity. The ground supports our weights. However, in addition, the ground provides a tangential force that tends to prevent us from slipping. Yet there is a limit to how much tangential force the ground can exert before we slip. While the force of the horse on me is exactly equal to the force I exert on the horse, the horse is capable of

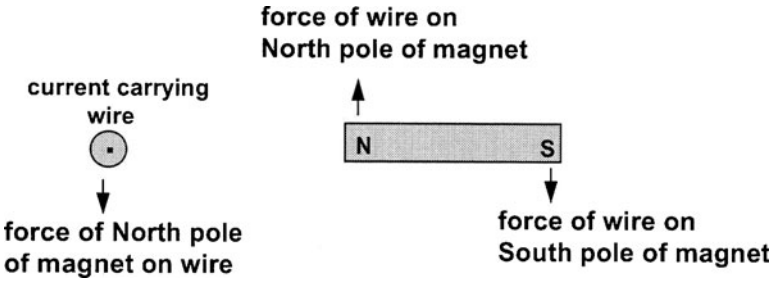


Fig. 5.20 Forces acting on a compass near a wire

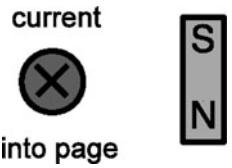


Fig. 5.21 A compass aligns in the opposite direction when the current is flowing into the page

experiencing a much greater force of the ground on it than the force of the ground on me. As a result, I will slip before the horse does and I will lose the competition.

What do we learn from the above? according to **Newton's Third Law**,

a magnet must also exert a force on a wire carrying a current!

A clarifying way to exhibit this second force is with the following configuration (Fig. 5.20).

> **Note**
>
> The force is perpendicular to both the direction of the current and the North–South axis of the bar magnet.

Now suppose that the direction of the current in the wire is reversed. Then:

1. The compass needles in Fig. 5.16 will tend to be oriented in the reverse direction as seen in Fig. 5.21.
2. The force of the wire on the poles of a bar magnet are reversed in direction (Fig. 5.22).

A very important configuration of electric current in a wire is the **solenoid**. This consists of a wire wound like a helix (Fig. 5.23).

Note that reversing the direction of the current *or* the orientation of the helix *clockwise* or *counter clockwise*) exchanges the poles of the solenoid. For example, we might have a solenoid as seen in Fig. 5.24.

Henceforth, we will represent the solenoid by a set of parallel line segments as illustrated in Fig. 5.25.

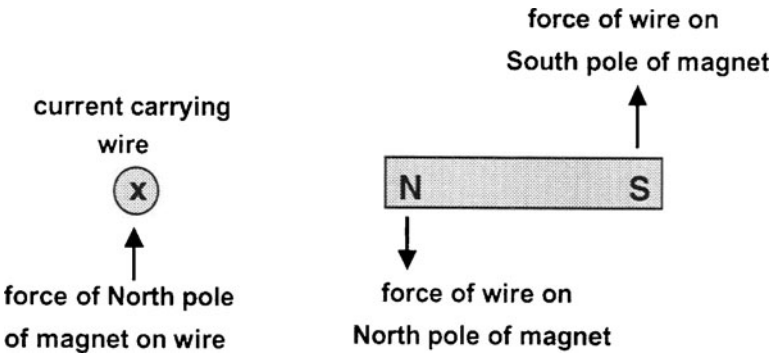


Fig. 5.22 Forces by a current-carrying wire on a bar magnet

Fig. 5.23 A solenoid

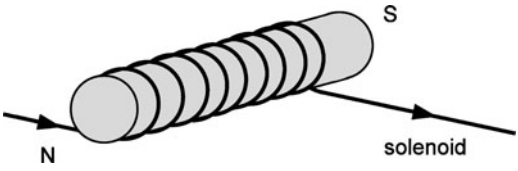


Fig. 5.24 A solenoid with reversed winding

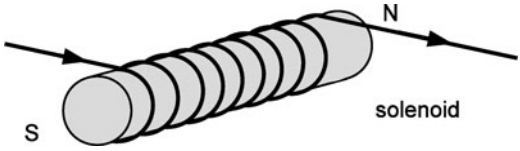
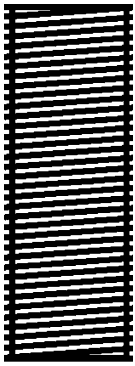


Fig. 5.25 Simplified figure of a solenoid



Experiments show that a long solenoid behaves very much like a bar magnet, with the poles as indicated in the above figure. As such, it is often referred to as an **electromagnet** – a device that owes its magnetization to an imposed electric current. When the current is turned off, the magnetization vanishes. In contrast, a piece of

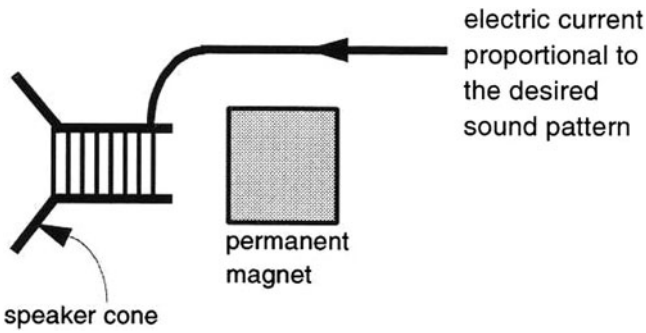


Fig. 5.26 A schematic of a loudspeaker

lodestone or a piece of iron that has been magnetized by a magnet and has retained its magnetization even after the magnet is removed is called **permanent magnet**.

5.7 The Loudspeaker

The solenoid provides the basis for the operation of a **loudspeaker** (represented in Fig. 5.26) as follows. A solenoid is attached to the cone of the speaker. The sound pattern is fed into the solenoid by an electric signal from the amplifier. The amount of displacement of the solenoid and cone depends upon a balance between the force of the neighboring permanent magnet on the solenoid and the force of a spring (not shown in the figure) which tend to pull the solenoid toward an equilibrium position.

The strength of the magnetic force on a solenoid can be increased greatly by inserting a cylinder of iron into the core of the solenoid. Such a unit is used in buzzers, bells, and the telegraph apparatus.

5.8 The Buzzer

In Fig. 5.27, we depict a **buzzer**. When the switch is closed, an electric current flows through the solenoid and attracts the iron plate.

Once the iron plate loses contact with the pointer, electric current ceases to flow through the solenoid and the iron plate drops back into its original position, thus allowing current to once again flow through the solenoid. This cycle is repeated at a high frequency, thus producing the sound of the buzzer. If the plate is attached to a hammer that can strike a bell-shaped piece of metal. The buzzing sound is replaced by the sound of a ringing bell.

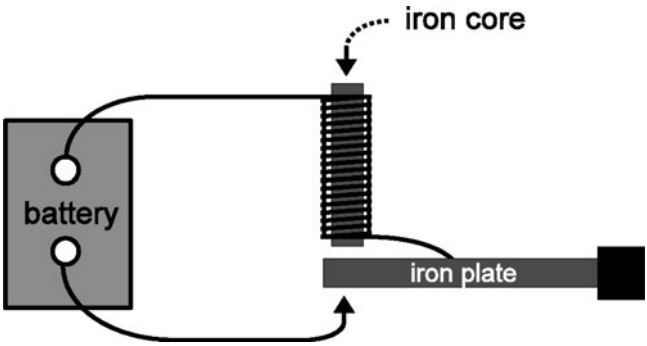


Fig. 5.27 A schematic of a buzzer

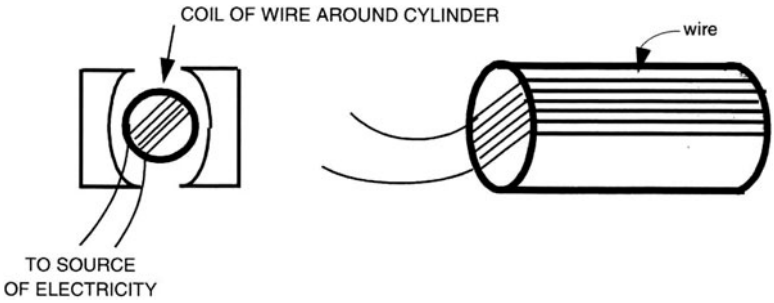


Fig. 5.28 Electric motor

5.9 The Electric Motor

The "**electric motor**" also makes use of the interaction between a permanent magnet and a solenoid electromagnet (Fig. 5.28).

A coil of wire is wound around a cylinder and placed between the poles of a permanent magnet. An electric current is caused to flow through the coil, which acts like a solenoid. The magnetic force on the coil causes the cylinder to rotate so that the "poles" of the solenoid line up with those of the permanent magnet. But by the time the poles line up, the cylinder has "rotational inertia," so that it cannot come to a dead halt. Instead, it moves on. As so far described, there would be a force that would tend to reverse the sense of rotation of the solenoid. Instead, the so-called brushes of the motor cause the current in the coil to reverse its direction, so that the poles are oppositely aligned and the cylinder is caused to rotate further in the same sense. This sequence is repeated again and again as the motor rotates.

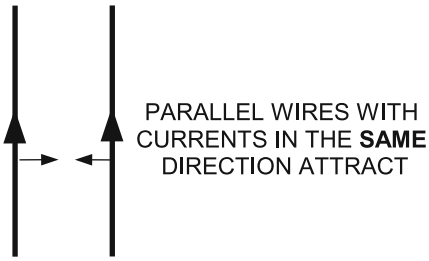


Fig. 5.29 Force between two current-carrying wires

5.10 Force Between Two Wires Carrying an Electric Current

What happens if two wires carrying an electric current are in each other's neighborhood?

They exert forces on each other too! The details were explored by Ampère in 1820, within months of Oersted's discovery. The experiment is depicted in Fig. 5.29. We see that if two parallel wires carry respective currents in the same direction, the wires **attract** each other. (Like attract, in contrast to electric charges!) If the currents are antiparallel, the wires repel each other, as you would expect.

The force between the two wires is not an electric force since the wires are electrically neutral. Instead, physicists concluded that electric currents behave magnetically, not only with respect to their interaction with permanent magnets, but also with respect to their interaction with each other. We do not need permanent magnets to observe magnetic forces. Eventually, experimental and theoretical studies revealed that even permanent magnets owe their magnetism to small, microscopic electric current loops made by electrons within the material. The conclusion was that

All magnetic phenomena are due to electric currents.

5.11 The Electromagnetic Force and Michael Faraday

The next major advance in electricity and magnetism was made by **Michael Faraday**; see Fig. 5.30. in the 1830s through his discovery of the **induced electromotive force** (EMF). Generally speaking, an **EMF** refers to a means whereby electric charges are given an electric force which enables them to move through a material against the presence of internal friction, called **electrical resistance**. Faraday's discovery led to the technologically revolutionary source of EMF called the "**electric generator**".

Before Faraday's discovery, electric currents were produced by attaching wires to an electric **battery** (called a "pile" in England and France, albeit with different pronunciations.) A battery was made by putting together a pile of discs of dissimilar

Fig. 5.30 Michael Faraday (source: http://en.wikipedia.org/wiki/Michael_Faraday)

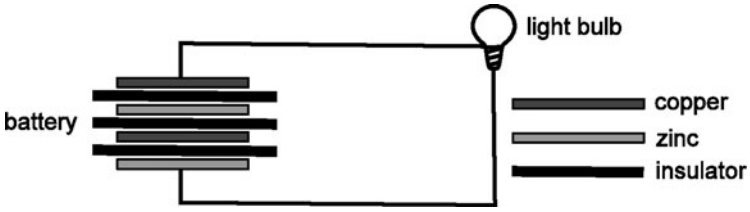


Fig. 5.31 A battery lighting up a light bulb

metals such as copper and zinc, arranged like a sandwich, with sheets of insulator in between the metal discs. In driving currents through wires, a battery is said to produce an **EMF** (Fig. 5.31).

Faraday discovered that if one moves a magnet through a loop of metal wire, an electric current will flow around the loop (Fig. 5.32).

An **induced EMF** is said to drive current around the loop.

If the direction of motion of the magnet is reversed, the direction of the current in the loop is reversed as depicted in Fig. 5.33.

What do you think happens if you reverse the orientation of the magnet?

Suppose now that the *magnet is stationary and the loop is moved to the left*. The relative motion is the same as in Fig. 5.32. The result is the same (Fig. 5.34).

Only the relative motion of the magnet with respect to the loop is relevant.

Given the above observations, electric current was eventually identified as constituting moving fundamental charges, such as electrons in a metal. Oersted's

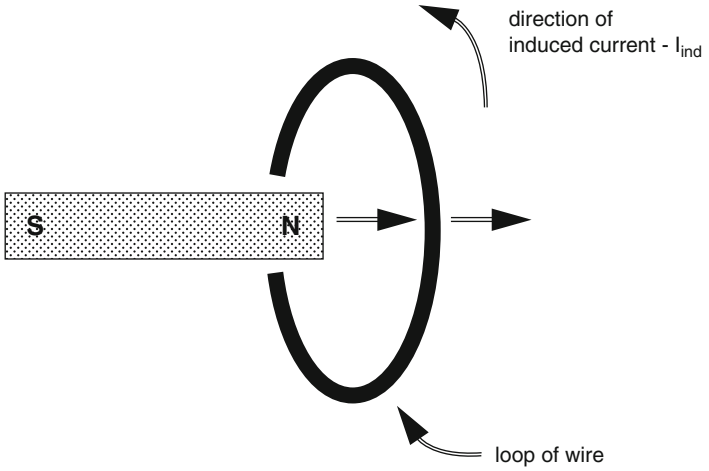


Fig. 5.32 Faraday induction of a current in a metal loop by a moving magnet

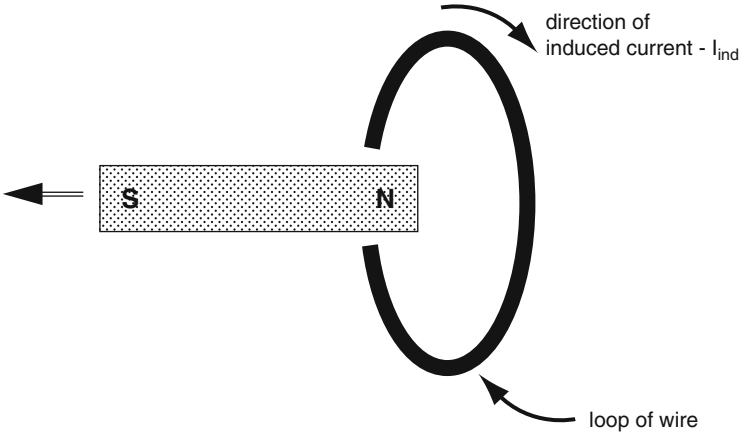


Fig. 5.33 Reversing the direction of motion of the magnet reverses the current direction

observation that a wire with an electric current experiences a force in the presence of a magnet can then be understood on the basis of the principle that:

A moving electric charge experiences a magnetic force in the presence of a magnet.

The fact that an electric current is induced to flow in a wire that is moving in the presence of a magnet can be understood on the basis of this same principle since the wire has electric charges which experience a magnetic force:

Consider a metal wire that is being pulled to the right so as to move between the poles of a magnet as shown in Fig. 5.35a. There will be an induced current

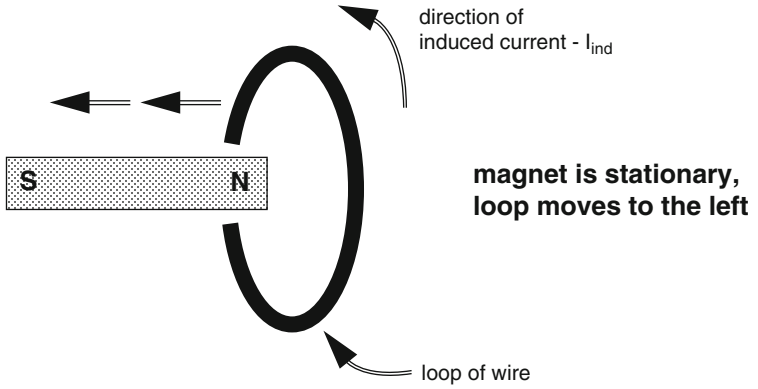


Fig. 5.34 A moving loop with a stationary magnet

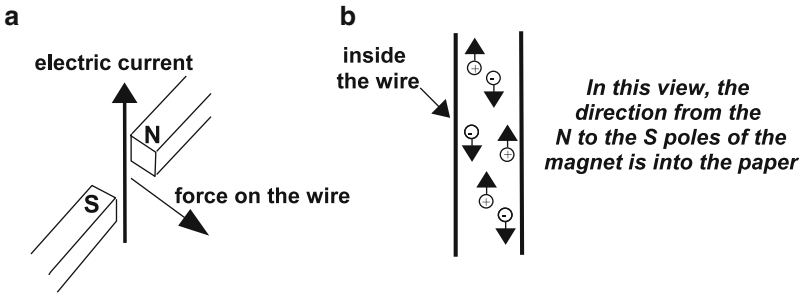


Fig. 5.35 Force on a current-carrying wire by a magnet

flowing upward. We see in the microscopic view presented in Fig. 5.35b that the free electrons experience a **downward** force, perpendicular to the direction of motion of the wire, and move downward so as to produce the upward current. On the other hand, the positive ions experience an upward force but remain close to their equilibrium positions and do not contribute to the current.

However: *From the viewpoint of a second observer who sees the wire stationary and the magnet moving, there is no moving charge and hence no magnetic force. This second observer describes the situation in terms of a changing magnetic field because the second observer observes a moving magnet. Accordingly, the charges in the wire experience an electric field, and therefore an* ***electric force****, in accordance with Faraday's Law.*[2]

[2]The reader who is extremely probing of this process will note that even for the second observer, the electrons in the wire are moving in association with the induced current. As a consequence, the second observer (as well as the first observer) predict that there will be an additional force – a magnetic force for both observers. This force on the electrons turns out to be directed to the left. As a result, an external force must be applied on the wire to prevent it from acquiring a motion to the left.

In the above example, the basic physical principle that accounts for the EMF depends upon the state of motion of the observer with respect to the wire and the magnet. In one case the basis is an electric force, while in the other case the basis is a magnetic force. (For an observer with respect to whom both the wire and the magnet are moving, both a magnetic force and an electric force must be used as a basis!)

Stated differently, the question as to whether a magnetic force and/or an electric force is present is meaningless in the absence of a specification of an observer along with a specification of the state of motion of that observer. We will refer to this aspect of observation as the **Relativity of Description**.[3]

We see then that Faraday's Induced EMF follows from Oersted's observations.

We should note another interesting aspect of the above system: The electrons must overcome electrical resistance as they flow through the wire. This resistance produces thermal energy. Where does this energy come from? Analysis reveals that to maintain the motion of the wire, a force must be exerted on the wire. This force provides the necessary energy.

5.12 Applications of Faraday's EMF

The **microphone** is essentially a loudspeaker in reverse (Fig. 5.36).

A diaphragm is attached to a solenoid. Sound waves impinge on the diaphragm, setting the diaphragm and solenoid in motion. A neighboring magnet induces an EMF in the solenoid which is passed on to an amplifier. The electrical signal from the amplifier can be used to drive a speaker or make a recording.

The **Electric Generator** is essentially a motor in reverse.

---

[3] **This is an application of the Principle of Relativity**. We experience one of its consequences when we sit in a subway train and watch a second train moving relative to us while having poor visibility of any other objects such as a train station. We wonder whether it is our train or the second train that is moving with respect to the tracks. For another common example of this principle, imagine yourself in a car stopped at a red traffic light on an upgrade. You see beside yourself a second car that is slowly moving forward. You then check whether your brake pedal is securely pressed because you worry whether it is in fact your car that is slowly moving backward. In this situation, until you discern the state of motion of the road or some other objects beside the road *relative to yourself*, you are finding it difficult to decide which of the two cars is actually moving with respect to the road. Now imagine yourself in a spaceship in outer space. You look out the window and see a second spaceship moving past you. Which of the two spaceships is moving, you might ask. Such a question has no answer. You can say that the second spaceship is moving with respect to yours, or vice versa. Or, you might decide to investigate the state of motion of both spaceships with respect to the earth and find that both are moving with respect to the earth! It is clear that the state of motion, that is, the velocity of an object is a relative one; it depends upon the observer.

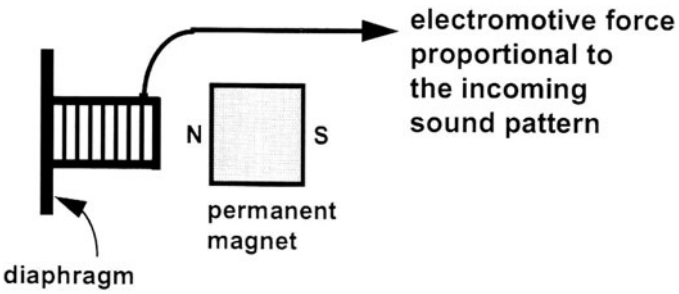


Fig. 5.36 A microphone

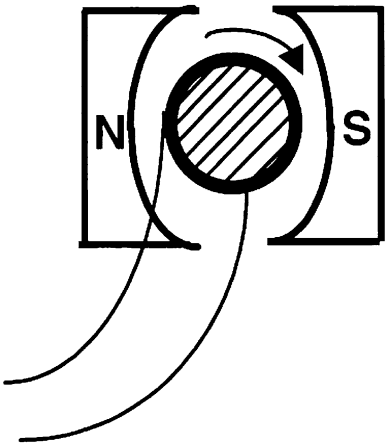


Fig. 5.37 An electric generator using an electromotive force

A cylinder has a coil of wire wound around it. The cylinder is rotated by an external force – say a waterfall or steam engine. The presence of the poles of a permanent magnet produces an EMF in the coil, which is used to run electrical devices. Electric companies use huge generators to "produce electricity." "Producing electricity" refers, in fact, to providing electric power needed to maintain the EMF used by devices hooked up to the companies electric lines (Fig. 5.37).

5.13 A Final "Twist"

We have realized that permanent magnets can be replaced by wires carrying a current; notably, a solenoid with a current behaves like a magnet. Thus, we can produce an EMF in a coil by moving a solenoid relative to the coil. Basically, the source of the EMF can be regarded as simply a magnetic force.

Consider the following alternative: Instead of moving the solenoid, *change the current in the solenoid*. It turns out that this change will also produce an EMF in the coil!

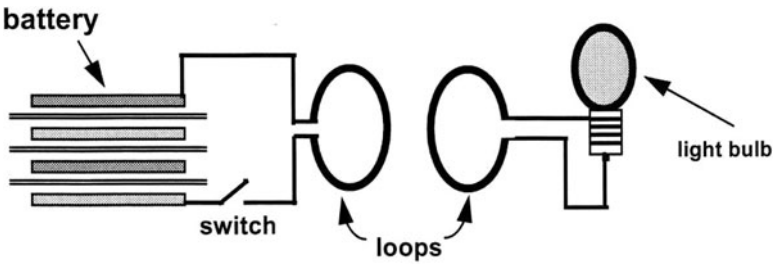


Fig. 5.38 Induction by a wire loop on another wire loop

In the figure below, we depict two wire loops close to each other. The left loop is connected to a switch and a battery. The second loop is connected to a light bulb. When the switch is closed the light bulb lights up briefly (Fig. 5.38).

What has happened is that initially there was no current in the first loop. After the switch is closed, the current in the first loop changes to some stationary value, albeit usually over a very short period of time, perhaps a hundredth of a second. *It is only during this short period of changing current that the second coil experiences an EMF*. In the graphs, we depict the variation in time of both the current in the solenoid and the induced EMF in the adjacent loop. The current does not rise instantaneously once the switch is closed. In this example, it is seen to take about 4 ms (milliseconds) to reach its final value. Also, notice that there is an induced EMF only when the current in the solenoid is changing in time, with a peak value at around 3/4 ms. The induced EMF is proportional to the slope of the graph of the current vs. time (Fig. 5.39).

What is the basic principle behind this phenomenon? It certainly is *not* a **magnetic force**, since no charges are moving initially in the second solenoid. No magnet is moving, so that it does not seem to be a Faraday EMF of the sort we introduced earlier. We have noted that, in the presence of a solenoid carrying a current, we can produce an EMF in a second solenoid *either* by moving one solenoid with respect to the other *or* by changing the current in the first solenoid. An interesting question is:

Are these two methods related?
If so, what is the unifying principle behind them?

5.14 Action-at-a-Distance and Faraday's Fields

In an effort to explain how electric and magnetic forces can act at a distance, Faraday proposed the existence of an "**electric field**" and a "**magnetic field**".

What is a **field**? To answer this question we will cite some familiar examples. Weather reports provide us with the value of temperature, pressure, and wind velocity at various points on a map. All *three* parameters are "fields". The first two

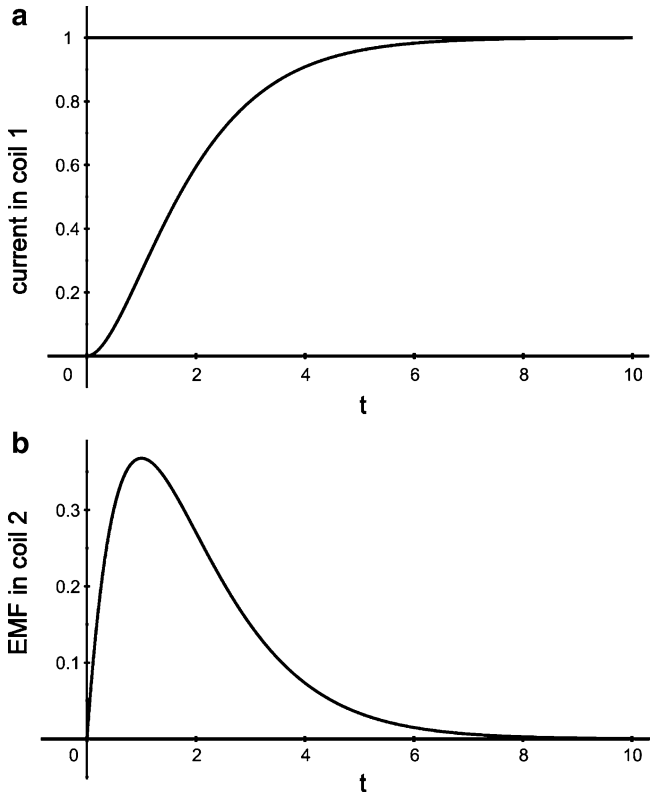


Fig. 5.39 Changing current in one coil inducing an EMF in the second coil

are specified by numbers alone, such as 20°C or 30°C for temperature, and 29 in. mercury or 31 in. mercury for pressure. Wind velocity is specified by a *direction* as well as a *number*, and is an example of a "**vector field**." As we will shortly see, both electric and magnetic fields are vector fields.

Below is an example of the symbol used on weather maps to provide both direction and magnitude for wind velocity (Fig. 5.40).

Meaning of the symbols: The wind velocity at Boston's Logan Airport is 15 knots $\approx$ 17 mph NW. The wind velocity in Springfield center is 20 knots $\approx$ 23mph NE.

5.15 The Electric Field

According to Faraday, an electric charge is accompanied by an **electric field** which "fills" (i.e., is present) throughout space. Below we depict a *positive point charge* with its electric field (Fig. 5.41).

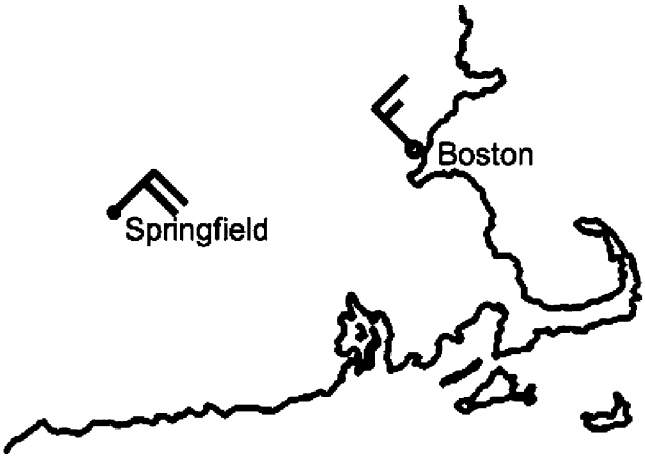


Fig. 5.40 Wind velocity map

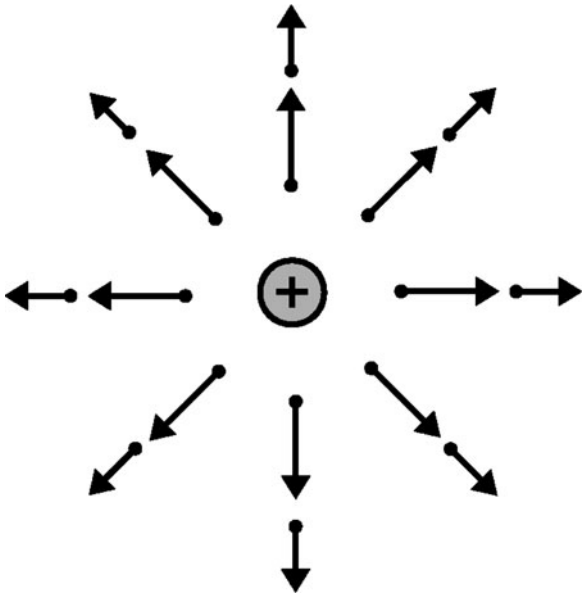


Fig. 5.41 Electric field of positive charge

The direction of the field at the base (the heavy dot) of the arrow is indicated by the direction of the arrow. All arrows point directly away from the point charge. The *length* of an arrow is proportional to the *magnitude* of the field at the corresponding point. We note that the magnitude decreases with increasing distance from the point charge and is the same for equal distances. For a *negative* point charge, the arrows point *toward* the charge (Fig. 5.42).

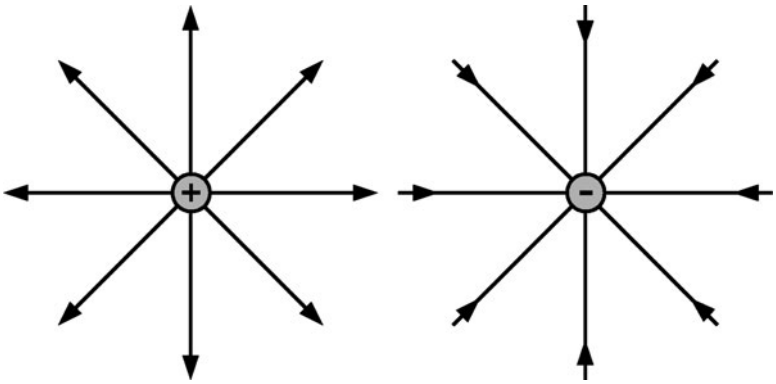


Fig. 5.42 Electric field lines of point charges – a positive charge on the *left* and a negative charge on the *right*

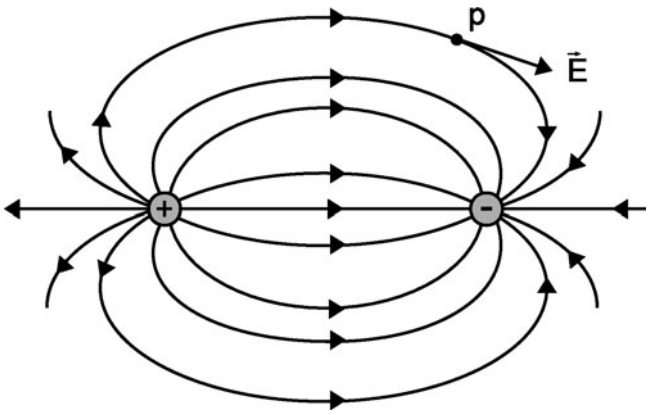


Fig. 5.43 Electric dipole field

We will represent the electric field by the symbol E. An alternative way to represent the electric field on a map is to use continuous "**electric field lines**". See Fig. 5.42.

Here, it is the relative closeness of the field lines that indicate the magnitude of the electric field at various locations in space. Remember that in fact a charge exists in three-dimensional space, so that the figure above shows the field lines in a plane running through three-dimensional space.

While the *magnitude* of the electric field in some region of space is proportional to the density of the field lines (how close together the lines are), the *direction* of the field at a point on a field line is along the *tangent* to the field line at that point (see Fig. 5.43, where the field E at point P is indicated).

If there are many charges present, the total electric field in space is a superposition, that is sum, of the contributions of each charge taken separately.

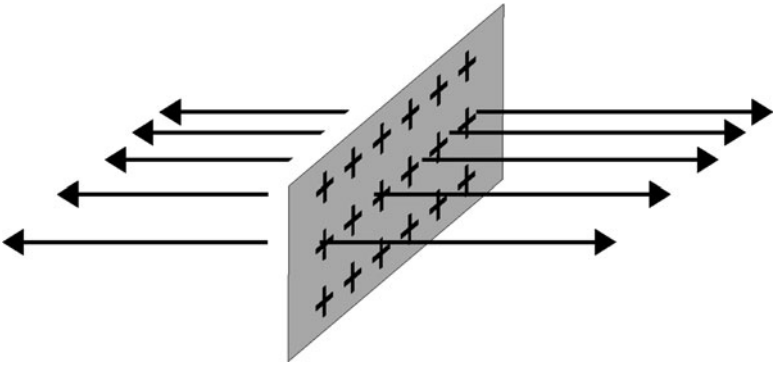


Fig. 5.44 Electric field of a sheet of charge

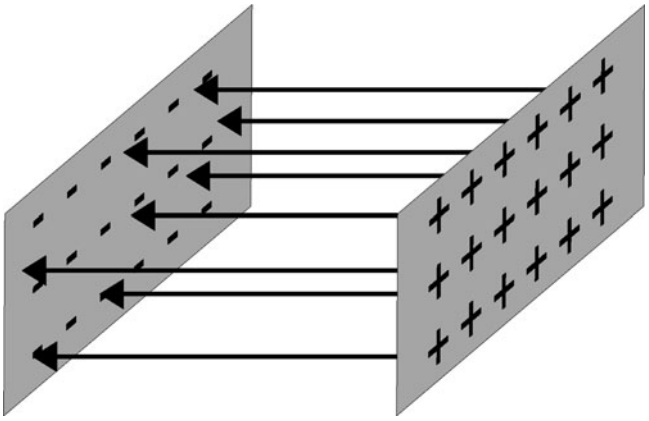


Fig. 5.45 Electric field of two charged sheets

Below are examples of the electric fields of interesting charge configurations.

1. A positive and a negative point charge, side by side – called a "**electric dipole**." See Fig. 5.43.
2. An infinite sheet with a uniform distribution of positive charge. See Fig. 5.44.
3. Two infinite sheets with positive and negative uniform charge distributions, respectively. The electric field is confined between the sheets. See Fig. 5.45.

What is the significance of the electric field? The field represents the *potential* of the charges associated with it to exert a force on an additional charge, say q, placed in the field. Thus, the force $\vec{F}$ (note the arrow on the symbol, since force is a vector) on q due to the changes producing the field $\vec{E}$ is given by:

$$\vec{F} = q\vec{E}, \tag{5.1}$$

where $\vec{E}$ is the electric field at the location of the charge q.

Comment: The electric field of a charge may be likened to the halo drawn by artists above a saintly person or a hero to indicate the potential of the person to influence others in a spiritual or holy way.

The electric field can be regarded as a modification of the space between two charges and thus deals with the philosophical action-at-a-distance issue: The electric force between two charges is mediated by the electric field. But is the electric field a *real* thing? What properties would give it reality? In my opinion, physics cannot answer such a question:

The essential goal of physics is to establish a theoretical framework for describing in a quantitative way what we decide to and are able to measure. That framework makes use of models, concepts, and images. However, its ultimate content is a set of mathematical equations, which we call laws. The laws are as simple and all-encompassing as possible, and provide relationships among measurable quantities.

5.16 The Magnetic Field

Now we turn to **magnetic phenomena**: A magnet or a current-carrying wire is understood to fill space with a magnetic field, which has a specific magnitude and direction at every point in space. We will represent the magnetic field by the symbol $\vec{B}$. There exists a prescription for determining the magnetic field for a given permanent magnet or a current-carrying wire – a prescription which is beyond this course. Below, we have sketched the magnetic field for a number of cases.

1. Bar magnet as seen in Fig. 5.46.
   Note that the magnetic field lines pass through the bar magnet itself. Also note the similarity of the pattern with the electric field lines of an electric dipole.
2. An infinite straight wire with current. Here current is directed out of the paper. See Fig. 5.47.

> **Note**
>
> **The relation between the magnetic field and the current has a complex form** and is referred to as **Ampère's Law** in honor of its discoverer , who was mentioned earlier in this chapter. Essentially, **the magnetic field $\vec{B}$ is proportional to the current I that produces it**. Thus,
>
> $$\vec{B} \propto I. \tag{5.2}$$

3. A long solenoid with tightly wound coils. Compare the field line configuration with that of a bar magnet. See Fig. 5.48.
4. A horseshoe magnet. See Fig. 5.49. Here we have drawn only the field outside the magnet. Noteworthy is the closeness of the magnetic lines between the poles: The field is most intense in this region. Also, between the poles the field lines tend to be straight.

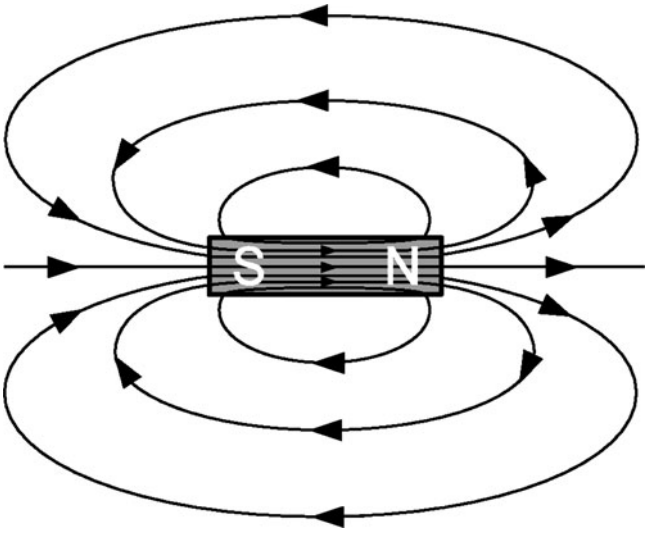


Fig. 5.46 Bar magnet

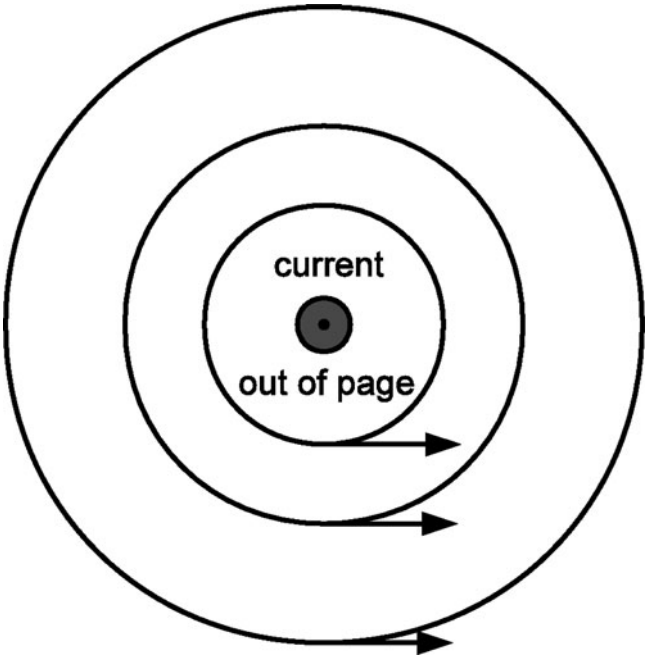


Fig. 5.47 Magnetic field of a straight wire

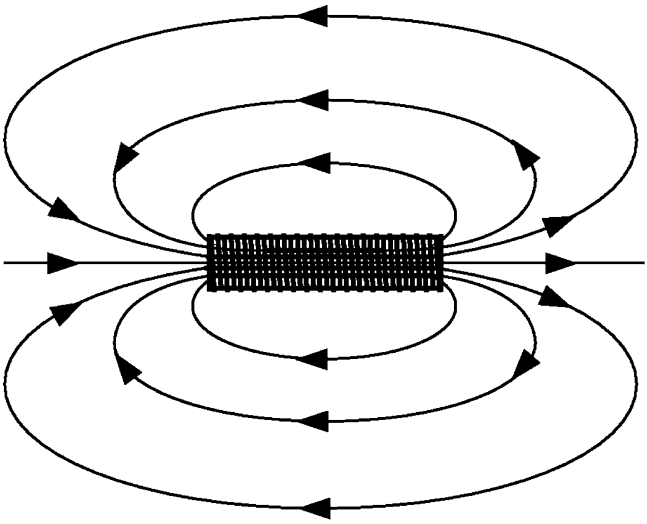


Fig. 5.48 Magnetic field of a solenoid

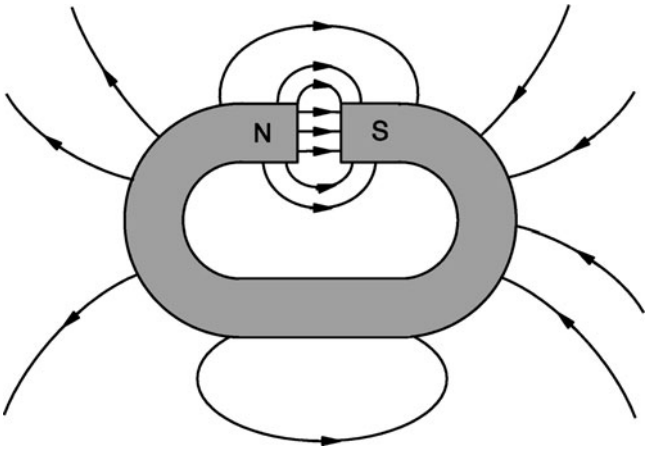


Fig. 5.49 Magnetic field of a horseshoe magnet

Note: Generally, in contrast with the electric field lines described so far, magnetic field lines are "closed" – that is, they have no beginning or end. This fact can be shown to be connected with their being produced by moving electric charge, rather than by what physicists refer to as "magnetic charges." It is also connected with the fact that when a permanent magnet is split in two, we end up with two whole magnets, both with north and south poles. (See Fig. 5.12.) Later we will discuss electric fields that have closed field lines.

How can we determine the direction of the magnetic field at some point in space? The answer is simple. A bar magnet experiences a torque (twisting force)

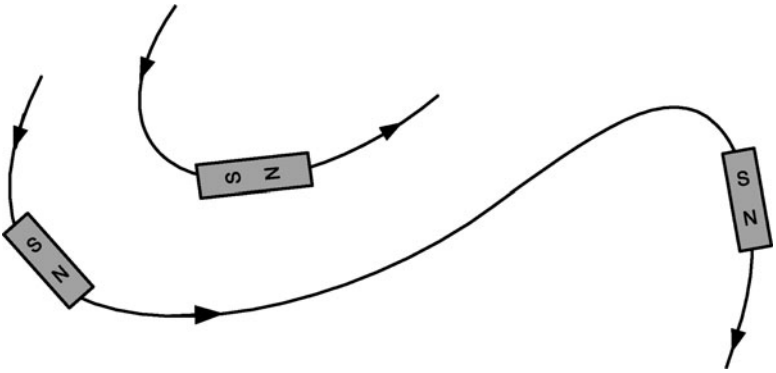


Fig. 5.50 Compass needles line up with the field

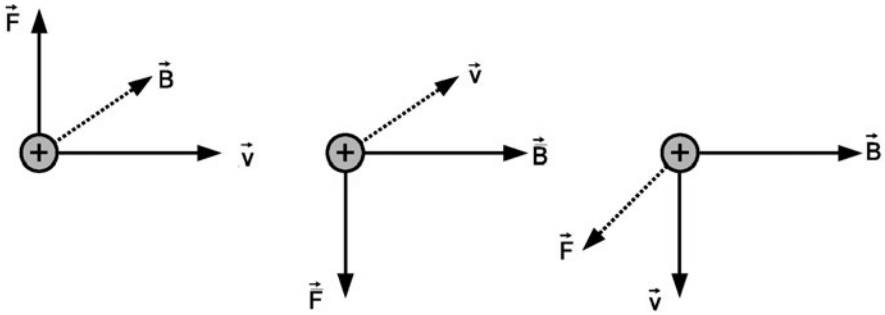


Fig. 5.51 Magnetic force on a point charge

in the presence of a magnetic field which lends it to line up with its *South to North direction* parallel to the direction of the magnetic field. Thus, the bar magnet tends to line up tangent with a magnetic field line. We can therefore use compass needles to determine the direction of the magnetic field. In Fig. 5.50, we depict the orientation of several compass needles in a given magnetic field.

5.17 Magnetic Force on a Moving Charge

Point charge: Generally, the force is perpendicular to both the velocity v of the charge and the magnetic field B at the location of the charge. Also, **the force vanishes if the velocity is parallel to the magnetic field** $\vec{B}$.

In Fig. 5.51, we depict the force on a positive charge with various velocity directions with respect to a magnetic field. (If the charge is negative, the direction of the force is reversed.) Note that the force $\vec{F}$ is perpendicular to the plane determined by $\vec{v}$ and $\vec{B}$.

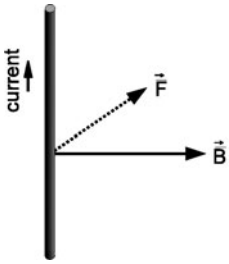


Fig. 5.52 Magnetic force on a wire

For the simple case that the velocity is perpendicular to the magnetic field, that is, $\vec{v} \perp \vec{B}$:

$$F = qvB. \tag{5.3}$$

In Fig. 5.52, we depict a wire that is carrying a current in the presence of a magnetic field that is perpendicular to the wire. (Note that a current in the upward direction can be produced by positive charges that are moving upward **and/or** negative charges moving downward.)

5.18 Force Between Two Parallel Wires Carrying Currents

Each wire produces a magnetic field which, in turn, accounts for a magnetic force on the other wire. With some rather painful analysis, it can be shown that

- The wires **attract** each other if the currents are in the **same direction**.
- The wires **repel** each other if the currents are in **opposite directions**.

The Analysis: In Fig. 5.53 we depict two wires that are carrying currents in the same direction. Let us label the wires #1 and #2, respectively. To find the force of wire #1 on wire #2, we need to find the magnetic field $\vec{B}_1$ due to wire #1 at the position of wire #2. We will label this field $\vec{B}_{1 \text{ at } 2}$. Now let us focus our attention on point P on wire #2.

From the direction of $B_{1 \text{ at } 2}$ and of I_2, we can determine the force $F_{1 \text{ on } 2}$ of wire #1 on wire #2 as being to the left. Thus we see that wire #2 is **attracted** by wire #1.

5.19 Generalized Faraday's Law

Recall that an EMF is produced when a wire is moving in the presence of a magnet – whether it be permanent or otherwise. This result can be understood as being associated with a magnetic force on an electric charge which is moving in the presence of a magnetic field. However, the EMF produced in a stationary wire, in the presence of a moving magnet **or** in the presence of a second wire which carries a current which changes in time, cannot be attributed to a magnetic force. In this case,

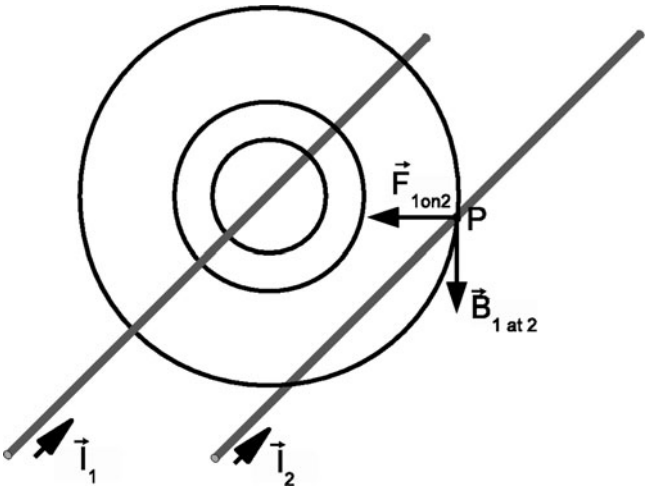


Fig. 5.53 Magnetic force between two long current-carrying wires

a new principle is needed. This principle is embodied in **Faraday's Law**, which can expressed as follows:

Any change in a magnetic field with respect to time is accompanied by an induced electric field.

In more concrete mathematical terms:

$$\vec{E} \propto -\frac{\Delta B}{\Delta t}. \tag{5.4}$$

That is, there is an induced electric field $\vec{E}$ that is proportional to the rate of change (symbolized by the Greek letters Δ) of the magnetic field $\vec{B}$ with respect to time. *The reason for inserting a minus sign will be clear later.*

This induced electric field can drive electric charges through a wire that may be present. Thus, in these cases,

*the EMF is attributed to an **electric** force!*

A number of very important observations are in order:

1. We have seen that a current-carrying wire produces a magnetic field. Thus, it is absolutely clear that changing the current will change the magnetic field and hence produce an electric field.

   Now, suppose a magnet is moving relative to me at some constant velocity v. The magnetic field at some fixed location will change, so I will observe the presence of an electric field. However, suppose you move *with* the magnet, at the same velocity v with respect to me, so that it is *motionless relative to you*. You will therefore observe no change in the magnetic field at any of your fixed locations and so *you will not observe an electric field*!

Thus, the observation of an induced electric field – even the very question of its existence – depends upon the state of motion of the observer. It is not appropriate to ask whether there is an induced electric field. All we can say is that there is an EMF as far as all observers are concerned. How this EMF is accounted for depends upon the state of motion of the observer.

2. Consider again the situation when a wire loop and magnet are moving with respect to each other. There will be a current induced in the loop with a magnitude which is usually very close to being independent of the state of motion of the observer[4].

   Now consider two specific observers: One is at rest with respect to the magnet and observes a moving loop, while the second observer is at rest with respect to the loop and observes a moving magnet. Both observe an induced current. However, they account for the induced current in two different ways: The **first observer** accounts for the current in terms of a **magnetic force** due to the charges in the wire loop moving in a **constant magnetic field**. For this observer, there is **no electric field**. On the other hand, the **second observer** accounts for the current in terms of an **electric force** due to an electric field that is brought about by the moving loop leading to a **changing magnetic field**.
3. We see that the question of the existence of an induced electric field depends upon the state of motion of the observer. So it is with a magnetic field. Consider, for example, that someone who observes an electric charge moving with a constant velocity v will perceive the presence of a magnetic field. Someone else, moving at the same velocity v as the charge, does not observe a moving charge and hence must account for the consequent observations without the presence of a magnetic field from the charge.

   You might feel upset that the question of the existence of an electric or a magnetic field can depend upon the state of the observer. However, we all have to deal with this apparent dilemma in asking the question as to whether an object at rest on the surface of the earth has kinetic energy. With respect to an observer at rest on the earth, the object has no kinetic energy. However, a person at rest with respect to the sun would say that the object is moving around the sun along with the earth and has kinetic energy. Thus, the amount of kinetic energy that an object has depends upon the observer.
4. The two ways of producing an EMF are used in two designs for a microphone depicted below. Both designs work! Microphone (1) works on the basis of a *magnetic* force on charges in the solenoid. Microphone (2) works on the basis of a *electric* force on charges in the solenoid (Fig. 5.54).
5. Let us use the following symbols to indicate the state of motion of an observer, represented by an eyeball.

[4]Einstein's Theory of Special Relativity does predict a dependence of the current on the state of motion of the observer. The dependence is small when velocities are much less than the speed of light in vacuum (3×10^8 m/s).

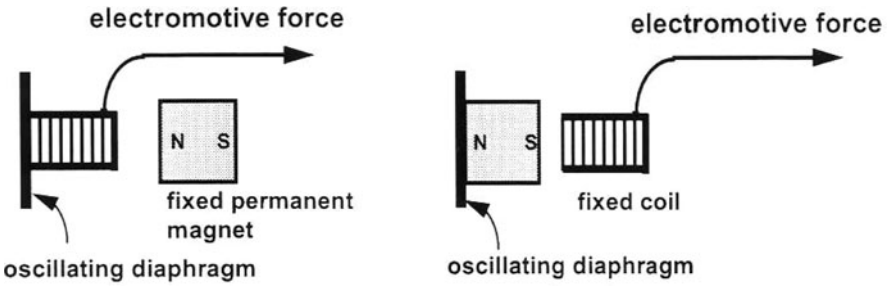


Fig. 5.54 Two designs for a microphone

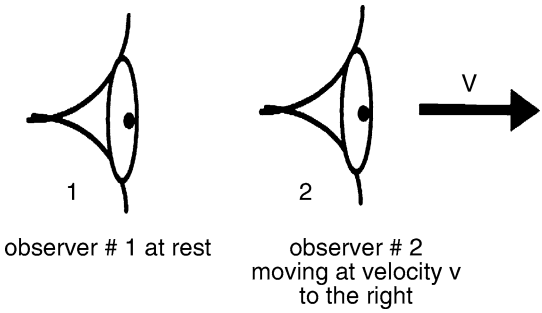


Fig. 5.55 Symbols denoting the state of motion of an observer

In Fig. 5.55 are the two different situations which we have just discussed in pictorial form.

Case I: A permanent magnet is in the presence of various observers (Fig. 5.56). How these observers account for their observations depends upon the relative velocity of the observer and the magnet.

Observer #1 accounts for his/her observations in terms of both a magnetic field and an electric field $\vec{E}_1$. The magnet produces a magnetic field by virtue of its being a magnet. It produces an electric field because of its motion with respect to the observer.

Observer #2 accounts for his/her observations in terms of a magnetic field alone since the observer and the magnet are moving at the same velocity with respect to the paper and hence have no relative velocity.

Observer #3 accounts for his/her observations in terms of both a magnetic field and an electric field $\vec{E}_3$ which is different from $\vec{E}_1$. Note that the velocity v' is at an angle with respect to the velocity v.

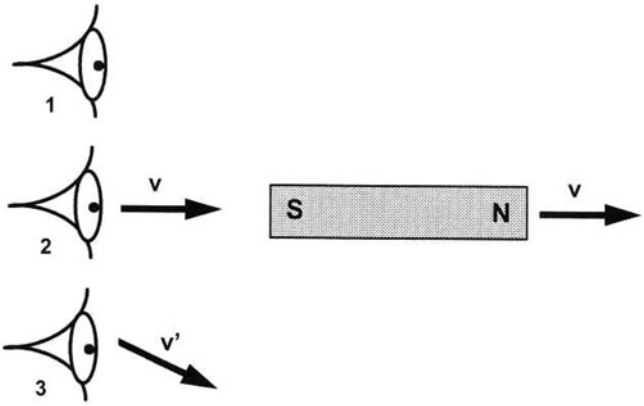


Fig. 5.56 Three observers of a magnet

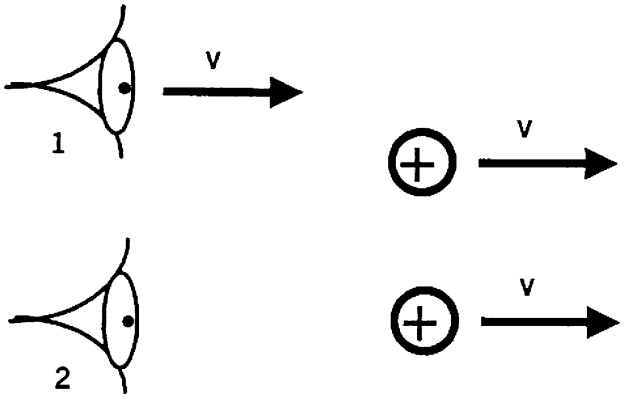


Fig. 5.57 Two different observers of two moving charges

Figure 5.53

Case II: An electric charge is in the presence of various observers. How these observers account for their observations depends upon the relative velocity of the observer and the charge (Fig. 5.57).

Observer #1 accounts for her observations in terms of an electric field alone: Both charges are at rest with respect to this observer. Thus, according to this observer, they exert only an electric force upon each other.

Observer #2 accounts for his observations in terms of both an electric field and a magnetic field. There are both an electric and a magnetic field due to the upper charge. Therefore, the lower moving charge experiences both an electric force and a magnetic force. One can reverse the roles of the two charges.

SUMMARY: Moving permanent magnets, moving electromagnets (wires with electric current), and stationary electromagnets with a changing electric current – all

produce a change in the magnetic field and therefore an induced electric field. If a loop of wire is present, this electric field can drive an electric current and people refer to the presence of an "induced EMF." It is important to realize that *the induced electric field is present whether or not a loop of wire is present*!

5.20 What Do Induced Electric Field Lines Look Like?

Below we show some examples.

1. Moving bar magnet or solenoid.
   In Fig. 5.58, I have drawn closed electric field lines on a rectangular surface. The direction of "rotation" of these lines is connected with the change $\Delta\vec{B}$ in the "B-field" being to the right. Thus, we would get the same direction of the induced electric field to the left of a bar magnet which is moving to the left, as shown in Fig. 5.59. Here, the negative magnetic field is to the right, but the magnitude of the magnetic field is increasing, so that the change $\Delta\vec{B}$ in the magnetic field is to the right.

   We can represent the above schematically in a very much simplified Fig. 5.60. Note that if $\vec{B}$ is pointing to the *right* and its magnitude is decreasing, $\vec{B}$ is pointing to the *left*.
2. Stationary solenoid with a changing current.
   Home exercise: Describe $\vec{E}_{\text{ind}}$ to the right and left of the solenoid.

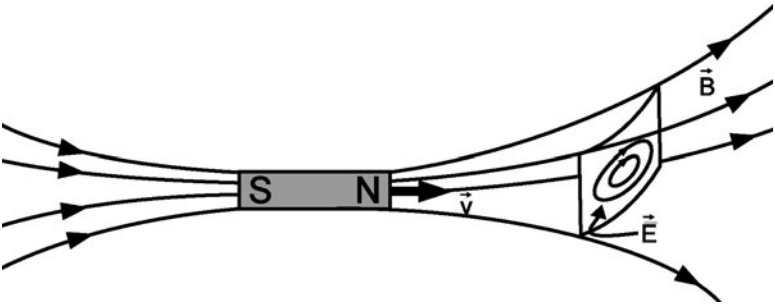


Fig. 5.58 Induced electric field from a moving magnet

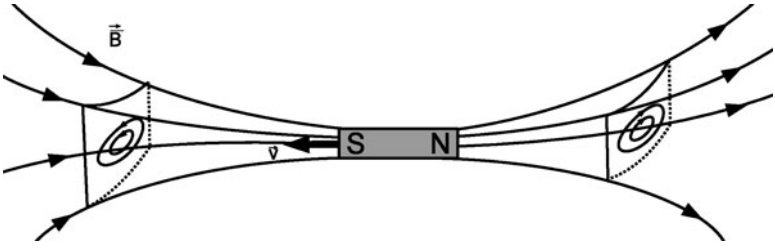


Fig. 5.59 Induced electric field from moving magnet – reversed direction

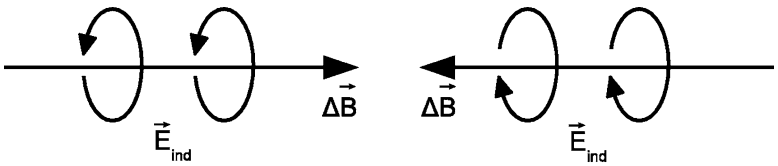


Fig. 5.60 Changing magnetic field leads to an electric field

3. Long straight wire with a changing current.
   Note that the change in current ΔI produces a change in the magnetic field $\Delta \vec{B}$, which is said to induce the presence of an induced electric field $\vec{E}_{\text{ind}}$.
4. Short segment of wire with a changing current.
   Because the wire is infinite in length in case (3), the lines of E_{ind} are parallel straight lines. In the present case, $\Delta \vec{B}$ is concentrated around the wire segment so that $\vec{E}_{\text{ind}}$ is also concentrated. We can also see how $\vec{E}_{\text{ind}}$ has closed.

5.21 Lenz's Law

Notice that in both figures – Figs. 5.61 and 5.62 – the induced **electric** field produced by a changing current is in a direction opposite to that of the change in current, as far as positions along the wire are concerned. Thus, E_{ind} opposed the change in current. The result is that to change the current in a wire, work has to be done to overcome the consequence-induced electric field. That is why extra power is needed while an electric motor is being started up.

Notice too, how the sense of rotation of loops is related to the direction of straight arrows: The relation between E_{ind} associated with $\Delta \vec{B}$ (determined by Faraday's Law) is different from the relation between $\vec{B}$ and the current I which produces that magnetic field. (Here a changing current ΔI is producing a changing $\Delta \vec{B}$.) This difference is reflected by the minus sign in Faraday's Law:

$$\vec{E}_{\text{ind}} \propto -\frac{\Delta \vec{B}}{\Delta t}. \tag{5.5}$$

The above behavior reflects what is referred to as **Lenz's Law**. It states that

The current induced by a changing magnetic field produces a magnetic field that opposes the change in magnetic field.

Let us examine a few experiments to see how Lenz's Law applies.

Figure 5.32: Here we see a magnet moving toward a loop of wire. The magnetic field is thereby increased to the right. The induced current has a direction that produces a magnetic field to the left along the axis. This figure corresponds to Fig. 5.58. To the right of the latter figure, we see the induced electric field lines labeled $\vec{E}$. If a loop of wire was present along an electric field line loop, there would be an induced

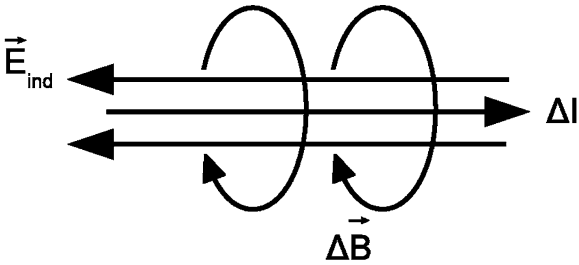


Fig. 5.61 Changing current leads to an induced electric field

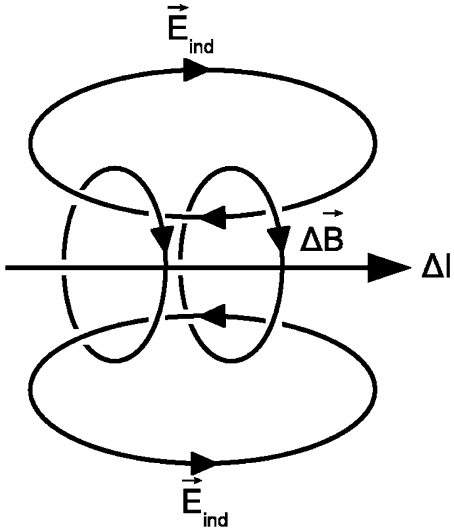


Fig. 5.62 Changing current in a short wire leads to an induced electric field

current around the loop, as seen in Fig. 5.32. This induced current will produce a magnetic field to the left, thus opposing the increased magnetic field to the right.

Figure 5.61: Here, the change in current, ΔI, produces a changing magnetic field, ΔB. In turn, this changing magnetic field produces the induced electric field, $\vec{E}_{\text{ind}}$. This induced electric field produces an induced current (over and above the original current ΔI). This additional contribution to the current is in the same direction as the electric field and thus in a direction **opposite** to the changing current ΔI.

NOTE: Lenz's Law is not an independent, new law. It merely accentuates the significance of the minus sign in Faraday's Law. It is an extremely useful tool for predicting the net qualitative result of Faraday's Law without having to carry out complete mathematical calculation. Furthermore, it has extremely important consequence regarding the stability of electromagnetic systems: Let us suppose that the sign in Faraday's Law was a PLUS sign. Let us consider a ring of metallic wire. It has a huge number of free (mobile) electrons that are moving at incredible speeds in random directions. On average, in equilibrium, their currents essentially cancel and we end up with an essentially vanishing net current. However, these individual currents do not exactly cancel. As a consequence, there are always fluctuating

changing net currents. With a PLUS sign, a changing current would produce an electric field that will increase that current in the same direction. We would have a **positive feedback**. The current can be shown to increase exponentially! The system would be unstable. The actual NEGATIVE sign provides a **negative feedback** and always brings the system back toward its equilibrium vanishing net current.

5.22 The Guitar Pickup

We will now discuss the design of a **guitar pickup** that makes use of **magnetic polarization**. It is based on the following phenomenon: Materials such iron or steel or nickel can be magnetized so as to form a permanent magnet, as described above. These materials are said to be **magnetic materials**. (Aluminum or copper cannot be magnetized and are said to be **nonmagnetic materials**.) Magnetic materials also exist in **nonmagnetized states** – witness a steel paper clip that you buy from the store. However, if a piece of magnetic material is placed in the vicinity of a magnet, the piece can become temporarily magnetic – the technical term is **magnetically polarized**. Often, when the magnet is removed, the piece will return to its nonmagnetic state, so that its magnetic state is dependent upon the presence of the other magnet. A guitar pickup depends upon a return to the non-magnetic state.

In Fig. 5.63, we depict details of a practical design for a common guitar pickup. The permanent magnet polarizes the steel string. The solenoid experiences the magnetic fields of both the permanent magnet and the magnetically polarized vibrating string. When the string vibrates, the magnetic field of the polarized string changes in time, resulting in an induced EMF in the solenoid pickup coil. The induced EMF that is passed on to the amplifier will have a pattern in time that mirrors that of the velocity of the string.

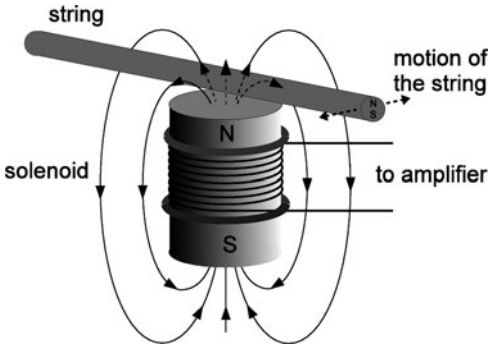


Fig. 5.63 Guitar pickup using magnetic polarization and a Faraday EMF

5.23 Maxwell's Displacement Current

Faraday's Law describes how a change in the magnetic field with respect to time is accompanied by an induced electric field. Around 1860, **James Clark Maxwell** (Fig. 5.64) discovered that **a change in electric field with respect to time must be accompanied by an induced magnetic field**. The rate of change of $\vec{E}$ with respect to time is known as the **displacement current**. In simplified form, we have:

$$\vec{B}_{\text{ind}} \propto + \frac{\Delta \vec{E}}{\Delta t}. \tag{5.6}$$

There is a situation where such a relation would not be surprising: A moving point charge will be associated with a magnetic field and an electric field which changes with respect to time as shown in Fig. 5.65.

Fig. 5.64 James Clerk Maxwell (source: http://en.wikipedia.org/wiki/James_Clerk_Maxwell)

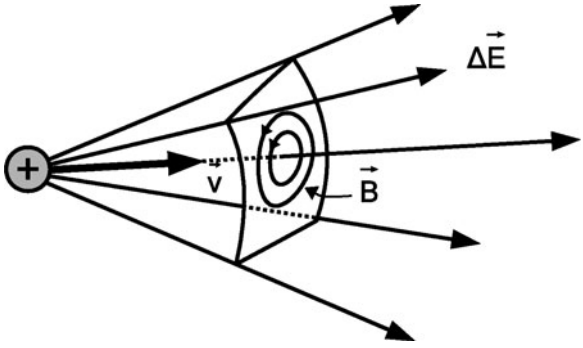


Fig. 5.65 Electric and magnetic fields of a moving charge

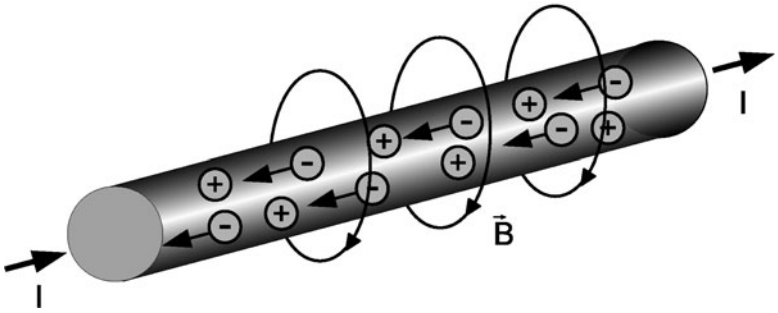


Fig. 5.66 Magnetic field from an electric current

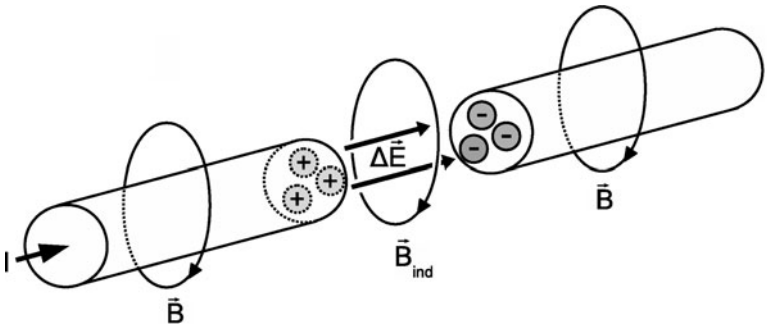


Fig. 5.67 Induced magnetic field due to changing electric field – the **Displacement Current**

However, this latter magnetic field is the one usually associated with Ampère's Law. In order to appreciate the difference between an "Amperian" magnetic field and one associated with the displacement current (and therefore to appreciate Maxwell's contribution), we will consider instead the following two situations:

1. Let us study more closely a metal wire carrying a current. Positive ions are stationary, electrons account for current, and the metal remains electrically neutral. Outside the wire, all we have is a magnetic field (Fig. 5.66).
2. Now suppose that a segment of wire is removed while the EMF which drives current through the wire continues to "pump" electrons in from the right and draw them off from the left. The ends of the wire will accumulate respective positive and negative charge, and we will have a changing electric field in the space between the two ends (Fig. 5.67).

   We observe a magnetic field $\vec{B}_{\text{ind}}$ in the gap region not due to a current, but rather due to a changing electric field, the **displacement current** $\Delta\vec{E}/\Delta t$. Maxwell's great contribution was that more generally, a changing electric field must be accompanied by a magnetic field. The resulting relation led to his deduction that there exists an electromagnetic disturbance which we call **electromagnetic waves** and that light is an example of such a wave.

5.24 Electromagnetic Waves

Now we are ready to see how Faraday's Law and Maxwell's Displacement Current can generate an electromagnetic wave pulse.

Consider first the propagation of a pulse on a long string under tension. The equilibrium state is a straight string. Figure 5.68 shows the string at some early stages after it is being plucked at its center.

Next we turn to the propagation of an electromagnetic wave pulse. Here the equilibrium state is absence of an electromagnetic wave. For a wave in vacuum, there is nothing. We can start the pulse by having a localized electric field $\vec{E}_1$. (One way to do this is by giving an electric charge a sudden jerk – in our example, a downward jerk on a positive charge.) This represents a change in the electric field, so that a magnetic field $\vec{B}_1$ is generated (via the mechanism of Maxwell's Displacement Current). See Fig. 5.69, where the change in field is observed at a position just **below** the charge. In turn, $\vec{B}_1$ generates a second contribution to the electric field, $\vec{E}_2$ (via Faraday's Law). Then, in turn, $\vec{E}_2$ generates $\vec{B}_2$, and so on. See the figure below.

If we add $\Delta\vec{E}_1$ and $\Delta\vec{E}_2$, we obtain an electric field which, along the horizontal axis, has a direction and qualitative magnitude shown in Fig. 5.70.

This process actually takes place continuously in time and space. The result is that an electromagnetic wave pulse propagates outwardly from the source.

Maxwell himself derived specific mathematical equations to express the laws of electricity and magnetism. These equations have an infinite number of solutions which describe combinations of motion of electric charge and electric and magnetic fields. Most significantly, Maxwell showed that there exists a class of solutions which describe the propagation through space of electric and magnetic fields as a so-called electromagnetic wave. Maxwell's analysis provides mathematical rigor behind our description above. In Fig. 5.71 we depict an electromagnetic wave traveling to the right.

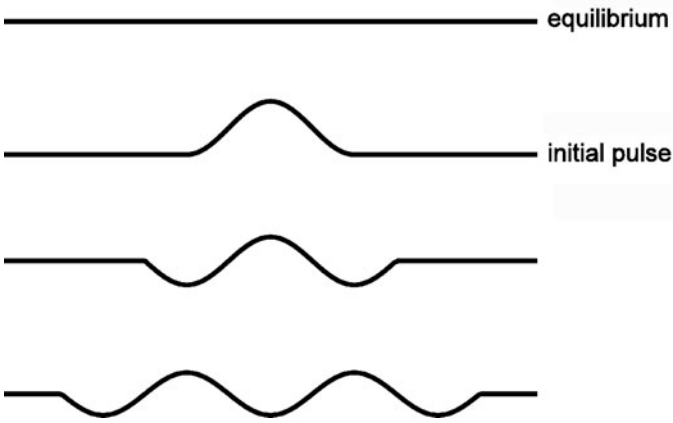


Fig. 5.68 Propagation of a pulse along a stretched string

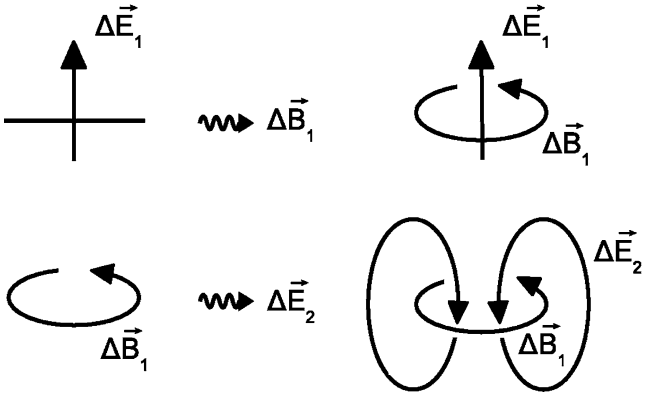
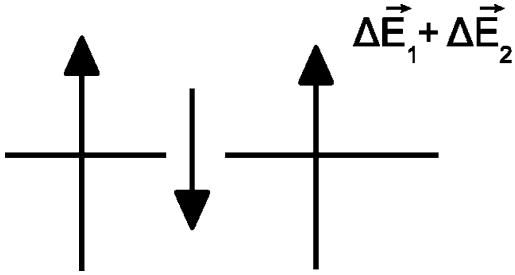


Fig. 5.69 Generating an electromagnetic field – stage 1

Fig. 5.70 Generating an electromagnetic field – stage 2



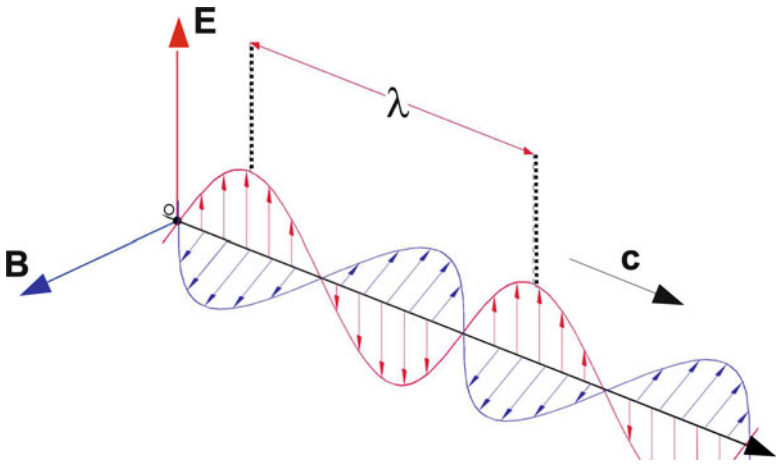


Fig. 5.71 Electromagnetic field – detail

The direction of both the electric and magnetic fields is perpendicular to the direction of propagation. Hence, an **electromagnetic wave is said to be transverse**. In addition, the electric field and the magnetic field are perpendicular to each other.

What is the wave velocity for these EM waves?

Remember the two relevant equations:

$$\vec{E}_{\text{ind}} \propto -\frac{\Delta \vec{B}}{\Delta t}$$

$$\vec{B}_{\text{ind}} \propto +\frac{\Delta \vec{E}}{\Delta t}. \tag{5.7}$$

In these equations, there appear two constants of proportionality involving two constants:

The **permeability of free space** and the **permittivity of free space**.

Maxwell found that the wave velocity of electromagnetic waves – c – is given by:

$$c = \frac{1}{(\text{Permeability of free space} \times \text{Permittivity of free space})^{1/2}} \tag{5.8}$$

The **permittivity of free space** is analogous to the mass density and the **permeability of free space** is analogous to the inverse of the effective force in (2.27).

When Maxwell evaluated this expression, he obtained

$$c = 3.0 \times 10^8 \,\text{m/s}.$$

This value is just the measured value for the speed of light in vacuum and confirmed his identification of a light wave as an electromagnetic wave. Can you imagine how Maxwell might have felt at this discovery!

Comment: The fact that Maxwell's equations have EM waves as solutions without the need for electric charge means that the waves are **self-sustaining**: We do not know of EM waves which were not originally produced by charge. But once the EM waves are produced, the charges may be removed afar. This self-sustaining feature is dependent upon the existence of Maxwell's Displacement Current.

A dramatic example of this self-sustaining property occurs in the phenomenon of **electron–positron pair annihilation**. When these two fundamental particles are close to each other, they have a high probability of both disappearing and being entirely replaced by EM radiation.

Heinrich Hertz (Fig. 5.72) is given credit for being the first to detect EM waves and being able to identify them as such in the framework of the then commonly accepted set of electromagnetic phenomena. A schematic drawing of his apparatus is shown in Fig. 5.73; a photograph of the actual apparatus is shown in Fig. 5.74. In the experiment, a high voltage source produced a spark across a gap in a circuit.

Fig. 5.72 Heinrich Hertz (source: http://en.wikipedia.org/wiki/Heinrich_Hertz)

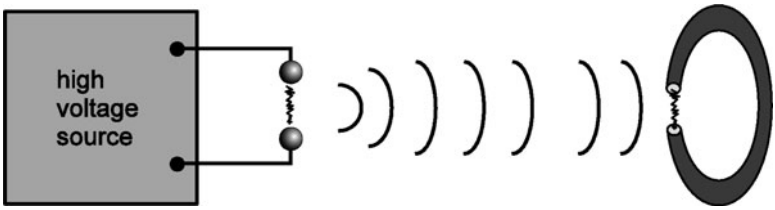


Fig. 5.73 Generating an electromagnetic field from an electric spark

Across the room was a metal ring with a gap. A spark jumped across the ring's gap in response to the original spark.[5]

The first spark involves a changing electric current which produces a changing magnetic field. Here it is a $\Delta\vec{B}$ which is the initial source of the EM pulse. Only an EM wave could account for the great distance traveled by the EM disturbance with such small attenuation.

In time, a wide variety of disturbances, produced under different circumstances, have been identified as EM waves. *The only difference among them is the range of frequencies*. See Fig. 5.75.

[5]Figure 5.74 of Hertz's apparatus was generously provided by John Jenkins, who directs the Spark Museum. For more details about the museum, see its website at: http://www.sparkmuseum.com.

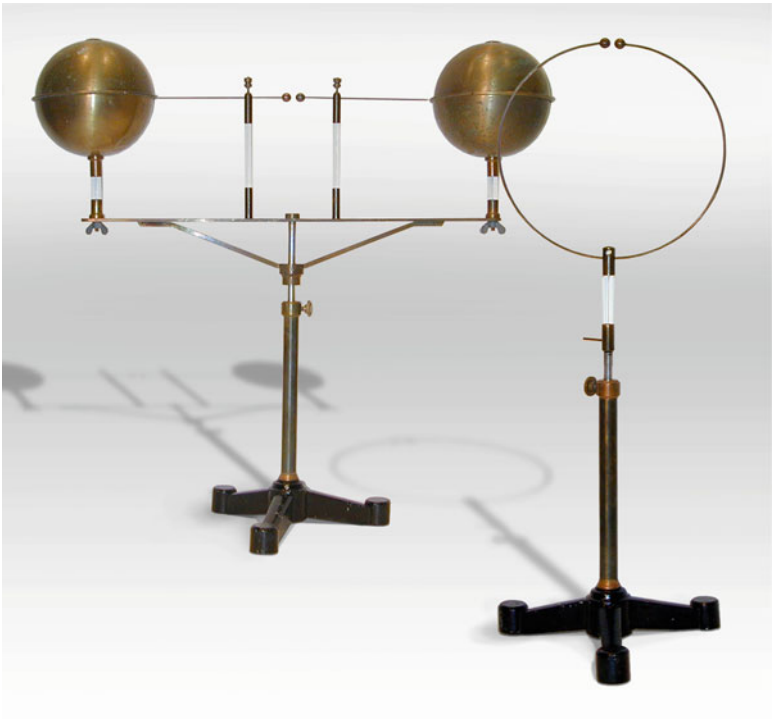


Fig. 5.74 Heinrich Hertz's apparatus for the detection of electromagnetic waves (photo provided by John Jenkins)

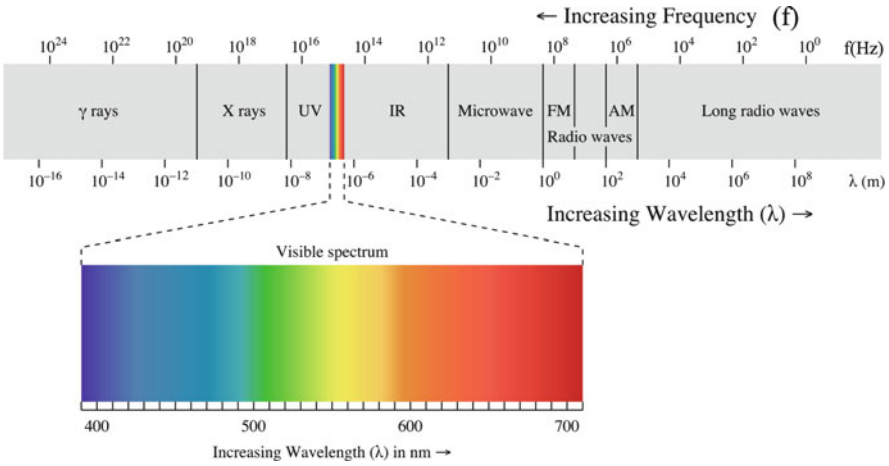


Fig. 5.75 Names of electromagnetic waves for various ranges of frequencies (source: http://en.wikipedia.org/wiki/Electromagnetic_radiation)

5.25 What Is the Medium for Electromagnetic Waves?

Aside from EM waves, all the waves we have discussed so far in the course involve the disturbance of a **medium**:

- Taut string
- Tuning fork
- Sound in air gas, liquid, or solid
- Surface wave on a liquid such as water
- Vibrating Chladni plate or wooden plates of a stringed instrument

Note

Physicists understandably were convinced that EM waves must also involve the disturbance of a **medium**. They called this as yet undiscovered medium the "ether." However, all efforts to reveal the existence of the "ether" failed. The question was settled in 1905 by Einstein's "Theory of Special Relativity."

This theory allows us to describe all observations in terms of a theory in which a medium for the propagation of EM waves, the **ether**, plays no role and may be regarded as nonexistent. The stupendous details, ramifications, and consequences of the Theory of Special Relativity are beyond the scope of this course. You are encouraged to read a layman's account of the theory. Suffice it to say that nuclear weapons and energy were a couple of "by-products."

In satisfying his need to understand the enigmas regarding electric and magnetic fields, Einstein presumably had no foresight as to the awesome consequences of his studies. In fact, the possibility of a nuclear bomb was regarded by most physicists as absurd, even into the 1930s.

Should physicists stop thinking? Some, perhaps many people, feel so. I do not want to belabor this issue much here. I would, nevertheless, encourage you to think about an entirely different area for the sake of comparison because the area is seemingly more benign to us:

About 100 years ago, Sigmund Freud ushered in the modern age of psychology. People have benefitted greatly from this development. However, psychiatry gave birth to the age of manipulation of the masses with highly sophisticated methods of propaganda and advertising. Most glaringly, the Nazi horrors, including tens of millions dead directly due to WWII and about 12 million exterminated victims of the holocaust, by far exceeds the losses to humankind, at least, so far, due to nuclear weapons. *Should we stop studying psychiatry and psychology?*

5.26 The Sources of Electromagnetic Waves

According to the laws of electromagnetism which were formulated by Maxwell, EM waves are produced by **accelerating charge**. Recall that acceleration is the measure of the rate of change in velocity. Thus, a charge that is moving at a constant velocity and therefore is not accelerating, does not produce EM waves, even though an observer uses both electric and magnetic fields to account for the forces produced by the moving charge.

We say that an accelerating charge "radiates" EM waves or emits EM radiation.

Example 1: If a charged particle is oscillating sinusoidally like a simple harmonic oscillator, EM waves will be emitted having a frequency equal to that of the frequency of oscillation. This is the principle behind the operation of radio and TV antennas. Electric charge is "pumped" in and out of the antenna. The electric current runs up and down the antenna. Compare this motion with that of air in the fundamental mode of a sound wave is a semi-closed tube! (Fig. 5.76).

Example 2: **Velocity** is specified by its speed and its direction. **Acceleration** represents a changing velocity, whether the speed is changing and/or the direction is changing. A charge that is moving around a circular path at constant speed still has an ever-changing direction. Therefore, it is accelerating. As a consequence, it will emit EM waves. Not surprisingly, perhaps, the waves have a frequency equal to the frequency of revolution of the charge. See Fig. 5.77.

Example 3: Resonance between two electrically charged simple harmonic oscillators.

In Fig. 5.78, we have depicted two SHOs whose masses have electric charge and whose frequencies of oscillation are identical. Suppose that the left SHO is set into oscillation. It will emit EM waves which reach the SHO to the right. At the position of the right SHO we have an electric field (associated with the EM waves) which oscillates up and down sinusoidally. It therefore produces a sinusoidal force on the

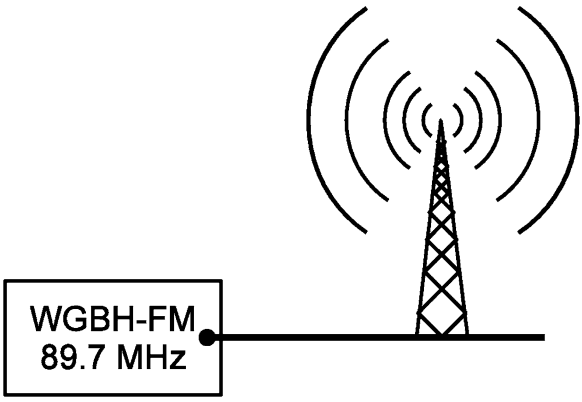


Fig. 5.76 Generating a radio wave

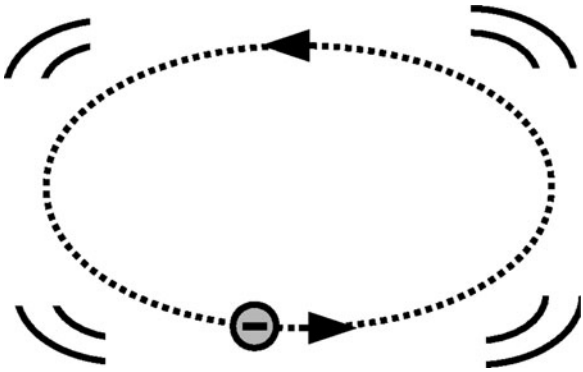


Fig. 5.77 EM field from a circulating charge

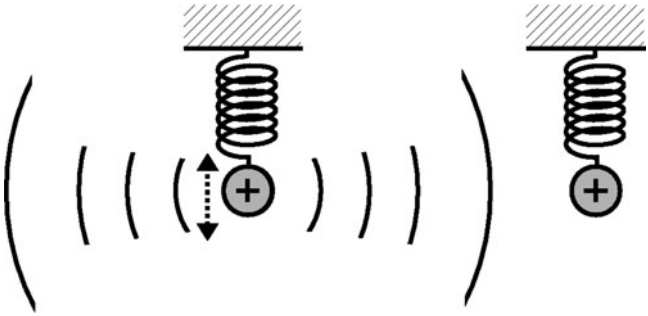


Fig. 5.78 Resonance between two charged SHOs

right charge at a frequency which equals the "natural" frequency of oscillation of the SHO and so causes the right charge to oscillate with a relatively large amplitude. We have a resonant response.

Summary of Electricity and Magnetism

Electromagnetic phenomena are ultimately manifested by forces between electric charges. These forces can be determined, in principle, by the state of the charges alone. Thus,

1. Two charges exert an electric force on each other that is determined by their relative position.
2. Two charges that are both moving exert a magnetic force on each other that depends upon both their position and their relative velocity.

These forces can also be understood in terms of electric and magnetic fields, a concept introduced by Michael Faraday. Thus,

1. A charge produces an electric field
2. A moving charge produces a magnetic field
3. A charge experiences an electric force in the presence of an electric field
4. A moving charge experiences a magnetic force in the presence of a magnetic field
5. An accelerating charge produces a combination of electric and magnetic fields referred to as an electromagnetic wave. Light is an example of an electromagnetic wave

Added note: The question of whether electromagnetic waves are real is not a question within the domain of Physics; it is a philosophical question. However you choose to think of an electromagnetic wave, it is a mathematical quantity that can be used to determine the behavior of electric charges.

5.27 Terms

- Action–reaction
- γ-ray
- *Ampère's Law* (relating current to magnetic field)
- Compass
- Conductor
- Electrically polarized
- Electric battery
- Electric charge
- Electric force
- Electric current
- Electric field
- Electric generator
- Electric motor
- Electromagnetic energy
- Electromagnetic field
- Electromagnetic force
- Electromagnetic wave
- Electromotive force ("EMF")
- Electron-volt (eV)
- Electroscope
- Faraday's law of induction
- Free electrons
- Galilean principle of relativity
- Gravitational force
- Induced EMF
- Insulator
- Lenz's law
- Loudspeaker
- Magnet
- Magnetic field
- Positive and negative charge
- Magnetic force
- Power
- Magnetite (or "lodestone")
- Maxwell displacement
- Metal
- Microphone
- Microwave
- Neutral (electrically)
- Newton's third law
- North and south pole of a magnet
- Nuclear force
- Optical fiber
- Relativity of description
- Solenoid
- Weak force
- X-rays

5.28 Important Equations

Electric force:

$$F \propto qE. \tag{5.9}$$

Magnetic force:

$$F \propto qvB. \tag{5.10}$$

Ampére's Law:

$$B \propto I. \tag{5.11}$$

Faraday's Law:

$$E \propto -\frac{\Delta B}{\Delta t}. \tag{5.12}$$

Maxwell's Displacement:

$$B \propto +\frac{\Delta E}{\Delta t}. \tag{5.13}$$

5.29 Problems for Chap. 5

1. What is an "electromagnet"?
2. Where is the magnetic N-pole of the earth? (Approximately?)
3. What was Oersted's discovery?
4. Draw the electric field lines of an electric dipole.
5. Draw the magnetic field lines of a bar magnet.
6. Magnetic field lines are: always closed / may be closed / never closed.
7. Electric field lines are: always closed / may be closed /never closed.
8. State Faraday's Law.
9. How do the following work? **motor**; **generator**; **loudspeaker**.
10. Describe the operation of a **microphone** with two different designs, using fundamental physics principles.
11. Describe a situation which illustrates the Principle of Relativity in electricity and magnetism.
12. Describe a situation wherein the presence of when a electric field depends upon the motion of the observer.
13. Describe a situation wherein the presence of a magnetic field depends upon the motion of the observer.
14. What is Maxwell's Displacement Current?
15. Electromagnetic radiation is produced when an electric charge is behaving in what way?
16. Calculate the wavelength of microwaves in a microwave oven having a frequency of 2,500 MHz. On the basis of your answer, discuss the effect on cooking of the nodal lines of a standing wave that could set up in the microwave oven.

Chapter 6
The Atom as a Source of Light

We have noted that according to Maxwell's theory of electromagnetism, light is nothing but a visible electromagnetic (**EM**) wave that has a frequency in the narrow range $\sim 4 \times 10^{14}$ Hz to $\sim 7 \times 10^{14}$ Hz. The corresponding range of wavelengths is 4,000–7,000 $\mathring{A}$. Furthermore, EM waves are produced and emitted by accelerating electric charge. There are two interesting questions that immediately confront us: (1) We accept the premise that animal eyes evolved so as to be sensitive to sunlight. Still, what characteristics of animal eyes makes them sensitive to this particular narrow range of frequencies? (2) What are the physical characteristics of the sun that causes sunlight to be concentrated in a particular range of frequencies?

The answer to the first question is connected with the fact that the eye uses conglomerates of atoms, that is, molecules, as detectors of EM waves. This chapter therefore focuses on the atom as a source and receiver of EM waves. The answer to the second question has to do with the fact that the sun is a body that is in equilibrium, with a surface temperature of about 5,800° above absolute zero (that is 5,800 Kelvin / 5,800 K = 6,073°C). That there is a connection between temperature and frequency must be surprising to the beginning student of physics; as we will see later on this chapter, the connection stems from their common link with energy.

6.1 Atomic Spectra

Matter is constantly emitting EM radiation and absorbing EM radiation which strikes it. Ordinarily this EM radiation is invisible and of low intensity. However, when matter is heated up from room temperature, it emits EM radiation which becomes more intense and eventually, in a certain range of temperatures, becomes visible. When dilute gases of atoms are heated up sufficiently, as they are in a star, we can see EM radiation which comes directly from the star's atoms. That radiation reflects the behavior of a single atom. The frequency spectrum of radiation around

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_6,

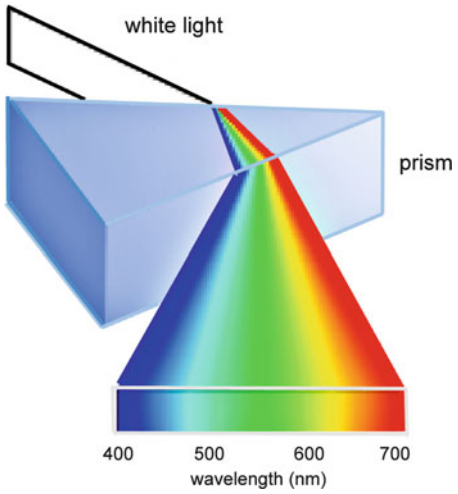


Fig. 6.1 Prism analyzing white light (source: http://en.wikipedia.org/wiki/File:Dispersive_Prism_Illustration_by_Spigget.jpg)

the visible range can be determined using a **diffraction grating** or a **prism**. (How these devices work will be discussed in Chaps. 7 and 8, respectively.)

The experimental setup is depicted in Fig. 6.1, wherein a beam of light from a light bulb is passed through a slit so as to produce a narrow column beam. This column beam is passed through a prism, out of which comes a column beam that is fanned out spatially. The higher the frequency, or the lower the wavelength, of a Fourier component of the incoming beam is, the more is it deflected away from the incoming direction. (See Sect. 8.7 in the text, which shows how this effect is a result of **dispersion** and **refraction**.) This outgoing beam is incident upon a screen, which can be viewed and analyzed.

Because the incoming beam is close to being white light, the spectrum will be a continuous **rainbow** spectrum. In Fig. 6.2, we exhibit the image that is produced on the screen by a beam of light emitted by various gases of atoms which have been heated up to a very high temperature (thousands of Kelvins) or has been subjected to a very high electric voltage. Because a column beam has been analyzed, each component shows up as a *line*. For this reason, scientists make reference to the **line spectrum** of an atom, which can be compared to the sound spectrum of a musical instrument. See Fig. 6.2 for the line spectra of hydrogen and of iron.

We see that the frequency spectrum of an atom is **discrete**, as opposed to the spectrum of sunlight, which is **continuous**. Each atom has its own unique spectrum, which can serve as its "fingerprint" for identification purposes. This fact allows us to determine which atoms and molecules are present in outer space, such as in stars, quasars, and interstellar gases.[1]

[1]The following website (12-29-2010) enables you to see the emission spectrum of elements shown in the periodic table. http://chemistry.bd.psu.edu/jircitano/periodic4.html.

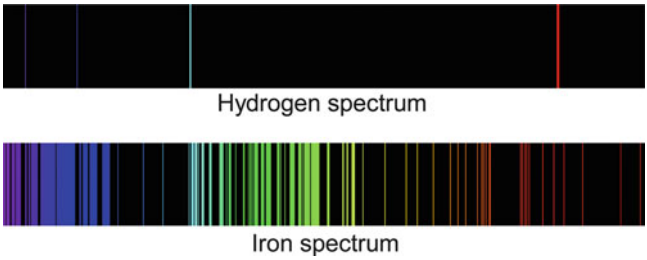


Fig. 6.2 Spectra of hydrogen and iron (source: http://en.wikipedia.org/wiki/Emission_spectrum)

6.2 The Hydrogen Spectrum of Visible Lines

The following are the frequencies of the visible spectral lines of hydrogen, which is the simplest element:

$$f_1 = 4.57 \times 10^{14}\,\text{Hz}$$
$$f_2 = 6.17 \times 10^{14}\,\text{Hz}$$
$$f_3 = 6.91 \times 10^{14}\,\text{Hz}$$
$$f_4 = 7.32 \times 10^{14}\,\text{Hz}.$$

The fundamental question is: What is the theoretical basis for these spectra?

To appreciate the meaning of the above question, let us review the case of the vibrating string with fixed ends. The frequency spectrum is a harmonic series:

$$f_1,\ f_2 = 2f_1,\ f_3 = 3f_1,\ \ldots.$$

What counts essentially are the ratios of the frequencies of the overtones to the fundamental frequency, $1 : 2 : 3 : 4 : \ldots$. The fundamental frequency sets the scale for the whole spectrum. Recall that we can express the frequency spectrum as follows:

$$f_n = nf_1 \quad \text{where } n = 1, 2, 3, \ldots. \tag{6.1}$$

We showed that this spectrum follows from the fact that the modes of vibration of a string have periods which are an integral *fraction* of the time that it takes for a pulse to make a round trip along the length of the string. Since according to Maxwell's theory, the source of EM radiation is accelerating charge, a reasonable hypothesis to account for atomic spectra would be that:

1. The charges in an atom have modes of vibration associated with atomic forces. In order to determine the modes an as yet unknown model of the atom would be needed.

2. Each particular frequency in the frequency spectrum of an atom is associated with a particular mode of vibration. In order to understand this idea, recall the charged SHO of Sect. 5.26 and exhibited in Fig. 5.78.

Unfortunately, no one could find a model for an atom that accounted for the observed atomic spectra. A hint was provided in 1884 by Johann Balmer, who discovered that the observed visible spectrum of hydrogen, that is, the four frequencies given above, was fitted well with the following complicated formula:

$$f_n = 3.29 \times 10^{15}\,\text{Hz} \cdot \left[\frac{1}{4} - \frac{1}{(n+2)^2}\right]. \tag{6.2}$$

Thus, for example,

$$f_1 = 3.29 \times 10^{15}\,\text{Hz} \cdot \left[\frac{1}{4} - \frac{1}{(1+2)^2}\right] = 4.57 \times 10^{14}\,\text{Hz}. \tag{6.3}$$

Two points must be stressed: *First*, it can be proved that any finite set of numbers (here four numbers) can be fitted precisely with any one of an infinite number of formulas such as (6.2).[2] *Second*, Balmer's formula had no theory to give it physical significance when it was first presented. It was purely **empirical**.

Formulas similar to Balmer's were later found to fit the observed spectral lines of hydrogen that lie in both the ultraviolet region (i.e., just *above* the visible frequency range) and in the infrared region (i.e., just *below* the visible frequency range). Later, Neils Bohr gave Balmer's formula a theoretical basis. As we will shortly see, the formulas did provide Bohr with a clue to his theory in which it involved a difference between two numbers (the two terms between the brackets of (6.2)).

It is interesting to note that after Maxwell showed that there must exist a displacement current contribution to the magnetic field, and identified light as an electromagnetic wave, it was believed that the existing set of fundamental laws of physics was complete; that is, the laws could in principle account for all observations thus far and henceforth to be made.[3] This set of laws, along with the accompanying concepts, is referred to as **classical physics**. Unfortunately, as far as atomic spectra were concerned, no one succeeded in finding a model for an atom that accounted for the observed spectra. In fact, it ultimately became clear that classical theory

[2] As an example, the set of numbers, $\{1, 2, 4\}$, can be represented by 2^{n-1} *or* by $[n(n-1)/2 + 1]$, where $n = 1, 2,$ and 3, respectively.

[3] The term **"in principle"** means that one merely had to solve the mathematical equations and one would find that the theory would be confirmed by experiment. In practice, there are many phenomena that require the solution of equations that are too difficult and complicated to solve, so that the theory cannot be tested. However, unsolvability does not imply that the equations and the theory that they represent are incorrect.

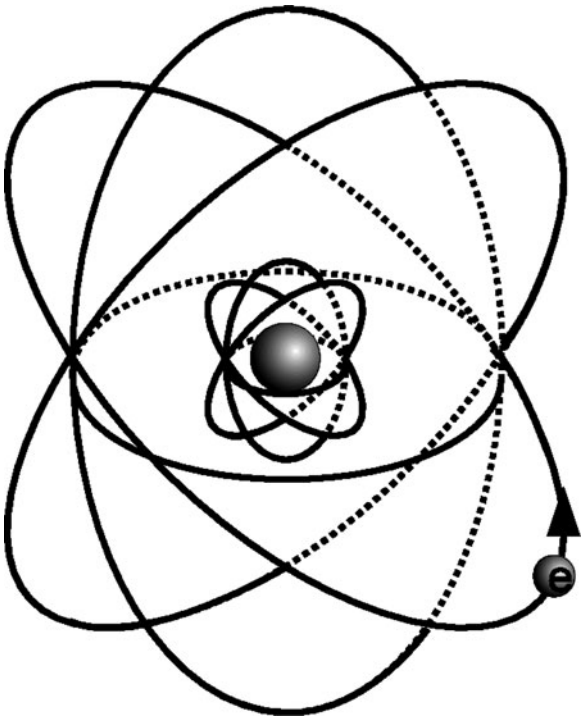


Fig. 6.3 Electron orbits in the Rutherford model

was inadequate and would have to be modified or improved so as to account for phenomena at the atomic level. Thus, limitations of mathematical solvability were not the issue here.

In 1911, on the basis of experiments of alpha particles scattered by atoms of gold, **Ernest Rutherford** proposed a model for the atom in which a collection of negatively charged electrons revolve in planetary-like orbits about a positively charged nucleus. The radius of the electron orbits is on the order of Ångstroms. ($1\ \mathring{A} = 10^{-10}$ m.) The nucleus is relatively utterly minuscule, with a diameter that is about 1/100,000 that of the atom as whole and therefore occupying only about one part in $(1/100{,}000)^3 = (1/1{,}000{,}000{,}000{,}000{,}000) =$ one-thousand-trillionth of the volume of the atom as a whole. Since the electrons are regarded as being much smaller than the nucleus, most of an atom's volume is empty space! We see a reflection of Rutherford's model, expanded upon by Bohr, in the common symbol for an atom (Fig. 6.3).

There was a major difficulty with the Rutherford model in the context of Classical Physics in which the atom is unstable.

Suppose an electron is orbiting a nucleus at some radius. Because of its acceleration, the electron will emit EM radiation having a frequency equal to the frequency of revolution. Also, the radiation has energy, so that the orbiting electron must lose energy, which here is a sum of kinetic energy (KE) and potential energy (PE).

Since the radius of the orbit decreases with decreasing energy, the electron will spiral into the nucleus. Furthermore still, since the frequency of revolution decreases with decreasing radius, the radiation will have an ever continuously decreasing frequency. And finally, classical theory predicts that an electron that starts out at a radius of $1\mathring{A}$ would spiral into the nucleus in about one-billionth of a second! Thus, an atom would collapse and thus be quite unstable.

Two responses were reasonable at this point in the search for a theory of the atom:

1. One could search for a new model while keeping the classical laws.
2. One could keep the basic Rutherford model but find new laws.

6.3 The Bohr Theory of the Hydrogen Atom

When there is disagreement with experiment, indicating a need for revised laws, physicists try hard to preserve as much of the essence of existing laws. Such was the case when, in 1913, **Neils Bohr** proposed a theory of the hydrogen atom that incorporated the Rutherford model but combined classical laws with a modification that restricted the orbits (Fig. 6.4).

We can summarize the theory as follows:

1. According to classical theory, the electron orbits are ellipses, with any **size** or degree of flatness (referred to as "eccentricity") being possible. Circular orbits can have any radius. Bohr proposed that only certain discrete orbits are possible. We exhibit Bohr's discrete concentric circular orbits in Fig. 6.5.

Fig. 6.4 Neils Bohr (source: http://en.wikipedia.org/wiki/Niels_Bohr)

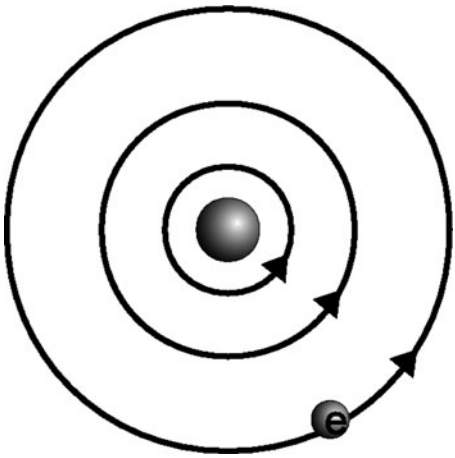


Fig. 6.5 Bohr orbits of the hydrogen atom

In Bohr's theory, the radii are equal to the following multiples of the so-called Bohr radius, which is equal to about $0.53\mathring{A}$, and to which we will give the symbol a_0:

$$r = a_0,\ 4a_0,\ 9a_0,\ 16a_0,\ \ldots.$$

$$\text{Generally, } r = n^2 a_0, \text{ where } n = 1, 2, 3, \ldots. \tag{6.4}$$

The allowed orbits correspond to certain allowed energies:

$$E = -13.6\,\text{eV}, -3.4\,\text{eV}, -1.5\,\text{eV}, -0.85\,\text{eV}, \ldots$$

$$\text{or generally, } E = -13.6\,\text{eV}/n^2, \text{ where } n = 1, 2, 3, \ldots. \tag{6.5}$$

Note that while the energies are **negative**, the energies of the orbits increase (are less negative) with increasing radius. We will see that the only importance of these energies is the difference between pairs of energies, so that their being given negative values is inconsequential.

2. Next we consider Bohr's theory of emission of EM radiation. According to Bohr, this takes place not because of the acceleration of charge, but rather in association with *transitions* of the electron from one orbit to another of lower energy. A transition from the $n = 2$ orbit to the $n = 1$ orbit is depicted in Fig. 6.6.

   A transition from one orbit to another of lower energy is accompanied by the emission of a specific discrete amount of EM radiation. This unit of EM radiation is called a **photon**, which is represented by the wiggly arrow in the figure. By the Principle of Conservation of Energy, the photon must have an energy equal to that lost by the atom.

   Thus, photon energy equals energy lost by an atom or

$$E_{\text{ph}} = E_{\text{i}} - E_{\text{f}} \tag{6.6}$$

which is also called the photon emission condition.

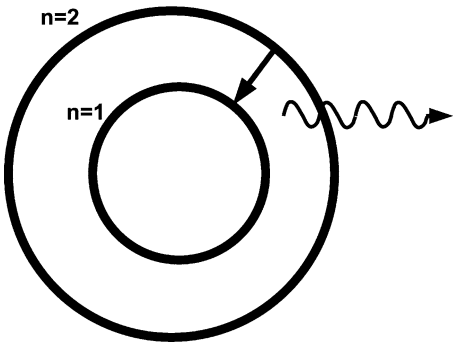


Fig. 6.6 Transitions in the Bohr model that lead to emission of EM radiation

The emission takes place without the need of an interaction of the atom with an external system. The process is therefore referred to as **spontaneous emission**. In (6.6), E_i and E_f are the initial and final energy of the atom, respectively. For example, if the atom makes a transition from the $n = 2$ orbit to the $n = 1$ orbit, the emitted photon has an energy

$$E_\mathrm{ph} = (-3.4\,\mathrm{eV}) - (-13.6\,\mathrm{eV}) = 10.2\,\mathrm{eV}.$$

What are the characteristics of a photon?

(a) A photon has a *specific frequency*. In fact, monochromatic (sine wave) EM radiation of a given frequency f consists of a collection of photons having the same frequency.

(b) A photon has a *specific energy* related to its frequency via the **Planck Relation**:

$$\text{Energy of photon} = E_\mathrm{ph} = hf. \tag{6.7}$$

In (6.7), h is a universal fundamental constant of nature known as **Planck's Constant** – named after **Max Planck**. It has the value

$$h = 4.1 \times 10^{-15}\,\mathrm{eV\ per\ Hz}. \tag{6.8}$$

Thus, suppose that the radiation has a frequency $f = 5.0 \times 10^{14}$ Hz. Then,

$$E_\mathrm{ph} = 4.1 \times 10^{-15} \cdot 5.0 \times 10^{14} = 2.0\,\mathrm{eV}.$$

The condition of (6.7) for photon emission then reads

$$hf = E_\mathrm{i} - E_\mathrm{f}. \tag{6.9}$$

Therefore, the emitted photon has a frequency given by

$$f = \frac{E_\mathrm{i} - E_\mathrm{f}}{h}. \tag{6.10}$$

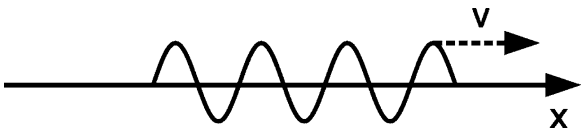


Fig. 6.7 A photon wave packet

Sample Problem 6-1

Find the frequency of the photon emitted by a hydrogen atom that makes a transition from the $n = 2$ orbit to the $n = 1$ orbit.

Solution

The energy of the photon is

$$E_{\text{ph}} = E_2 - E_1 = (-3.4) - (-13.6) = 13.6 - 3.4 = 10.2\,\text{eV}.$$

So that its frequency of the photon is

$$f = E_{\text{ph}}/h = \frac{10.2}{4.1 \times 10^{-15}} = 2.5 \times 10^{15}\,\text{Hz}.$$

This frequency is in the invisible ultraviolet region.

(c) A photon also has an associated *wavelength*, which is related in the usual way to the frequency:

$$\lambda = \frac{v}{f} = \frac{c}{f}. \tag{6.11}$$

(d) A photon usually is *localized in space*, both longitudinally, in the direction of its motion, and transversely. We will describe only the longitudinal extent, which we will refer to as the **length of the photon**. A photon that is emitted by an atom is about ten million oscillations in extent. Such a finite segment of a sine wave is generally called a **wave packet** and is depicted in Fig. 6.7.

Sample Problem 6-2

Find the wavelength and the length of a photon of frequency 5.0×10^{14} Hz and ten million (10^7) oscillations in extent?

Solution

From (6.11), we have a wavelength

$$\lambda = \frac{c}{f} = \frac{3.0 \times 10^8}{5.0 \times 10^{14}} = 6.0 \times 10^{-7}\,\text{m}.$$

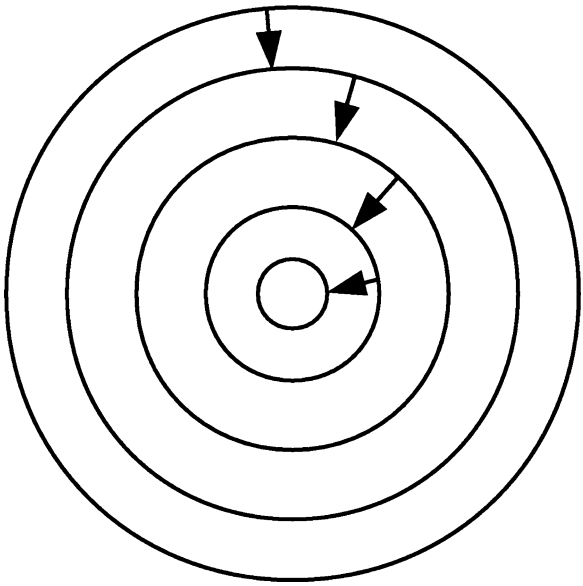


Fig. 6.8 A sequence of quantum transitions representing the collapse of the atom

So that the length is given by

$$\begin{aligned}\text{Photon length} &= \text{Number of oscillations} \times \text{Wavelength} \\ &= 10^7 \cdot 6.0 \times 10^{-7} = 6\,\text{m}.\end{aligned}$$

Note

The spatial extent of a photon is connected with the time it takes for the atom to emit the photon, which is typically on the order of 10^{-8} s: Thus, since the period is the inverse of the frequency,

$$\begin{aligned}\text{Photon emission time} &= \text{Number of oscillations} \times \text{Period} \\ &= 10^7 \times \frac{1}{f} = \frac{10^7}{5 \times 10^{14}} = 2 \times 10^{-8}\,\text{s}.\end{aligned}$$

3. The spiraling of an electron into the nucleus according to Classical Theory is replaced by consecutive **spontaneous transitions** of the atom from one orbit to another orbit of lower energy, as shown in Fig. 6.8. The process is usually referred to as **spontaneous emission**, in contrast to stimulated emission, to be discussed later in this chapter.

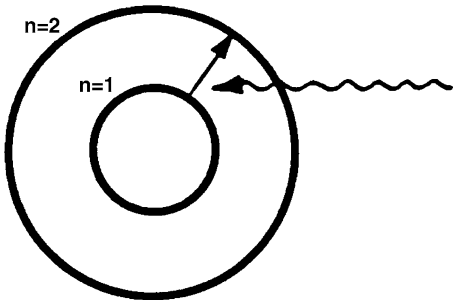


Fig. 6.9 Absorption of a photon by an atom

4. The stability of the atom against total collapse is provided by the existence of the orbit having the lowest energy, the $n = 1$ orbit. This orbit is referred to as the **ground state**.
5. It is observed that an atom will absorb monochromatic EM radiation only if its frequency is equal to one of the frequencies in the *emission spectrum*. That is, the *absorption spectrum* is the same as the emission spectrum. According to Classical Theory, **absorption** of EM radiation is understood to result from the force of the electric field of EM radiation that is incident upon the electrons of an atom and thus transfers its energy to the electrons. Preferential absorption of certain frequencies is due to resonance (see Sect. 2.16): The incoming EM wave has a frequency equal to one of the modes of vibration of the collection of interacting charges in the atom.

   According to the Bohr Theory, *absorption* occurs when a photon is incident upon an atom and has an energy equal to the difference between the energy of the initial orbit and the energy of an orbit of higher energy. See Fig. 6.9.

   According to the Principle of Conservation of Energy, we must have

$$E_{\text{ph}} \equiv hf = E_{\text{f}} - E_{\text{i}} \tag{6.12}$$

   which is also known as the photon absorption condition.

   Thus, for absorption of a photon of frequency f to take place,

$$f = \frac{E_{\text{f}} - E_{\text{i}}}{h}. \tag{6.13}$$

Sample Problem 6-3

What must the frequency of a photon be for it to be absorbed by a hydrogen atom that starts out in the $n = 2$ orbit and is to make a transition to its $n = 3$ orbit?

Solution

$$E_{\text{ph}} = E_3 - E_2 = (-1.5) - (-3.4) = 1.9\,\text{eV}$$

$$f = \frac{E_{\text{ph}}}{h} = \frac{1.9}{4.1 \times 10^{-15}} = 4.6 \times 10^{14}\,\text{Hz}.$$

> **Note**
>
> An atom in the $n = 3$ orbit can make a spontaneous transition to the $n = 2$ orbit and simultaneously emit a photon of the same energy and frequency. It is clear from 6.10 and 6.13 why, in the framework of the Bohr Theory, the absorption spectrum is the same as the emission spectrum.

6.4 Quantum Theory

The Bohr Theory accounted for the atomic spectrum of hydrogen very well. But it failed to quantitatively account for the spectrum of any other atom except those having only one electron, such as a Helium atom that has lost one of its two electrons. Ultimately the Bohr Theory was supplemented by a more comprehensive theory called **quantum theory**.

Recall that according to the Bohr Theory, an electron orbits a nucleus as do the planets about the Sun. However, this picture is not substantiated by experiments. Then what is the path of an electron? We will be describing the behavior of electrons in an atom that defies our understanding of the way particles should behave. Quantum Theory is precise in accounting for our observations. Most significantly, it predicts that

No experiments can allow us to describe how electrons move about.[4]

Instead, we account for experimental observations in terms of the atom's being in one of a set of so-called quantum states, whose significance will be elucidated below:

Suppose an atom is known to be in a certain quantum state. Rather than knowing the precise orbit of the electrons, quantum theory provides us with the **relative**

[4]This statement is likely to strike the uninitiated reader as being preposterous since it implies that Physics cannot answer questions that are fundamental to its own ultimate purpose. Originally, Physics provided us with two components for describing the behavior of the physical world:

1. Accounting for our observations
2. Providing us with a picture as to what systems look like. With respect to the solar system, the Laws of Physics allow us to predict where we will find a planet at any instant of time; and, we can draw pictures and produce cinema simulations of a planet in orbit

According to Quantum Theory, we have to give up the second component. We cannot predict where an electron will be nor can we draw any picture describing its motion. The electron cannot be described in terms of anything we have observed with our eyes. The electron is what it is, unknowable in our terms. There is an interesting comparison in the Torah, where it is written: *Moses said to God, "Suppose I go to the Israelites and say to them, 'The God of your fathers has sent me to you,' and they ask me, 'What is his name?' Then what shall I tell them?" God said to Moses, "I AM WHO I AM". This is what you are to say to the Israelites: 'I AM has sent me to you.'*

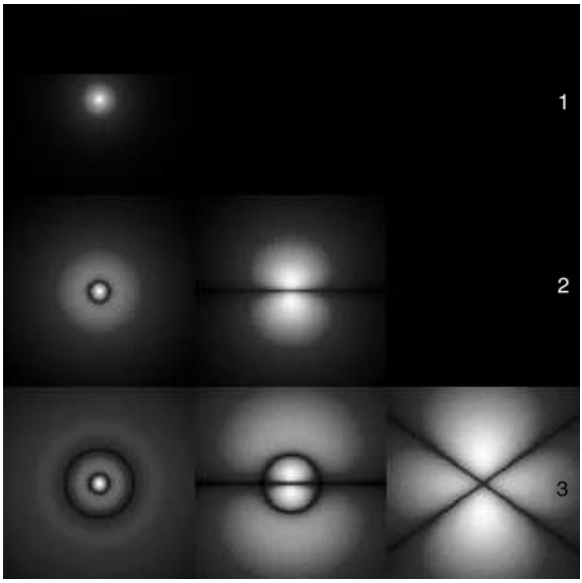


Fig. 6.10 Electron probability clouds (source: http://en.wikipedia.org/wiki/File:HAtomOrbitals.png)

probability for finding an electron at various locations, called the **probability density**. The probability density for some of the quantum states of the hydrogen atom is depicted in Fig. 6.10.[5]

It is very important to note that while a classical mode of vibration has a specific *frequency* of vibration, a quantum probability mode has a specific *energy*. At the simple level of description provided by the Bohr theory, the Quantum Theory of absorption and emission of EM radiation is essentially the same as that of the Bohr Theory except that quantum states are not associated with well-defined orbits of electrons. We have merely to replace the word *orbits* with the term *quantum state* in our text.

The reader might ask how a discipline can discover by reasoning internal unto itself its own limitations? A full discussion of this issue is beyond the scope of this text. My best recommendation is that you view the video of Richard Feynman lecturing on *Probability and Uncertainty, The Quantum Mechanical View of Nature*. You can see the lecture by Feynman on this website (1-12-2011):
http://www.clicker.com/web/richard-feynman-the-messenger-series/Probability-and-Uncertainty:-The-Quantum-Mechanical-View-of-Nature-404149/.

The lecture is contained in the book by Richard Feynman entitled *The Character of Physical Law* [MIT Press, Cambridge, MA, 1967].

[5]The numbers to the right of the figure correspond to the quantum numbers $n = 1$, 2, and 3. You see one probability density for $n = 1$, two for $n = 2$, and three for $n = 3$. Absent are two others for $n = 2$ and five others for $n = 3$. Thus, in the place of Bohr's one state for each n, there is one state for $n = 1$, four for $n = 2$, and nine for $n = 3$.

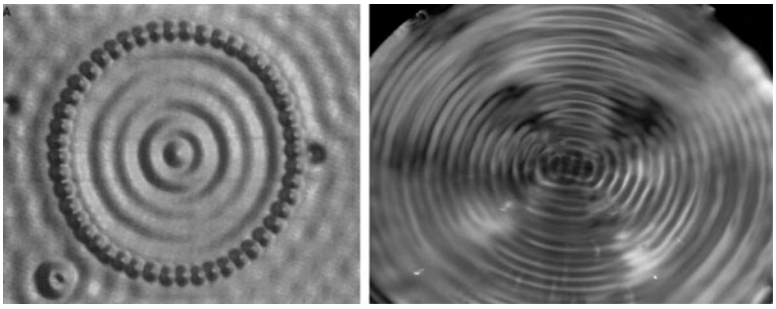


Fig. 6.11 The "Iron Corral": 48 iron atoms on a copper surface, with a standing wave of probability density; Standing surface wave in a mug of water (sources: Corral: From M. F. Crommie, et al, *Science* **262**, 218 (1993); Reprinted with permission from AAAS; photo of water surface by Konstantinos Metallinos)

In Fig. 6.10, the brighter the region, the more likely is it to find an electron in that region. Notice that the patterns of these probability densities bear a remarkable resemblance to the patterns of the modes of vibration of a brass plate. For this reason, we will also refer to these quantum states as **probability modes**.

At the top of the figure is the probability cloud for the ground state. While the Bohr theory describes this state as a classical orbiting electron at the Bohr radius, quantum theory and experiment describe the electron as not having a well-defined orbit. To the incredible contrary, the electron is most likely to be found at the origin. The second and third rows display probably clouds corresponding to $n = 2$ and $n = 3$, respectively.

Currently, it is possible to lay down a number of atoms on a substrate made of other atoms, arranged in a pattern of choice. One such example is shown in Fig. 6.11. We see 48 atoms of iron arranged in a circle – known as a **quantum corral**. Each peak represents the probability distribution of electrons on a single iron atom. An added delight is the presence of concentric crested rings, which represent a standing wave of probability density! Note the similarity with a surface wave of water that is produced by placing a mug of water in a sink and turning on the garbage disposal. If you count the crests, you can determine which circularly symmetric mode is being excited.

Quantum Theory provides us with a mathematical means (that is far beyond the scope of this text to present) of correctly determining the quantum states of all atoms, molecules, and, indeed, macroscopic samples of matter containing huge numbers of atoms, that is, solids, liquids, and gases. Quantum theory can account for the properties of all materials.[6] Ultimately, quantum theory enables us to explain such questions as to why, at room temperatures, copper is solid, is opaque with a

[6]The theory provides us with equations that need to be solved to calculate the material properties. In practice, these equations are so complex that they can be solved only approximately. However, any discrepancy between calculated values and observed values is accountable by the approximation of the calculation and not any shortcomings of the theory.

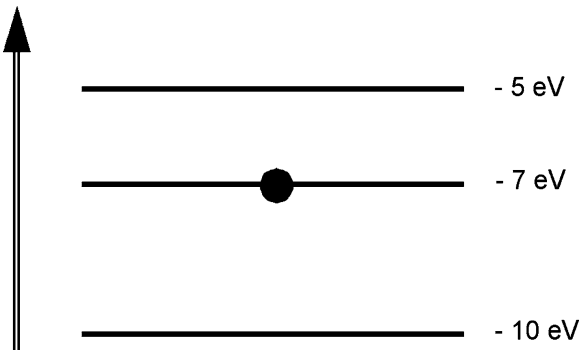


Fig. 6.12 Energy level diagram

shiny orange appearance, is pliable, and conducts electricity well, while water is liquid, is transparent and colorless, and conducts electricity very poorly.

The quantum theory of the emission of EM radiation can be summarized as follows:

1. Any system has a set of Quantum States (Probability Modes), each with a specific energy.
2. **Emission of EM radiation** occurs via a transition from one quantum state to another having a lower energy, accompanied by the emission of a photon. One might regard a quantum state as a **probability mode**.
3. **Absorption of EM radiation** is the reverse process: A photon of frequency f is absorbed by an atom making a transition from one quantum state to another of higher energy.
   For both emission and absorption, the photon energy and hence frequency is accounted for by the change in energy of the system, as in the Bohr Theory of the hydrogen atom. as given by (6.9), (6.10), (6.12), and (6.13).
4. Among all the states of any system, there exists a *state having the lowest energy*. This state is called the **ground state**. The states with higher energy are called **excited states**.
5. Sometimes more than one quantum state has the same energy. We then refer to this set of states as a **quantum energy level**, or simply, **energy level**. (In particular, even the so-called ground state can consist of more than one state.) The set of energies of an atom is referred to as the system's **energy spectrum**, in analogy with the frequency spectrum of a vibrating system.

It is useful to summarize an energy spectrum in the form of an **energy level diagram** (Fig. 6.12). Such a diagram is depicted below for a *fictitious* energy spectrum that has, to simplify our discussion, integer values of energy when expressed in electron volts. The system is in its first excited level, as indicated by the large dot.

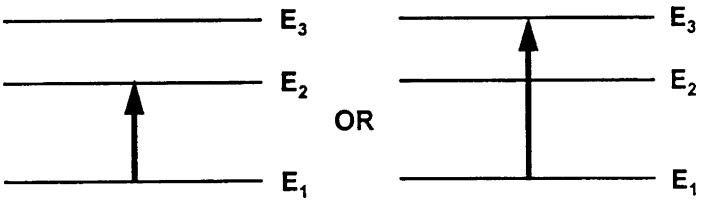


Fig. 6.13 Transitions from the ground state

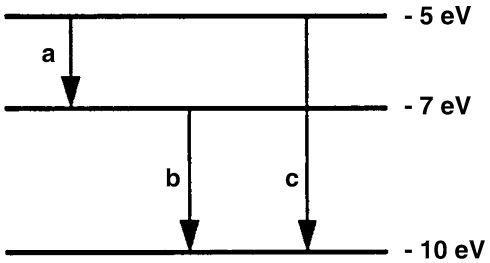


Fig. 6.14 Transitions to lower levels for a three level system

Normally, that is at room temperature, essentially all atoms are in their ground states. An atom can make a transition to an excited state by absorbing energy through a collision with another atom. Transitions from the ground state to one or the other of two excited states are shown in Fig. 6.13. Here, E_1 is the ground state energy.

Once in an excited sate, an atom can spontaneously make a transition to a state having a lower energy. Each such transition is accompanied by the emission of a photon. Thus, referring to the fictitious energy level diagram of Fig. 6.12, with the atom initially in the second excited state, three transitions can take place, with (a) and (b) being consecutive ones (Fig. 6.14).

Photons having three different energies and frequencies can be emitted, corresponding to the three possible quantum transitions (a)–(c), as follows:

1.

$$hf_a = E_3 - E_2 = -5 - (-7) = 2\,\text{eV}$$

$$f_a = \frac{2}{h} = \frac{2}{4.1 \times 10^{-15}} = 5 \times 10^{15}\,\text{Hz}.$$

2.

$$hf_b = E_2 - E_1 = -7 - (-10) = 3\,\text{eV}$$

$$f_b = \frac{3}{h} = \frac{3}{4.1 \times 10^{-15}} = 7 \times 10^{14}\,\text{Hz}.$$

3.

$$hf_c = E_3 - E_1 = -5 - (-10) = 5\,\text{eV}$$

$$f_c = \frac{5}{4.1 \times 10^{-5}} = 1.2 \times 10^{15}\,\text{Hz}.$$

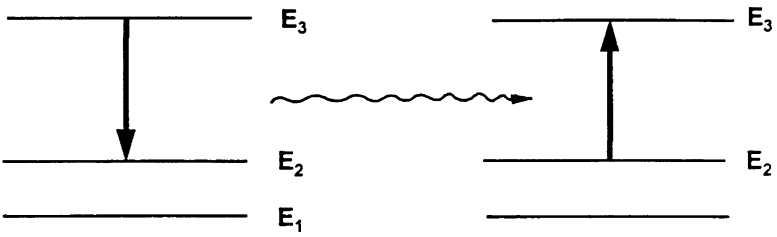


Fig. 6.15 Resonance between two atoms

Photons (a) and (c) are in the ultraviolet regime. Photon (b), being at the boundary between the visible and ultraviolet regimes, is barely visible.

Question: Suppose there are four quantum states. How many transitions are possible from the third excited state? Beware; there are more than four!

Resonance can occur between two identical atoms via the emission of a photon by one atom and the subsequent absorption of that photon by a second atom. This process is depicted on the following pair of energy level diagrams, depicted in Fig. 6.15.

Compare this process with that of the two charged SHOs in Sect. 8.9. In the classical case, resonance requires the equality of the frequency of the respective modes of the two systems. For quantum systems, resonance requires the equality of the difference in energy between a pair of energy levels.

Recall now that we mentioned in our introductory remarks that animal eyes are sensitive to EM radiation because of the behavior of atoms. We need to expand a bit on this remark: The eye is sensitive to a broad band of wavelengths. If individual atoms were the visual sensors in the eye, the absorption spectrum would consist of a small number of spectral lines in the visible regime and the absorption spectrum would be far from continuous. Instead, molecules are used that have a band of a great many excited states, all with energies above the ground state that correspond to visible photons. Absorption with an energy band is quite continuous.[7] We can surmise that evolution produced visual sensors that are sensitive to light from the Sun. As we will see below, the spectrum of sunlight is centered in the visual region of animal eyes.

6.5 Complex Scenarios of Absorption and Emission

We have discussed absorption and emission of photons as two independent processes. We now describe a number of scenarios that are more complex, involving both absorption and emission.

[7] See Sect. 14.12 for more details.

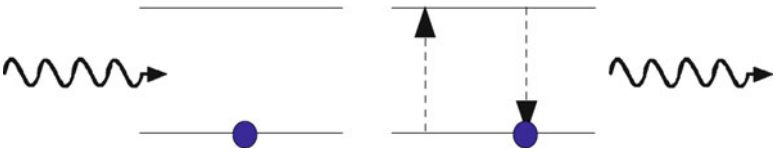


Fig. 6.16 Rayleigh scattering in an energy diagram

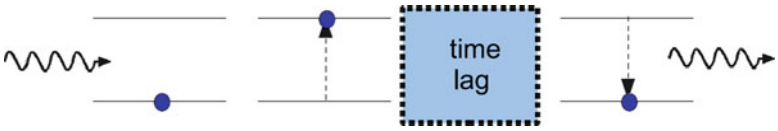


Fig. 6.17 Resonance fluorescence in an energy diagram

6.5.1 Rayleigh Scattering

The simplest obvious complex scenario is the absorption of a photon by an atom (or a general quantum system such as a molecule) followed by the emission of a photon with the same energy. The atom makes a transition from an initial state "i" to an excited state "f" and then returns to the initial state "i". Thus we have "i" to "f" and back to "i". We see this process depicted in Fig. 6.16.

The photon emitted is coherent (has a definite phase relation) with the incident photon. When we have a huge rate of photons incident on many such atoms, the overall process is observed as the scattering of a light beam – hence the insert of the term "scattering" in the name **Rayleigh scattering**. Since the outgoing photon has the same energy as the incoming photon, the scattering is regarded as being elastic. It was **Lloyd Rayleigh** who identified this type of scattering of sunlight as being responsible for our blue skylight – at the time when quantum theory was not yet formulated.

6.5.2 Resonance Fluorescence

Another possibility is that the atom remains in the excited state for an unpredictable time – the atom is described as being in the excited state. Some time later, the atom makes a transition back to the initial state. This phenomenon is referred to as **resonance fluorescence** (Fig. 6.17).

6.5.3 General Fluorescence

The last scenario we discuss here is **general fluorescence** . In this case, we have absorption of a photon by a molecule from an initial state to an excited state.

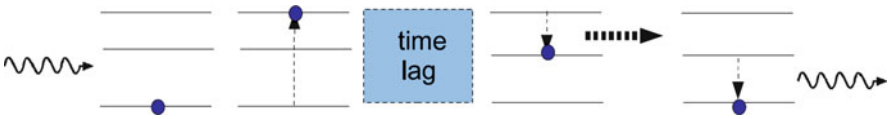


Fig. 6.18 General fluorescence in an energy diagram

After an unpredictable time, the molecule gives up its internal energy to the environment and makes a transition to a state of lower energy. Concurrently, or shortly thereafter, the molecule makes a transition to a state of lower energy along with emission of a photon. See Fig. 6.18. The thick dashed arrow represents the process whereby a part of the excitation energy is replaced by forms of energy such as vibrational or thermal energy. If the time for emission of a photon is relatively long – as long as hours – the phenomenon is referred to as **phosphorescence**. This process occurs when the intermediate state from which the final transition to the ground state occurs has a long lifetime.[8] A familiar example is the result of shining ultraviolet radiation on a rock and observing visible radiation (light) afterwards.

The energy given up to the environment can simply be an increase in thermal (heat) energy. In the case of the cones of the retina of an eye, some or all of the energy given up goes into the nerve impulse that travels down the optic nerve. See Chap. 12 on The Eye.

6.5.4 Stimulated Emission

Consider an atom that is in an excited state. We know that it can make a transition to a state of lower energy along with the emission of a photon having an energy given by the difference of the energies of the two states. We mentioned that the process is called **spontaneous emission**. Now imagine that initially the atom is in the above excited state and that a photon of the above energy is incident on the atom. The incident photon can then induce the transition to the lower state, with the emission of a photon. We will then have two photons of the same energy. See Fig. 6.19 in which we see depicted an incident photon of energy hf equal to the energy difference ΔE between two atomic states.

At first you might think that nothing has been gained. We could view the process as a simple combination of spontaneous emission that occurs coincidently with the passage of a photon of the same energy past the atom. The actuality is that the two photons are now correlated: they are in phase and therefore said to be **coherent**.

Stimulated emission is at the heart of the **Laser**. A beam of photons is created by excitation via electric discharge and subsequent de-excitation with photon emission.

[8]Note that typically the average time for a atomic transition is on the order of 10 nanoseconds (10^{-8} s).

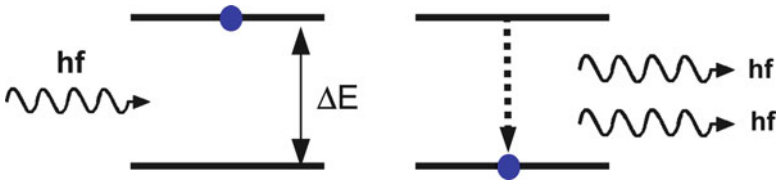


Fig. 6.19 Stimulated emission

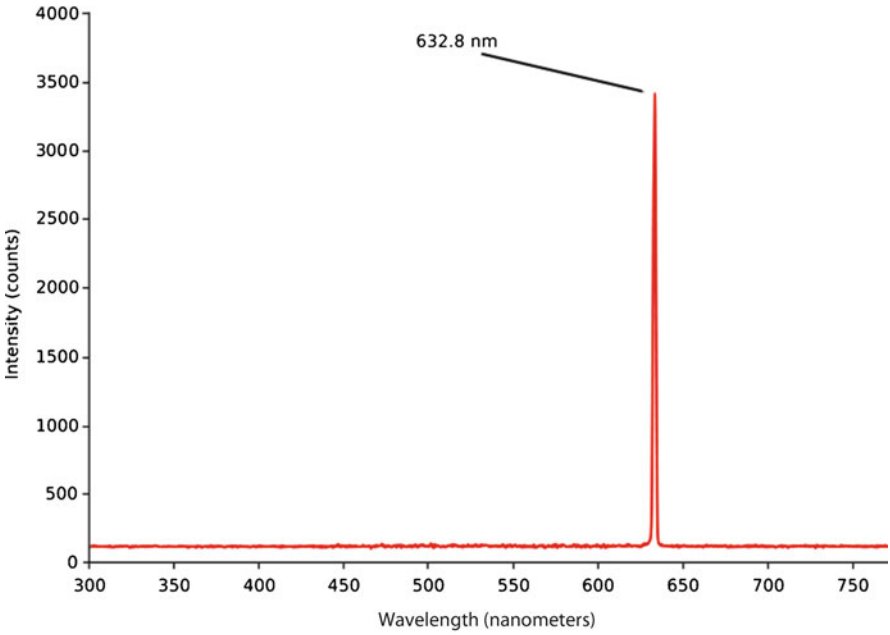


Fig. 6.20 Spectrum of a Helium–Neon laser (source: http://wapedia.mobi/en/Laser)

The photons travel down the length of the laser and contribute to photon production via stimulated emission from excited atoms that have not yet been de-excited spontaneously. Many round trips produce the clusters of coherent photons.

Lasers have very special characteristics. First, the spectrum of the laser beam is extremely sharply peaked, as we see in Fig. 6.20 for a Helium–Neon laser. Second, the laser beam consists of many clusters of photons in each of which all photons are in phase (coherent) with respect to each other. Photons of different clusters aren't in phase. We say that the laser beam is highly coherent.[9] On the other hand, light from

[9]See the website http://hyperphysics.phy-astr.gsu.edu/hbase/optmod/qualig.html for more details about lasers.

most common light sources, such as an incandescent bulb or a fluorescent bulb, is quite **incoherent**. The phases of the photons are randomly distributed because they are emitted in an uncorrelated way. For example, fluorescent light is a result of a huge number **independent** atomic transitions from many atoms.

6.6 Is Light a Stream of Photons or a Wave?

Until we discussed the Bohr Theory of the atom, we treated light as a wave – which is a continuous disturbance. However, in the context of the Bohr Theory or Quantum Theory of absorption or emission of radiation by an atom, we stated that the radiation can only be absorbed in discrete, indivisible "quanta" of radiation called **photons**. These are the so-called particles of light. Which picture is correct? How can they both be correct?

Controversy as to whether light was a beam of discrete particles or a wave was significant in Newton's day ∼300 years ago. Newton himself believed in the particle theory. His contemporary, Christian Huyghens professed the wave theory of light and is responsible for much of what we know about wave propagation today.

The argument regarding the nature of light propagation seemed close to being settled in the 1830s, when Young performed the first wave "interference" experiments[10] which clearly demonstrated the wave nature of light. Further support for the wave theory came from the demonstration that the speed of light in an "optically dense medium" (e.g., glass) is smaller than that in vacuum: The particle theory predicted the contrary. Finally, Maxwell had produced a theory of light as an electromagnetic wave and Hertz confirmed some of its predictions. It became accepted that light propagated as a wave and not as a stream of particles.

Unfortunately the controversy was not yet settled. In the last decades of the nineteenth century, results of studies of the spectrum of EM radiation emitted by ovens heated to high temperatures – so-called Black Body radiation experiments – disagreed with the predictions of classical laws. This led Max Planck, around 1901, to propose that EM radiation is emitted and absorbed in discrete multiples of hf – first called **quanta** and then later called **photons**. Nevertheless, Planck believed that EM itself was not quantized. Rather, it was quantized only in the process of emission or absorption. Einstein emphasized that since all means of detecting radiation involve absorption, and the absorption process is **quantized**, we might as well regard the radiation itself as being quantized, that is, consisting of a beam of quanta. We have already seen how, at the beginning of the twentieth century, the quantum theory of the atom and EM radiation invoked particle attributes to light.

Thus, we see that some experiments indicate that EM radiation propagates as a wave while others indicate that light is propagated as a stream of particles. Which is

[10] See Chap. 7 for details.

the correct description, one might understandably ask? At the level of this text, the question will, unfortunately, have to remain unanswered. All we can say for now is that light, or EM radiation, exhibits *both* wave-like properties and particle-like properties. No experiment has ever, nor, most physicists believe, will ever, be able to reveal the "true" nature of EM radiation in terms of common, familiar concepts. The bewildering wave-particle nature of light is akin to that of the electron.

6.7 The Connection Between Temperature and Frequency

We are now prepared to explain why it is that the frequency spectrum of the Sun is concentrated in the visible region. We mentioned above the role that the study of blackbody radiation played in the development of the quantum theory of radiation. In Fig. 6.21, we see the spectrum of blackbody radiation graphed against wavelength (*not* frequency) for various absolute temperatures ($T = {}^\circ\text{C} + 273$). (The graph

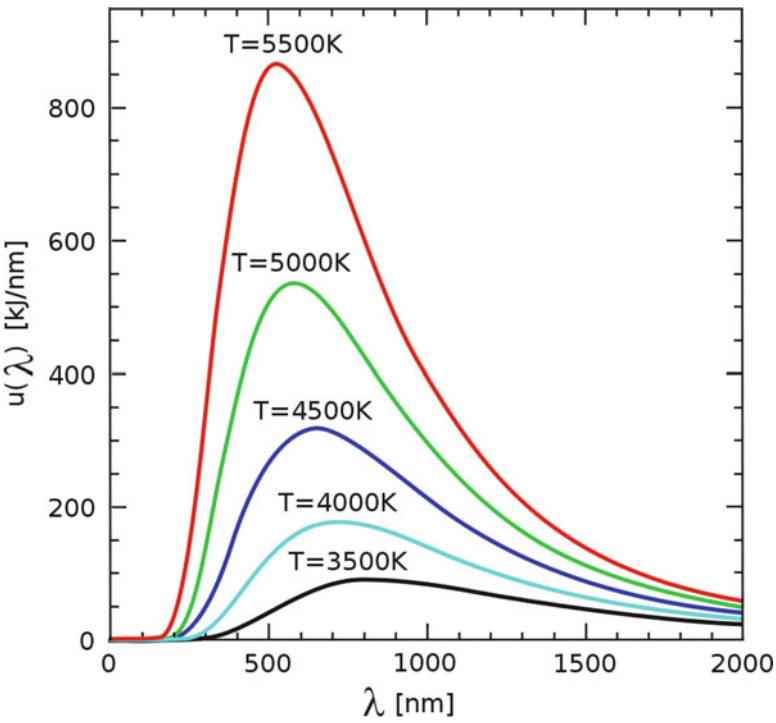


Fig. 6.21 Blackbody frequency spectrum (source: http://en.wikipedia.org/wiki/File:Wiens_law.svg)

refers to the **incoming flux**, which is proportional to the intensity.) Notice how the curves change as the temperature increases.[11]

First, we note that the *height* of the peaks increases with increasing temperature. This is so because it is a general property of matter that energy increases with increasing temperature. In the case of black body radiation in particular, the average energy of the photons is proportional to the absolute temperature T. According to the Planck equation, frequency is proportional to the energy $-E = hf$. As a consequence, the average frequency of the photons increases with increasing temperature.

Now let us turn to the position of the peak. Note that it moves toward lower wavelengths as the temperature increases. Here is the explanation: As the temperature increases, the frequency at the peak increases. The wavelength at the peak decreases correspondingly because of its inverse relation with the frequency: $f = c/\lambda$.

Correspondingly, the frequency associated with the peak increases with increasing temperature, being proportional to the temperature T.[12]

The decreasing peak wavelength with increasing temperature is associated with the changing color of materials that are heated up. For example, as a flame gets hotter, its color changes from red (at the low frequency end of the visible spectrum) to white (a broad spectrum centered around the middle of the visible spectrum) and finally to blue (at the high frequency end of the visible spectrum). (See Chaps. 13 and 14 for details regarding the connection between color and the frequency spectrum.) Notice that the spectrum of the Sun is very close to that of blackbody radiation at a temperature of 6,000 K, which is the surface temperature of the Sun. This connection between color and frequency and temperature enables astronomers to determine the temperature of distant matter in outer space.

One of the most outstanding applications of our knowledge of blackbody radiation is in the connection with the 3 K blackbody radiation of the **Big Bang**. Its discoverers, Penzias and Wilson, had detected unexpected radiation that they attributed to *noise* (random radiation) from the Earth. Ultimately, this radiation was shown to be associated with radiation produced as a result of the Big Bang. The first step in this identification was their fitting the spectrum of this radiation to a blackbody curve as shown above and fitting the temperature to 3 K. Coincidentally, cosmologists, who study the origin and evolution of the Universe, had already predicted the existence of such radiation with an estimated temperature of 3 K. This result gives support to the validity of Einstein's theory of Gravitation – **General**

[11]Regarding Fig. 6.21 you can see the Black Body curve for any chosen temperature as well as the corresponding color of the spectrum using the following physlet: (12-29-2010): http://ephysics.physics.ucla.edu/physlets/eblackbody.htm.

[12]The interested reader can carefully check from the curves that the wavelength at the peaks is inversely proportional to the absolute temperature.

Relativity. Recent penultimately precise measurements of the *variation of the temperature with direction in the sky* have led to further significant cosmological information.

Our study of the world of atoms and light has enabled us to see how physics weaves an intricate web of concepts and establishes quantitative relationships among various physical parameters, such as frequency, wavelength, temperature, energy, pressure, displacement, and time. We can appreciate why physicists feel that they are obtaining a knowledge of the ultimate truths about the Universe.

6.8 Terms

- Absolute temperature (Example: 100 K=100 Kelvins)
- Absolute zero temperature
- Atomic spectra
- Black body radiation
- Bohr radius – a_0
- Bohr Theory of the hydrogen atom
- Classical physics vs. modern physics
- Diffraction grating
- Discrete spectrum
- Empirical formula
- Energy level
- Energy level diagram
- Ground state
- Line spectrum
- Photon
- Photon absorption
- Photon emission
- Planck relation
- Planck's constant
- Probability density
- Probability mode
- Quantized process
- Quantum state
- Quantum theory
- Spectral analysis
- Wave packet

6.9 Important Equations

Planck relation:

$$E_{\text{ph}} = hf. \tag{6.14}$$

Photon emission:

$$f = \frac{E_{\text{f}} - E_{\text{i}}}{\hbar}. \tag{6.15}$$

Photon absorption:

$$f = \frac{E_{\text{i}} - E_{\text{f}}}{\hbar}. \tag{6.16}$$

6.10 Problems for Chap. 6

1. What are the frequencies and wavelengths of all photons that would be ultimately emitted by an atom that has energy levels with the following energies: 3, 2, and 0 eV?
2. What key feature of Balmer's formula was used by Bohr in his theory of the hydrogen atom? Explain.
3. What does the term "empirical" mean? Why was Balmer's formula merely empirical?)
4. What catastrophe plagued the Rutherford model within the context of classical theory? How does the Bohr Theory deal with this catastrophe?
5. Determine the frequency and wavelength of a photon having an energy of 1 J; of 1 keV.
6. Determine the energy of a microwave photon of frequency 2,500 MHz.
7. Read problem (8) in Chap. 4 that deals with the **stroboscope**.
   Estimate the number of photons emitted by the stroboscope in a single flash.
8. What is the length of a photon whose frequency is 7×10^{14} Hz if it consists of 100 oscillations?
9. Describe resonance between two atoms according to quantum theory.
10. Given that the wavelength at the peak of the black body intensity curve (Fig. 6.21) is 4,300 $\mathring{A}$ at a temperature of 7,000 K, calculate the corresponding wavelength at a temperature of 5,000 K using the relation: peak photon frequency $\propto$ absolute temperature. Compare your result with that taken from Fig. 6.21.
11. Suppose that a 100 Watt yellow light bulb has an efficiency of 3% for producing visible light. (That is, only 3% of the input electric power is converted into light). An observer looks directly at the bulb at a distance of 10 m from the bulb.

    (a) Assuming that the light is isotropic, what is the intensity of the light incident on the eye?
    (b) Estimate the area of the pupil of an average human eye.
    (c) Neglecting reflection and absorption by the pupil of the eye, determine the power of the light entering the eye through the pupil.
    (d) How much visible light energy enters the eye in one second?
    (e) For simplicity, suppose that all the visible light entering the eye has a wavelength of 6,000 $\mathring{A}$. What is the frequency of the EM radiation?
    (f) How much energy does a single photon of this radiation have?
    (g) How many photons enter the eye in one second?

12. Lawrence Livermore National Laboratories reported making a laser that could produce a laser pulse that had a peak power of about one petawatt ($=10^{15}$ watts) and that lasted for about one-half picosecond ($= 0.5 \times 10^{-12}$ s). (See the website 12-26-2010: https://www.llnl.gov/str/Petawatt.html).

    (a) Calculate the total energy of one pulse.

(b) Given an intensity of 700 Giga (10^9) Watts/cm^2 during the pulse, calculate the area of the laser beam.
(c) How long would a 100-W light bulb have to burn to produce the same energy? Note how the laser concentrates in many ways: frequency with respect to spectral intensity, intensity with respect to area, phase with respect to coherence and here, energy with respect to time.
(d) If the wavelength of the laser light were 530 nm, how many photons are there in a single pulse?

Chapter 7
The Principle of Superposition

Suppose there are two sources of waves of a given type. For example, there may be two loudspeakers emitting sound waves or two accelerating charges emitting EM waves. What is the resultant wave? Or, suppose that two pulses are sent down a string, one after the other, so that the second pulse "collides" with the first one after the first one has been reflected from the opposite end. What happens as a result of this collision? Such questions are answered by the **principle of superposition**, which states that:

The wave that results from two independent sources – the so-called **resultant wave** – is a simple sum of the two waves that would in turn be produced by the respective sources if each were present alone.

This principle will be illustrated by numerous applications in this chapter.

7.1 The Wave Produced by Colliding Pulses

In Fig. 7.1, we depict the wave pattern at five different times that results from a "collision" of two pulses that are traveling in opposite directions along a string. The solid curve represents the observed wave. Initially we see the two pulses far apart. In the next figure, we see the two pulses having just begun to overlap – we might say they are "colliding." The two individual pulses are shown as dashed curves, as they would appear if they were present alone, each without the other. What is actually observed is the solid curve, which is a sum of the two dashed curves. Notice how, in the third curve of Fig. 7.1, the displacement vanishes everywhere along the string. There is no potential energy at this instant. What has happened to the energy that was in the pulses? All of the energy of the wave resides in kinetic energy. Finally, we see how the two pulses survive the collision intact, moving away from each other, as if the collision had not taken place.

Below, in Fig. 7.2, is a second example of two colliding pulses – here square pulses. They have the same amplitude and move past each other at the same speed of 1 unit/s. We see the resulting wave at four different times. At $t = 1$ s, they are

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_7,

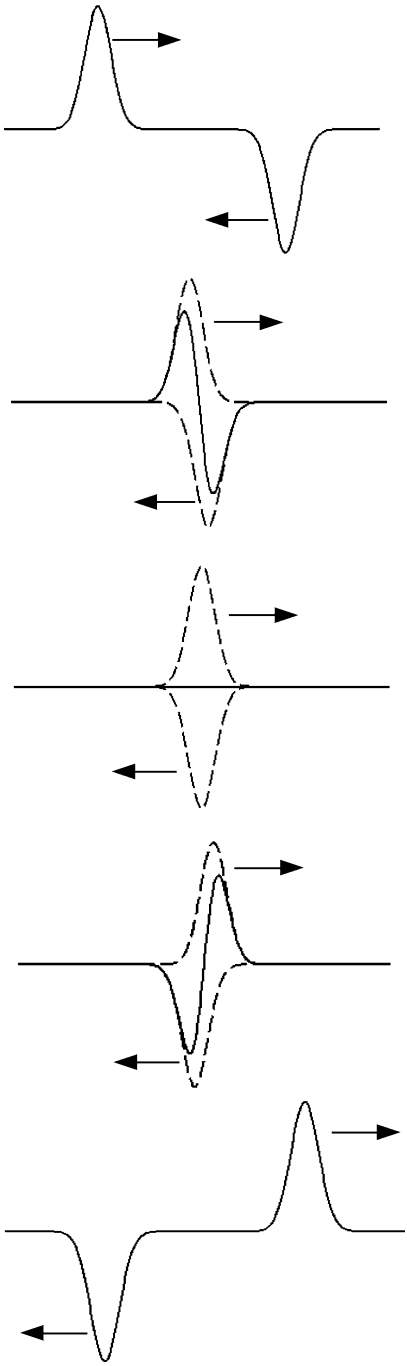


Fig. 7.1 Collision of two pulses

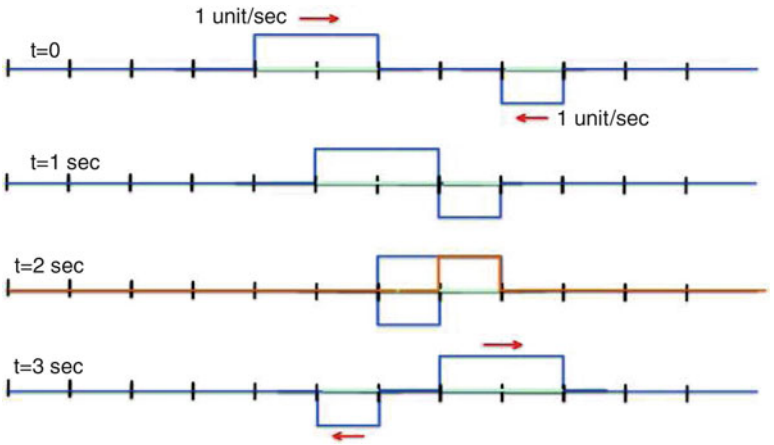


Fig. 7.2 Collision of two square pulses

about to "collide." At $t = 2$ s, we see how each of the pulses would look if each was travelling alone. However, the actual wave is obtained by adding the two graphs. There is complete cancellation in the middle segment. The resultant is shown as the orange graph. At $t = 3$ s and thereafter, the two pulses are seen retreating from each other as if neither were affected by the other.

7.2 Superposition of Two Sine Waves of the Same Frequency

By itself, a sine wave is characterized by its amplitude and either its period in time or its wavelength in space. However, when two sine waves are to be added, the relative position of their peaks matters too. Recall from Chap. 2 that this characteristic is called the **phase difference**, or equivalently, the **relative phase**.

In Fig. 7.3, we see three sine waves having the same period in time (1 s) and the same amplitude, but a non-zero relative phase with respect to each other. We simply say that they are **out of phase** or *have different phases*.

Clearly, since the crests of wave A coincide with the troughs of wave B, the waves will cancel each other when added. Thus $A + B = 0$.

Generally, the relative positions of the waves can be expressed in fractions of a cycle or as an angle ranging from 0° to 360°, with 360° representing a full cycle.

Thus, relative to wave A:

- Wave B is $1/2$ cycle or 180° ahead or behind A.
- Wave C is $1/4$ cycle or 90° ahead of A.

NOTE: Two waves with no phase difference are said to be **in phase**.

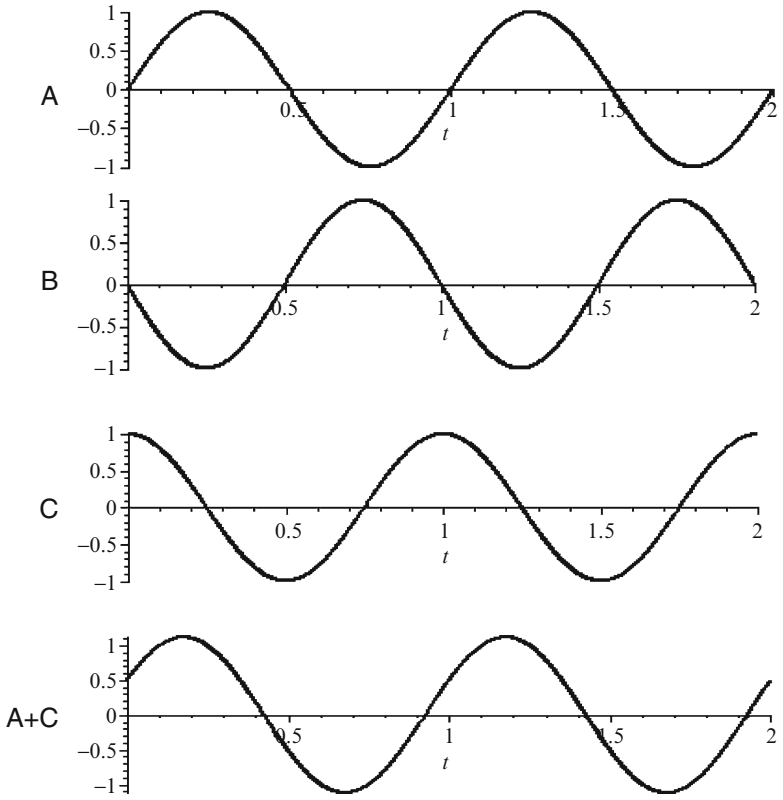


Fig. 7.3 Superposition of two sine waves

Note

Sine waves have the following amazing property that we mentioned in Chap. 2 and that we will review in greater detail here:

Two sine waves having the same period add up to a sine wave of the same period.

The amplitude of the resulting sine wave depends upon the amplitudes of the component sine waves and their relative phase.

Thus, we see in Fig. 7.3 that the sum of waves A and C is a sine wave with the same period as A and C.

Note

Above, we have described the superposition of two sine waves as they vary in time. The behavior is purely mathematical. Thus, if we have two sinusoidal patterns in space, we can apply the above results by replacing the word *period* by the *wavelength*.

Let A_1 and A_2 be the amplitudes of the respective components and A the amplitude of the sum or **resultant** amplitude. If the components are *in phase*, the resultant amplitude is given by

$$A = A_1 + A_2. \tag{7.1}$$

Note that if the two amplitudes are equal, the resultant amplitude is double the individual amplitude.

If the components are 180° *out of phase*, the resultant amplitude is given by[1]

$$A = |A_1 - A_2|. \tag{7.3}$$

Thus, referring to the figure on the previous page, waves A and C are out of phase and have the same amplitude. If added, they would therefore cancel and we would be left with no wave at all![2]

7.3 Two Source Interference in Space

Consider the resultant traveling wave from two **point sources**[3], S_1 and S_2 respectively, which emit sine waves of the same frequency (and hence wavelength), with amplitudes A_1 and A_2, respectively. We will assume here that the waves start out from S_1 and S_2 in phase. In Fig. 7.4, we have captured an instant when the two waves have a crest at their respective sources. The two sets of concentric circles represent the crests of the two component waves.

The key to determining the resultant wave at any particular position P in space lies in the *difference between the respective distances from the point to the two sources* – which we will refer to as the **path difference**.

In the figure, where two circles cross, the path difference is a multiple of the wavelength. At such points, the component waves arrive in phase. Then we have

[1] Generally, the amplitude is given by

$$A = \sqrt{A_1^2 + A_2^2 + 2A_1A_2\cos(\phi_1 - \phi_2)}. \tag{7.2}$$

Note that the phase difference is $\phi_1 - \phi_2$. You can plug into this expression the two phase differences, $\phi_1 - \phi_2 = 0$ or 180°, to reproduce the special cases – in phase or out of phase.

[2] For a general phase difference ϕ (which ranges from 0° to 360°), and when the two source amplitudes are equal, the total amplitude is given by

$$A = 2\cos^2\frac{\phi}{2}.$$

[3] A real source takes up space. A point source is a term used for a source that is so localized that we can regard it as taking up no space.

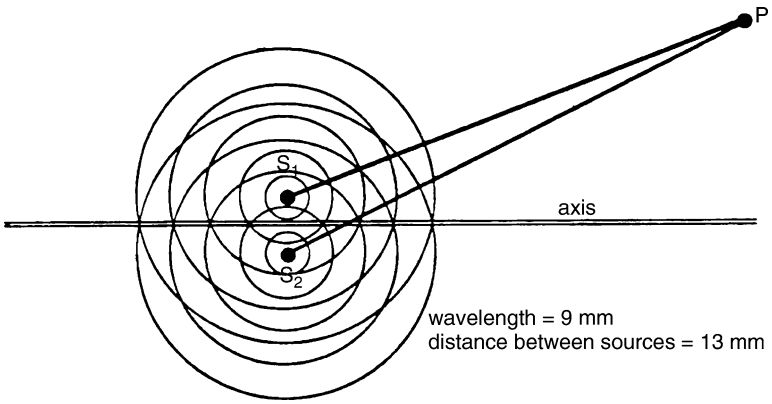


Fig. 7.4 Two point sources

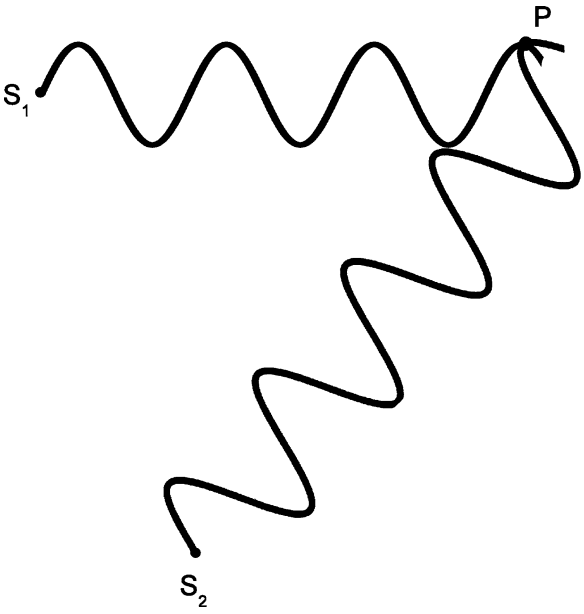


Fig. 7.5 Point where waves arrive in phase

what is called **constructive interference**. The path difference satisfies the equation, with an infinite sequence

$$\overline{S_2P} - \overline{S_1P} = 0, \lambda, -\lambda, 2\lambda, -2\lambda, \ldots. \tag{7.4}$$

which is condition for constructive interference. Here, $\overline{SP}$ refers to the distance between two points, S and P.

Such a point P is shown in Fig. 7.5. We see two sine waves leaving the sources in phase. The distances to the point P are $\overline{S_1P} = 5 + 1/4$ wavelengths and $\overline{S_2P} = 6 + 1/4$ wavelengths. (They leave the source with a value of zero and end up at a peak.)

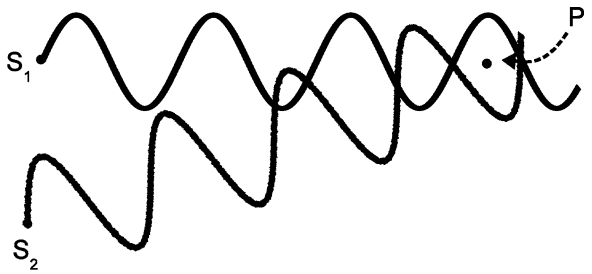


Fig. 7.6 Point where waves arrive out of phase

Since the components are sinusoidal in time, their resultant sum is sinusoidal. The resultant amplitude is given by $A = A_1 + A_2$ (7.1).

Now consider the opposite extreme, wherein the path difference is a multiple of λ plus an additional half wavelength (Fig. 7.6). (The path difference is said to be an **odd-half-integer** number of wavelengths.) The two sine waves, which started out in phase, arrive out of phase at point P. We have **destructive interference** and the amplitude at point P is given by $A = A_1 - A_2$, as seen in (7.3). In particular, if the component amplitudes are equal, the resultant amplitude vanishes.

The condition for destructive interference[4] is

$$\overline{S_2P} - \overline{S_1P} = \frac{\lambda}{2} - \frac{\lambda}{2}, \lambda + \frac{\lambda}{2} = 3\frac{\lambda}{2}, -3\frac{\lambda}{2}, \ldots \tag{7.5}$$

which is condition for destructive interference.

In Fig. 7.7, we depict the curves in space all along which there is constructive interference. These curves are called **hyperbolas** by mathematicians. The two sources are a distance d apart. Far away from the two point sources, the curves approach straight lines, indicated by the dotted lines, at various angles θ with respect to the axis. They are labeled $m = 0, 1, 2, \ldots$ corresponding to the path differences $0, \lambda, -\lambda, 2\lambda, -2\lambda, \ldots$.

The angles satisfy the equations:

$$\begin{aligned} \sin\theta_1 &= \frac{\lambda}{d} \\ \sin\theta_2 &= 2\frac{\lambda}{d} \\ \sin\theta_n &= n\frac{\lambda}{d}, \end{aligned} \tag{7.6}$$

where $\sin\theta$ is the trigonometric sine function.

Each angle specifies an **order of interference**.

[4]When $A_1 = A_2$, and for general path differences, the resultant amplitude is given by $2A\cos^2\phi/2$, where the phase difference $\phi = 2\pi(\overline{S_2P} - \overline{S_1P})/\lambda$. Compare this expression with (7.1).

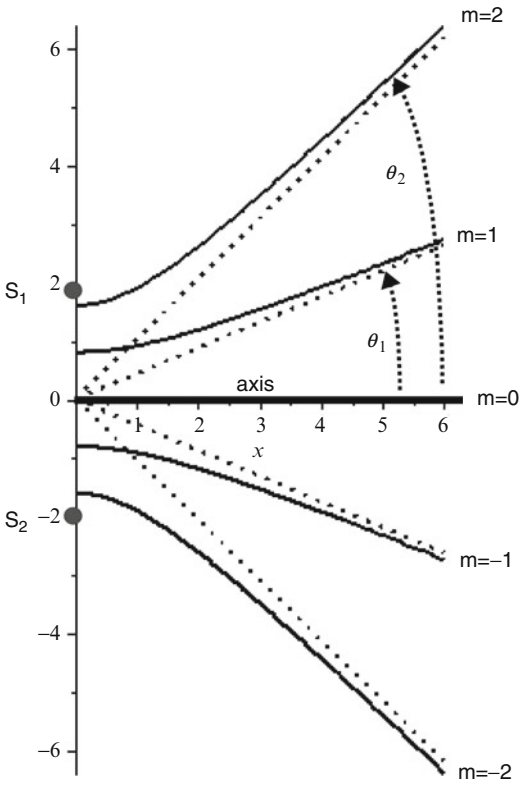


Fig. 7.7 Hyperbola curves along which there is constructive interference

Sample Problem 7-1

Suppose that $\lambda = 6 \times 10^{-7}$ m and $d = 15 \times 10^{-7}$ m. Find all of the angles of constructive interference.This ratio $\lambda/d = 0.4$ of the wavelength to the separation of sources is shown in Fig. 7.7, where d is equal to four units on the y-axis.

Solution

We have

$$\sin\theta_1 = \frac{\lambda}{d} = \frac{6}{15} = 0.4,$$

so that $\theta_1 = 24°$, corresponding to the 1st **order**.

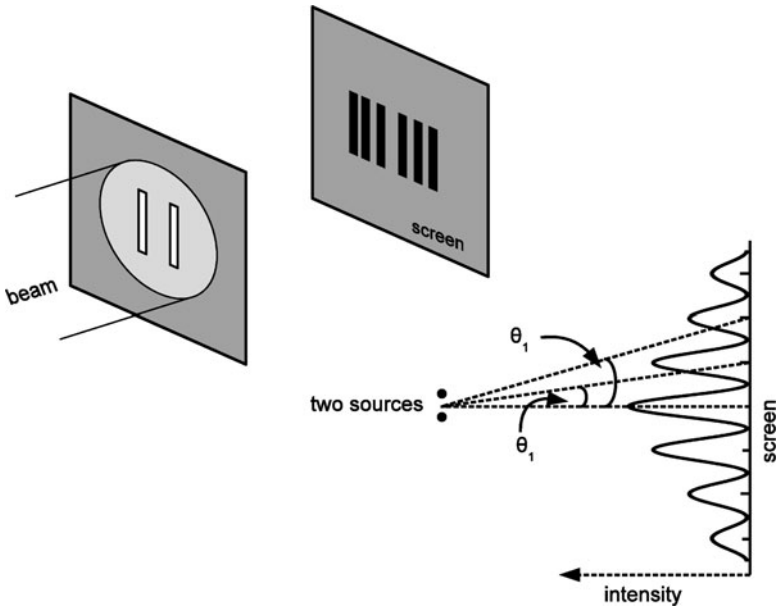


Fig. 7.8 Two slit interference

Also,

$$\sin\theta_2 = 2\frac{\lambda}{d} = 2 \times 0.4 = 0.8,$$

so that $\theta_2 = 53°$, corresponding to the 2nd **order**.
Note that $\theta_2 \neq 2\theta_1$**.**

Next, note that according to our (7.6),

$$\sin\theta_3 = 3\frac{\lambda}{d} = 3 \times 0.4 = 1.2,$$

which is impossible, since the sine of an angle cannot be greater than one. The equation cannot be satisfied, and thus, $m = 2$ is the highest order present.

One might ask how we can produce two sources that start out in phase since two independent sources typically do not have a known and controllable phase relation. Here is one commonly used method: We start with a single source of light, preferably light from a laser. Laser light is an intense beam of monochromatic light that is an ensemble of a many clusters; each cluster has a huge number of photons that are in phase with one another. We say that the light is extremely **coherent**. The beam of light is projected onto an opaque surface that has two parallel slits. These slits produce our two sources starting out in phase with each other, as seen in Fig. 7.8. We then view the light projected on a distant screen.

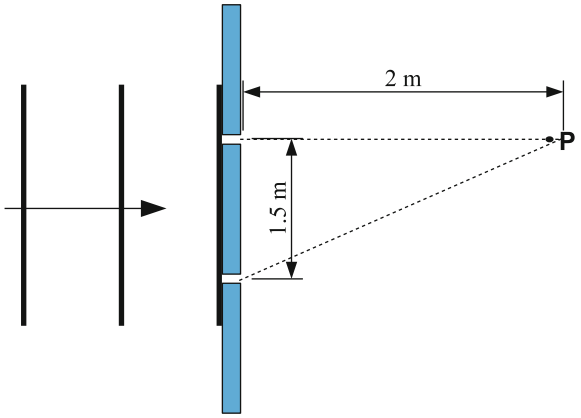


Fig. 7.9 Wave incident upon an opaque plane with two slits

The image on the screen consists of a strip with varying brightness. Analysis predicts that the brightness will vary along the screen as shown in the figure below. We note the set of vertical lines representing bright light. These lines are referred to as **fringes**. According to (7.6), the **closer** the two sources are (the smaller d is), the **greater** the angles $\theta_{1,2,3,...}$ are and the **further** apart will be the fringes on the screen.

Sample Problem 7-2

Suppose that a sound wave of wavelength 1 m is incident on a surface with two holes that are 1.5 m apart, as depicted in Fig. 7.9. You are to determine whether the waves arriving at point P are in phase or completely out of phase. The distance of point P from the surface is 2 m.

Solution

We cannot use a simple formula for the path difference. We must calculate it in association with this particular problem. Thus, the shorter path (from the upper hole) – call it ℓ_1 is simply 1-m. The second path has a distance ℓ_2 equal to the hypotenuse of the right triangle, namely

$$\ell_2 = \sqrt{1.5^2 + 2^2} = 2.5\,\text{m}. \tag{7.7}$$

Since the wavelength is 1-m, the path difference $[\ell_2 - \ell_1]$ is two and one-half wavelengths. Therefore, the waves arrive completely out of phase and we have destructive interference.

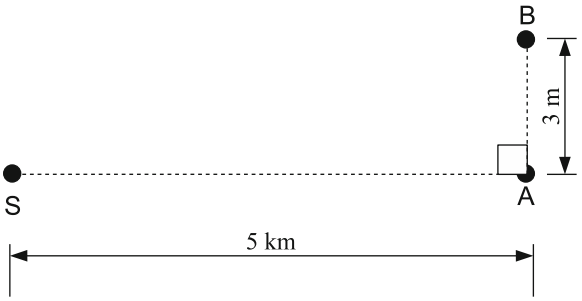


Fig. 7.10 Interference of a radio wave by a neighboring car in traffic

Sample Problem 7-3

Look at Fig. 7.10. It represents the following scenario: Waves from an FM radio station at point S with a carrier wave frequency of 100.1 MHz are received by a radio in a car A that is 5 km away. In addition, a neighboring car B that is a distance of 3 m from car A reflects the radio waves from the radio station toward car A, so that car A's radio receives two interfering waves. The indirect wave travels along the path S to B and then from B to A.

IMPORTANT INFORMATION: When the radio wave reflects off car B, the phase of the wave is flipped by one-half cycle.

Determine whether the waves, the direct wave, and the indirect wave arrive in phase.

Solution

We first calculate the wavelength of the carrier wave. We have

$$\lambda = \frac{c}{f} = \frac{3.00 \times 10^8}{100.1 \times 10^6} = 3.00\,\text{m}. \tag{7.8}$$

Next we calculate the path difference. We have $\ell_1 = 5\,\text{km} = 5{,}000\,\text{m}$. The second path length is the sum of the hypotenuse and the vertical leg of the right triangle (=3 m). The hypotenuse is given by

$$\sqrt{(5{,}000)^2 + 3^2} \sim 5{,}000\,\text{m}. \tag{7.9}$$

Therefore, the second path length is $\ell_2 \sim 5{,}003\,\text{m}$. Finally, the path difference is ~3 m and equal to the wavelength. Without the flipped phase due to reflection, the waves would arrive in phase. With the flip, the waves arrive OUT of PHASE and we have destructive interference.

As the cars move along, there will be evolving relative positions that will result in the two waves having varied phase relations.

7.3.1 Sound Level with Many Sources

In Chap. 4, we saw how the sound level is defined in terms of the intensity. Suppose that the wave has **many sources**. What is the resulting sound level? What matters is the intensity. The intensity depends upon whether the sources are **coherent**. Two waves are said to be **coherent** if they have a definite phase relation. In this case we also say that the sources are coherent.

Consider first two coherent sources that have the same frequency and produce two individual waves that have equal amplitude and are in phase at some location in space; we have constructive interference. In this case, we know that the amplitude is doubled. Consequently, the intensity is **quadrupled**! How can we obtain four times the intensity instead of the doubling we would expect? Are we gaining energy from nowhere, thus violating the principle of conservation of energy? The resolution of this dilemma is that there are other places in space where we have destructive interference. There, the intensity vanishes. Thus, on the average, the intensity is double the intensity of one source.

The result is that the sound level is increased by $10\ \log(4) = 6\,\text{dB}$ at points of constructive interference. If we have n-coherent sources that produce waves that are equal in magnitude and are all in phase, the sound level is increased by $10\log(n^2) = 20\log(n)dB$.

Next, suppose that we have two sources that are **incoherent** and produce the same intensity at some location in space. The result is that the total intensity is doubled, so that the sound level is increased by $10\ \log(2) = 3\,\text{dB}$. With n-incoherent sources, each with the same intensity at some location in space, the intensity is multiplied by a factor of n, so that the sound level is increased by $10\ \log(n)dB$.

As an application, suppose that we have a string orchestra with 100 instruments each producing a sound level of 1 dB at some location. The resulting sound level will be $(1 + 10\ \log(100)) = 21\,\text{dB}$.

7.3.2 Photons and Two-Slit Interference

We learned in Chap. 6 that while light propagates as if it were a continuous wave, in fact the wave characteristics represent the distribution of photons in space. The lack of continuity is reflected by the fact that if we have an ultra-low intensity of light, so that there is an ultra-low density of photons, one can use a photon detector to look for photons. The detector always detects individual photons. It never detects part of a photon. If two detectors are in different locations, only one of the detectors detects a particular photon.

Let us consider a two-slit experiment with a light beam of such low intensity that only one photon at a time can be found between the two slits and the screen. The result is that the photons will impinge on the screen in a random fashion but be distributed according to the interference pattern we observe with a high intensity of light.

We can place two detectors in the apparatus, one behind each of the slits. The result is that only one or the other detector registers that a photon has passed through a slit. It would seem fair to say that a photon exits one slit or the other – never both slits at a time; it does not split into two pieces. The question is how the photons can land on the screen with a pattern of interference.

It is beyond the scope of this book to discuss the answer to this question here. My goal is only to try to entice you, to get you to think and wonder. For a detailed discussion of this experiment, I recommend that you to see the Feynman video referred to in Sect. 6.4.

7.4 Many-Source Interference

We will now discuss the interesting behavior of various ensembles, each of which has many sources that are arranged in a periodic manner in a plane or in three-dimensional space. There are cases where the sources are not point sources. Nevertheless, it can be shown that if each of the multitude of sources is spread out while the sources are identical, the result is an interference pattern that is distributed in space the same way as if the sources were point sources.

7.4.1 Gratings

Suppose now that we have a long line of point sources, all of the same wavelength and equally spaced at a distance d (Fig. 7.11). At any given point in space, the resulting wave in time is a superposition of many sine waves. Again the resultant wave will be sinusoidal in time.

It can be shown that *far away from the sources*, the curves of constructive interference are again straight lines satisfying the same (7.6) as with two sources. The difference is that the regions of brightness are much more sharply defined. Thus, in the case of a light wave, projection on a screen gives a brightness pattern shown in the figure below.

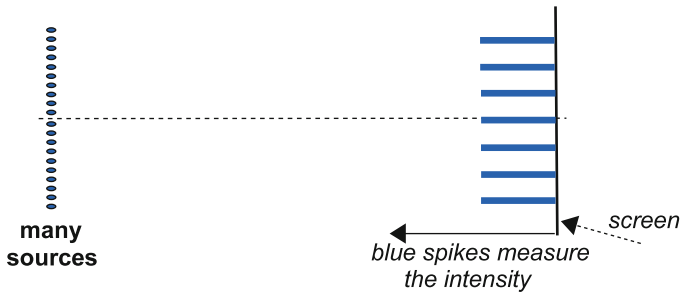


Fig. 7.11 Interference from a linear array of many sources that are equally spaced

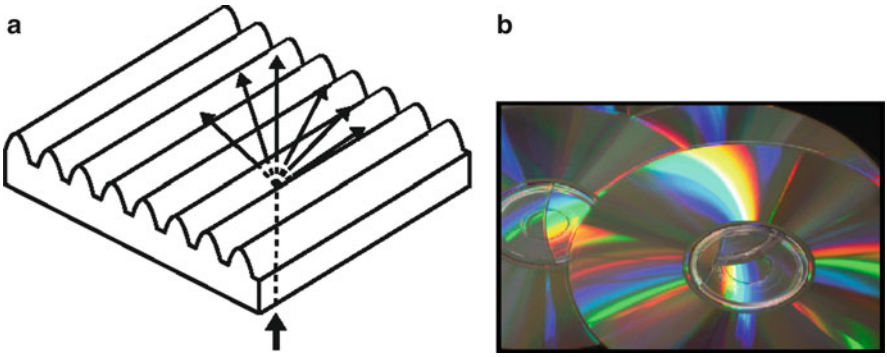


Fig. 7.12 Diffraction grating (photo by Leon Gunther)

We can produce a similar result without using point sources by having a light beam pass through or reflect off of a **diffraction grating**: In the first type, the grating consists of a plate of transparent material, such as glass or plastic, which has a surface upon which identical and equally spaced parallel grooves have been engraved. In the second type, parallel grooves are coated with a reflecting material such as silver. Such is the case with a compact disc (CD), which has closely and equally spaced concentric circles of holes that characterize the digital storage of information, be it audio or data storage. White light that is incident on the surface of the CD is reflected off the grooves, with interference of light depending upon the wavelength – hence leading to the rainbow of colors. Whether the gratings transmits light or reflects light, we essentially have an array of identical, equally spaced sources. See Fig. 7.12.

For a grating, instead of having a periodic array of **point sources**, we have a periodic array of identical **extended sources**: It can be shown that in this case, we obtain the same arrangement of fringes as we would from point sources except that the shape of each fringe is modified according to the wave coming from each groove. The same holds true for the interference pattern discussed below for a two-dimensional mesh and for a crystal.

*7.4.2 Diffraction Through a Mesh***

Let us pass a light beam through **silk screen**, which is a piece of "mesh" material that is a weave of threads arranged in two mutually perpendicular parallel arrays. The result is an interesting interference pattern shown in Fig. 7.13.

In the figure there are two silk screens above, labeled L and R; below them are two diffraction patterns, labeled (a) and (b), that were produced by the two lasers that were set side by side and can be seen through the silkscreens. The number of threads per inch is greater on the right (R); correspondingly, the distance between

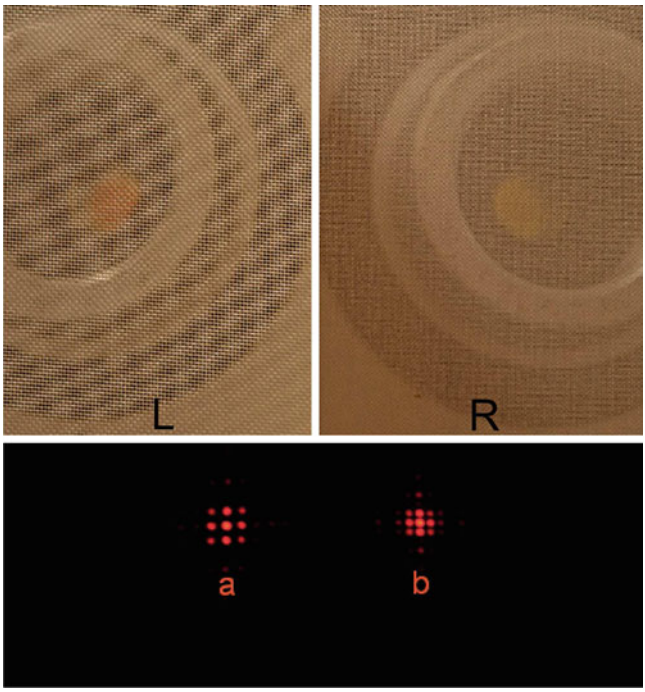


Fig. 7.13 Diffraction through a mesh (photo by Konstantinos Metallinos)

neighboring threads is smaller. One might think that each spot is produced by a single hole in the mesh through which the laser beam passed. This is not the case at all; each spot is produced by light that has passed a number of holes in the silk screen. Without looking at the answer in the footnote, can you determine how the pairs are matched up? L with (a) R with (b) or the converse?[5]

Finally, notice that the number of spots is limited. If the holes were infinitesimal in diameter, the light coming through a single hole would exhibit strong diffraction – See Sect. 8.1 on diffraction. Then the hole would act like a point source. In this case, the spots would go on far the right, left, above, and below. With a finite size hole, the angle of spreading (in radians), due to diffraction of the outgoing wave, is on the order of the wavelength divided by the diameter d of a hole. (See (8.1).) As a result, the number N of spots along a line can be shown to be given by

$$N \approx \frac{\text{Spacing } d \text{ between adjacent holes}}{\text{Diameter of a hole}}. \tag{7.10}$$

[5]Image (a) was produced by the screen labeled R and (b) by the silk screen labeled L. The reason is analogous to the fact that in two source interference, the closer the sources, the further apart are the fringes on a distant screen.

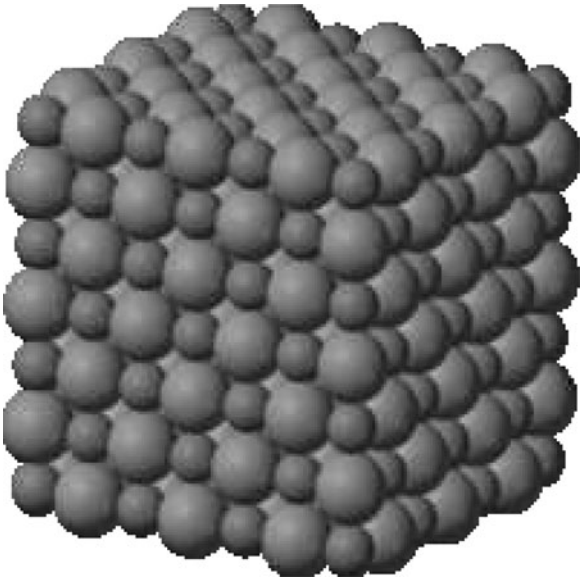


Fig. 7.14 Crystal lattice structure of sodium chloride (source: http://en.wikipedia.org/wiki/Sodium_chloride)

7.4.3 X-ray Diffraction of Crystals

We have just made the transition from a linear array of sources to a two-dimensional array of sources, the mesh material. We now move on to three dimensions (3D): In a **crystal**, atoms or groups of atoms are arranged in a periodic array in 3D space, as depicted by the model of a sodium chloride crystal in Fig. 7.14.

Suppose that we cast a beam of monochromatic EM waves on the crystal. We will get an interference pattern that is determined by the crystal structure. The outgoing wave is strong only in certain directions. One requirement is that the wavelength λ be on the order of or less than or about equal to the spacing d between neighboring atoms (the distance between a large ball and a small ball). This distance is typically on the order of Ångstroms.

In the case of sodium chloride, $d = 5.63/2 = 2.82 \mathring{A}$. We would then need a wavelength $\lambda < 2.82\ \mathring{A} = 2.82 \times 10^{-10}$ m. The corresponding frequency of the EM wave is then $f = c/\lambda = (3 \times 10^8)/(2.8 \times 10^{-10}) = 1.1 \times 10^{18}$ Hz, a frequency in the X-ray region.

The intensity pattern produced on a screen by X-rays scattering off a crystal of sodium chloride is shown in Fig. 7.15.[6]

[6]The figure is the black–white inversion of the original, so as to enhance the appearance of the pattern of spots.

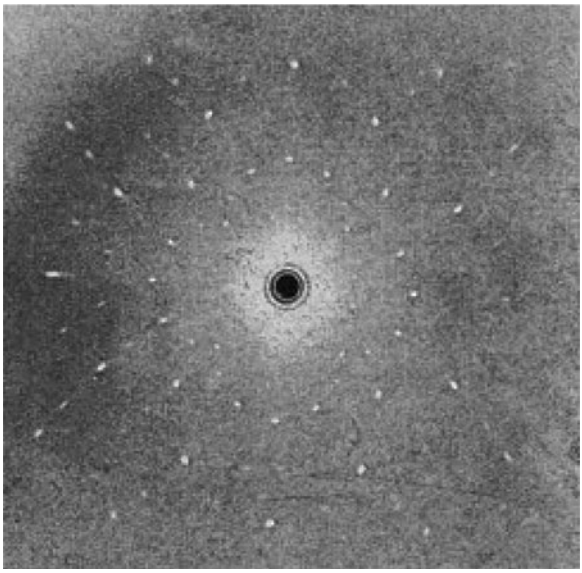


Fig. 7.15 X-ray diffraction pattern of a sodium chloride crystal (source: http://faculty.fullerton.edu/cmcconnell/304/X-Ray_Crystallography.htm)

7.5 Beats

Suppose we superimpose two sine waves having nearly the same frequency. Below we see one sine wave that has ten cycles in a 1 s interval, while a second sine wave has nine cycles during that 1 s interval. Their frequencies are 9 and 10 Hz, respectively. Notice how the two waves are in phase initially. Halfway, at one-half second, the waves are out of phase. The first wave has gone through five oscillations, while the second has gone through but four and a half oscillations. At the end of the entire interval, the first has gone through ten oscillations and second nine oscillations. It takes 1 s for them to be in phase again so that 1 s is the beat period (Fig. 7.16).

The wave pattern is that of a sine wave modulated by an envelope that oscillates at a frequency of 1 Hz, which is the difference between the two input frequencies. (Compare the pattern with that of the coupled SHOs, in Fig. 2.24.)

This phenomenon is called **beating**. The peaks in amplitude are called **beats**. The frequency of the envelope is called the **beat frequency**. Generally, it can be shown that if f_1 and f_2 are the two respective input frequencies, then the beat frequency is given by

$$f_{\mathrm{B}} = |f_2 - f_1|. \tag{7.11}$$

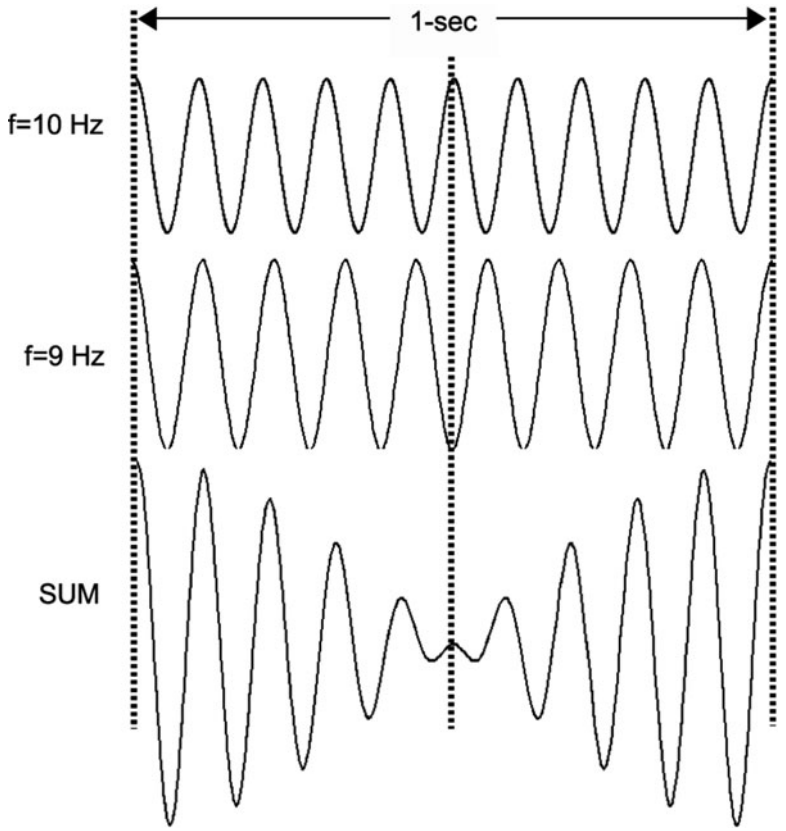


Fig. 7.16 Beats

Note the absolute value sign, since the beat frequency is a positive quantity and its value does not depend upon which of the two frequencies is the greater one.

This result can be obtained from trigonometric identities.[7]

---

[7]

$$\cos(a+b) = \cos\ a\ \cos\ b\ - \sin\ a\ \sin\ b. \tag{7.12}$$

Now let $\overline{f} = (f_1 + f_2)/2$. This is the average of the two frequencies. Assume also that $f_2 > f_1$. *The result does not depend upon which frequency is greater.* Notice that

$$f_1 = \overline{f} - f_B/2. \tag{7.13}$$

and

$$f_2 = \overline{f} + f_B/2. \tag{7.14}$$

We can then show that

$$\cos(2\pi f_1\, t)\ +\ \cos(2\pi f_1\, t) = 2\cos(\pi f_B\, t)\ \cos(2\pi \overline{f}\, t). \tag{7.15}$$

The significance of this equation is that the resulting wave is a sine wave with frequency $\overline{f}$ modulated by a sine wave having the beat frequency.

Sample Problem 7-4

What is the beat frequency if the two input frequencies are 440 and 442 Hz, respectively?

Solution

$f_B = 442 - 440 = 2\,\text{Hz}$

Sample Problem 7-5

A tuning fork whose fundamental is 440 Hz produces a beat frequency of 5 Hz when its sound is added to the sound of a violin A-string. What can one say about the frequency of the A-string?

Solution

We only know the absolute value of the difference between the two input frequencies. Thus, the violin string's frequency is either 435 or 445 Hz.

The phenomenon of beating has a number of *applications*. We will mention three of them here:

1. The tuning of stringed instruments requires high level of accuracy in matching the frequency of a string with a standard frequency source (such as a tuning fork) and/or with the frequency of other strings. Beating provides a means to attain the required accuracy.
2. In Chap. 8, we will discuss how beating is used by radar detectors to measure the speed of an automobile.
3. If a laser beam is passed through a medium consisting of a transparent liquid that has large molecules or very small sub-micron size particles in suspension, the outgoing beam will have a frequency component that is very slightly shifted from the input frequency. That shift reflects valuable information about the properties of the medium. The shift can be determined by beating an unaffected portion of the input laser beam with a portion that has been shifted. See Fig. 7.17, where a typical experimental setup is depicted. We note that the material has shifted the frequency of the laser beam from f_1 to f_2.

   Four mirrors are used to direct the laser beams. Splitting and combining are accomplished by using three **half-silvered mirror**: There are two outgoing beams, one that has been reflected by the mirror and the other that has passed straight through the mirror. They each have an intensity equal to half the incident intensity. What is remarkable is that a beat frequency of about 15 Hz can be detected as against a laser beam frequency that is on the order of 6×10^{14} Hz. This represents a sensitivity of two parts in 100 trillion!

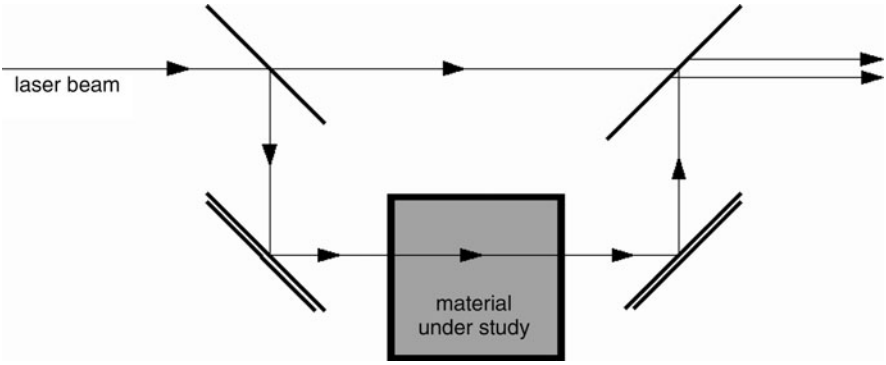


Fig. 7.17 Using beats to study materials

7.6 Terms

- Beat frequency
- Beats
- Constructive interference
- Destructive interference
- Diffraction grating
- In phase
- Order of interference
- Out of phase
- Path difference of two waves
- Phase difference
- Principle of superposition
- Relative phase
- Resultant amplitude
- Resultant wave

7.7 Important Equations

For constructive interference:

$$\overline{S_2P} - \overline{S_1P} = 0,\ \lambda,\ 2\lambda,\ \ldots. \tag{7.16}$$

Angles for constructive interference:

$$\sin\theta_1 = \frac{\lambda}{d},\ \ \sin\theta_2 = 2\frac{\lambda}{d},\ \ldots. \tag{7.17}$$

Angles for destructive interference:

$$\overline{S_2P} - \overline{S_1P} = \lambda/2,\ \lambda + \lambda/2 = 3\lambda/2,\ 2\lambda + \lambda/2 = 5\lambda/2,\ \ldots. \tag{7.18}$$

Beat frequency:

$$f_B = |f_2 - f_1|. \tag{7.19}$$

7.8 Problems for Chap. 7

1. Fig. 7.18 shows two specially shaped pulses traveling along a string, represented by thick lines segments. The blue pulse is moving to the left, while the red pulse is moving to the right. Eventually they pass through each other on the string. List below which of the five figures – (a)–(e) – represents the shape of the string at some future time.
2. Fig. 7.19 shows waves that are passing through two slits in the barrier. The solid lines are crests, and the dashed lines are troughs. Therefore, at points A, B, and C, there will be

   (a) Constructive interference, and the water will be still.
   (b) Constructive interference, and the water will be in motion.
   (c) Destructive interference, and the water will be still.
   (d) Destructive interference, and the water will be motion.
   (e) Alternating constructive and destructive interference, so the water will be in motion.

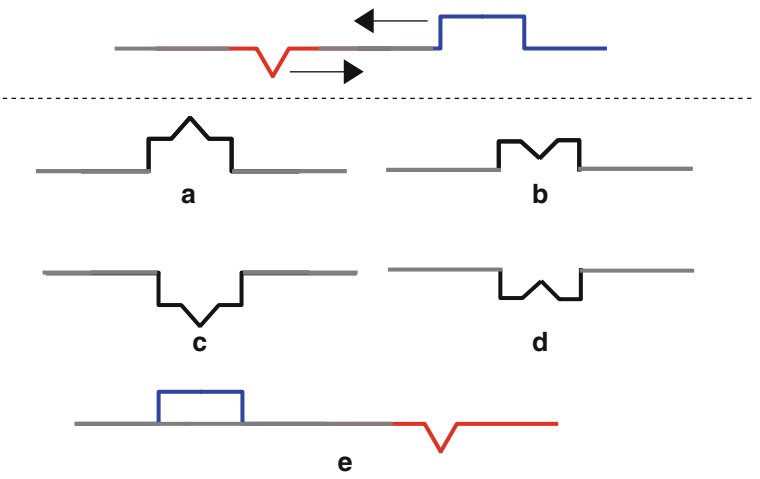


Fig. 7.18 "Colliding" waves

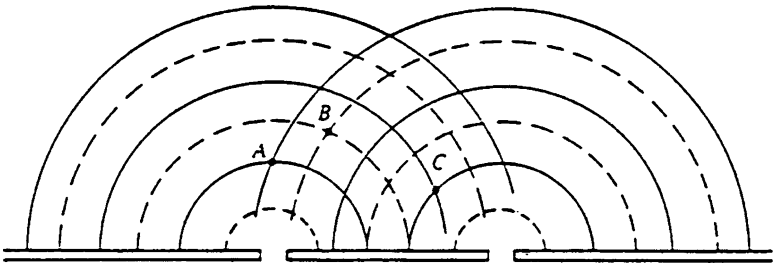


Fig. 7.19 Two slit interference

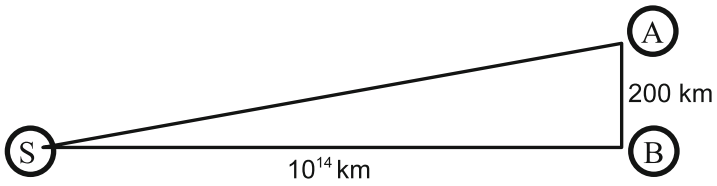


Fig. 7.20 Interference of light from a star

3. Suppose that a light wave of wavelength 4×10^{-7} m is incident upon a pair of slits that are separated by a distance of 14×10^{-7} m. Determine the angles of all the orders of interference of the outgoing wave.
4. A two-slit interference pattern of light is observed on a screen which is at a distance of 4 m from the slits. The slit separation is 0.2 mm, while the distance between neighboring fringes is 0.9 cm. Find the wavelength of the light.
5. A certain star is situated at point S, a distance 10^{14} km from the earth. An atom in the star emits a light wave of wavelength 4,000 $\mathring{A}$, and this wave is detected at two locations on earth – points A and B in Fig. 7.20.
   A and B are 200 km apart and form a triangle with the star. The signals at points A and B are added together.

   Do they add constructively or destructively? Explain using a calculation.

> **Note**
>
> You will need to use the following approximation: Given a right triangle, with legs a and b, where $a \ll b$, the hypotenuse c is given approximately by
>
> $$c \cong b + \frac{a^2}{2b}.$$

6. A piano tuner finds that two strings produce a beat frequency of 3 Hz, when one of the strings has a known frequency of 440 Hz. What can the tuner conclude about the frequency of the second string?
7. (a) Suppose that you want to determine your heart's pulse rate. You are seated next to a digital clock that is blinking at a rate of one blink per second. You cannot read the change in time. You find that when the clock has blinked 11 times, your heart has beat 13 times.

   Determine your pulse rate.[8] We will now see how we can increase our accuracy even when we never observe a beat coincide with a blink of the clock:

[8]The technique used in this problem is a simple application of the technique that Galileo is conjectured to have used to study the motion of a ball down an inclined plane. See the applet on the website (2-11-2011): http://www.joakimlinde.se/java/galileo/.

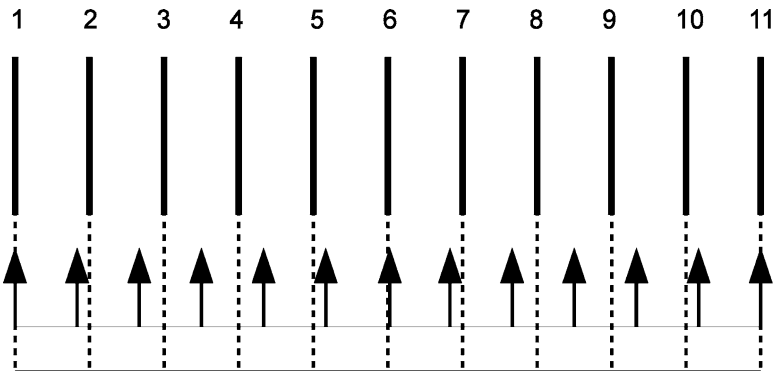


Fig. 7.21 Heart beats vs. clock blinks

In Fig. 7.21 we see marks representing the blinks of the clock, starting from the first blink at the initial time at $t = 0$. Below these marks are small marks representing the beats of my heart; we have assumed that the first beat is at $t = 0$.

(b) Suppose that we estimate that the events (blinks and beats) coincide at 4 s. What would be the corresponding estimate of my pulse?

(c) Suppose that we estimate that the events coincide at 7 s. What would be the corresponding estimate of my pulse?

(d) Suppose that we estimate that the events coincide at 11 s. What would be the corresponding estimate of my pulse?

(e) In fact, the marks were made using one-half inch spacings for the clock blinks and one centimeter spacings for the beats of my heart. What would be my pulse on this basis?
Notice how the results of parts (b)–(d) approach this last actual value.

(f) In the above example, a pair of events eventually coincides after 11 s. It is possible that a pair of events will never coincide.
Can you determine the condition on the frequencies of the two sequences of events that is necessary for events to eventually coincide?

This exercise and study has the purpose of giving us a bit of insight into the nature of the phenomenon of **beats**. Whether or not the events of two periodic sequences do coincide as displayed above, the formula for the beat frequency holds:

$$f_{\mathrm{B}} = |f_1 - f_2|. \tag{7.20}$$

(g) What is the beat frequency for the two sequences above?

8. A diffraction grating has 5,000 lines per cm. Determine the angles of the various orders of the interference pattern produced by a light beam having a wavelength of 4.4×10^{-7} m.
9. Add graphically the two waves in each of the two figures, (a) and (b), in Fig. 7.22.

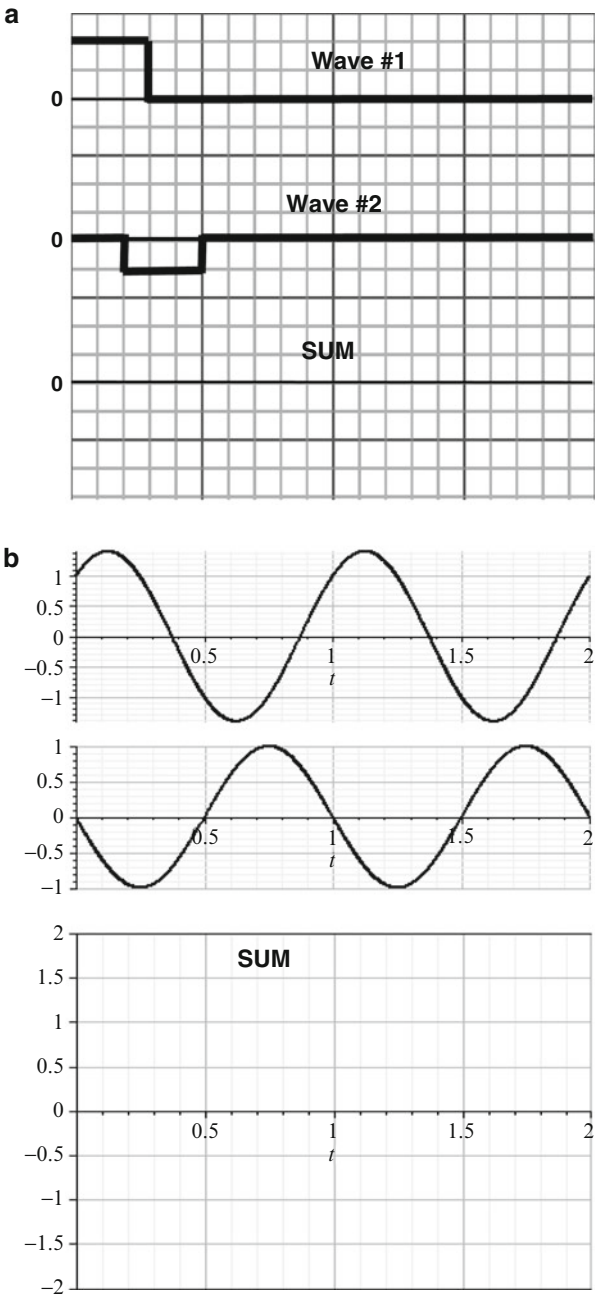


Fig. 7.22 Addition of waves

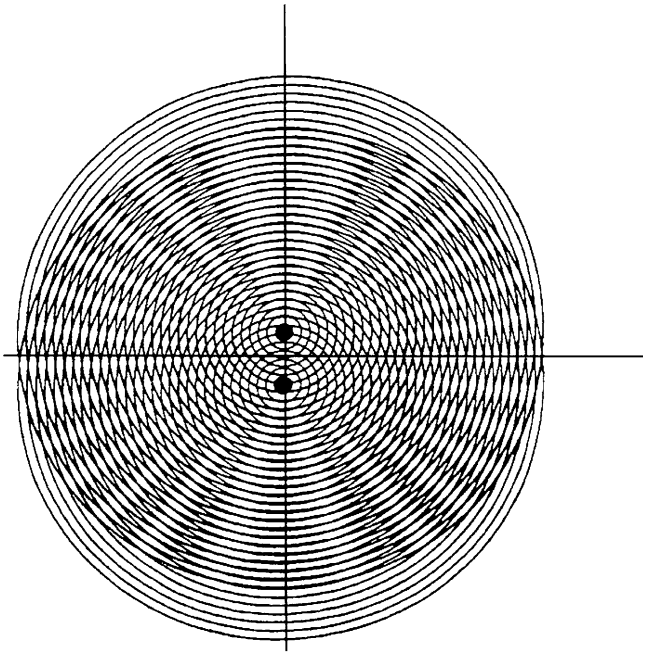


Fig. 7.23 Two spherical sources

10. In Fig. 7.23, we exhibit a drawing of two sources that are emitting waves having the same wavelength. The two sets of concentric circles are the crests. The equal spacing between neighboring circles equals the wavelength. Note the light rays that radiate from the region between the two sources. One such ray is horizontal, being the perpendicular bisector of the line joining the two sources. Note that all of the crossings between a pair of circles emitted by the two sources lie along the bright rays. These crossings are points of **constructive interference**. The bright rays are separated by dark rays; we have **destructive interference** along these rays. Both sets of rays are straight lines far from the sources.

    You are to study the second and fourth orders of interference shown in the figure as follows. Lengths should be measured using a ruler.

    (a) Determine the distance between the two sources.
    (b) For each order, draw a line from the midpoint between the sources along the center of the corresponding bright rays (at the crossings of the circles) where the rays are straight.
    (c) For each order, determine the angle that each ray makes with the horizontal. You might choose to measure the slope by completing a right triangle. The angle is the inverse tangent of the slope.
    (d) For each order, use (7.6) to determine the wavelength. Compare the value you obtain with the value you measure directly from the figure.

Chapter 8
Propagation Phenomena

Up until this point in the text, we have been focusing our attention on the sources of sound waves and light waves. The Principle of Superposition of Chap. 7 dealt with the waves produced by more than one source. In this chapter, we deal with effects on waves when they are not propagating through a single homogeneous medium that is infinite in extent. The phenomena to be studied are:

- **Diffraction**, which refers to the way waves bend around obstacles.
- **Reflection** of waves off interfaces between two media (such as sound off a wall or light off a mirror or rough surface).
- **Refraction**, which refers to the way waves behave when they are transmitted (pass on) from one medium to another (such as light from air to glass or sound from air to water). The operation of lenses, which are used in eyeglasses, microscopes, and telescopes, relies on the phenomenon of refraction.
- **Scattering** of waves by a tenuous distribution of obstacles, such as light off air molecules.
- **Doppler Effect**, which characterizes the effect on the frequency of a sine wave that is observed by a receiver that is moving with respect to the source of the wave.

It is important to keep in mind that the above phenomena are exhibited by **all types of waves**, including sound waves, EM waves, and waves propagating along the surface of a liquid, such as ocean waves. This is not true for **polarization** of electromagnetic waves, which reflects the fact that light is a transverse wave whose effective displacement, the electric field, can be in any direction perpendicular to the direction of propagation.

8.1 Diffraction

In order to study the phenomenon of diffraction, we will focus our attention on the simplest possible wave in three-dimensional space, the **plane wave**. A plane wave moves in a straight line. As a result, in filling three-dimensional space, a single crest occupies an entire plane, as seen in Fig. 8.1.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_8,

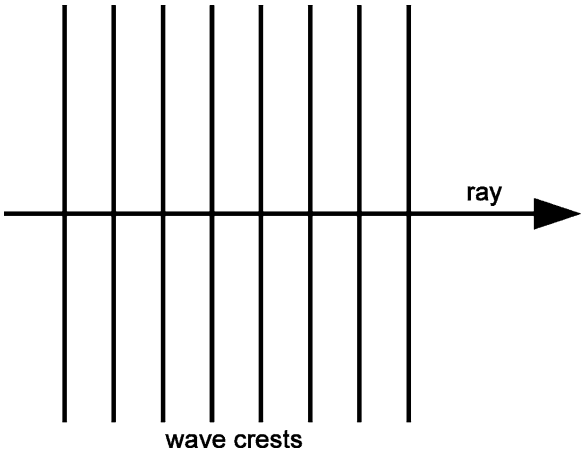


Fig. 8.1 Wave crests

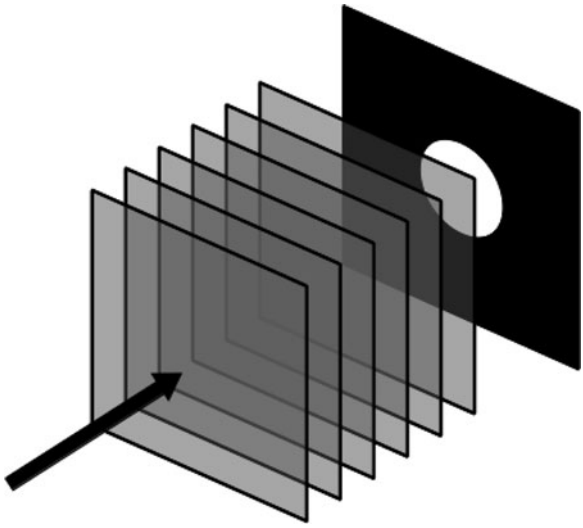


Fig. 8.2 Plane wave incident upon a hole

Suppose that a plane wave is incident on an opaque sheet of material. In the sheet we cut a round hole of diameter d. What will be the nature of the wave that progresses on the opposite side of the hole (Fig. 8.2)?

If the wave consisted of a beam of particles streaming along, we would observe a beam with a circular cross-section having a sharp boundary (Fig. 8.3):

Our normal experience with light agrees with this prediction. However, careful observation reveals this prediction to be false, or at least, in these normal

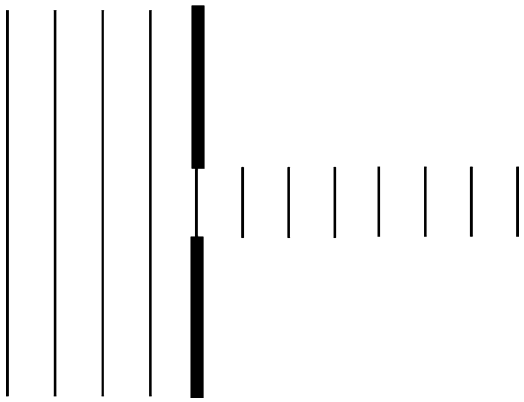


Fig. 8.3 Wave going through a hole with negligible diffraction – side view

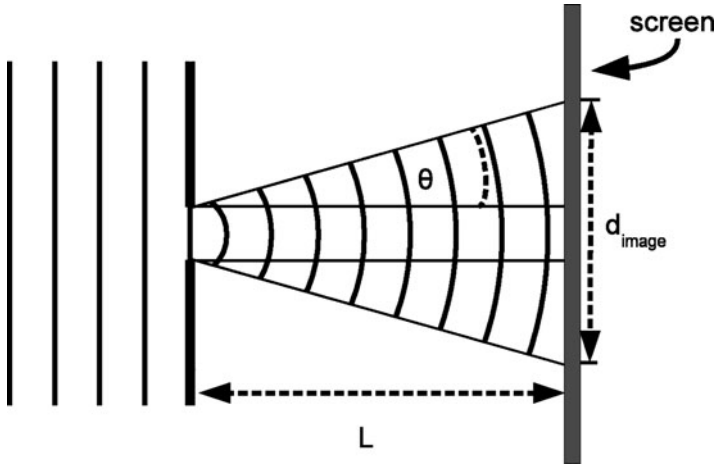


Fig. 8.4 Schematic of a wave going through a hole exhibiting diffraction

circumstances, only approximately true. What we actually observe is a "fanning out" of the beam. This phenomenon is referred to as **diffraction** and is exhibited in Fig. 8.4.

The angle θ in the figure is called the **diffraction angle**. It depends upon the wavelength λ and the diameter d of the hole. Suppose that we express the angle θ in radians. Then for small λ such that $\lambda \ll d$, it can be shown that

$$\theta \sim \frac{\lambda}{d}. \tag{8.1}$$

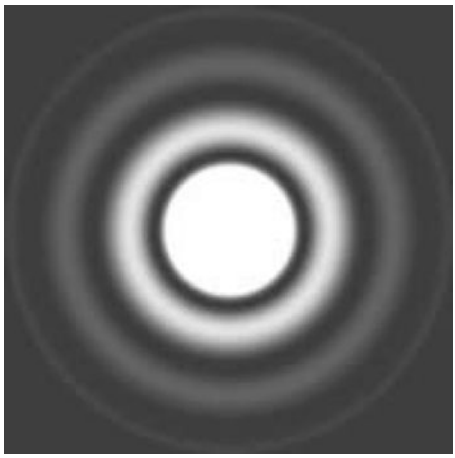


Fig. 8.5 Image produced as a result of diffraction by a hole

Recall that 2π radians equals 360°, so that one radian is about 57°. Thus, an angle θ that is about 1° or less is much less than one radian.[1]

Notice that the larger the wavelength is, the greater is the amount of diffraction.

Also, the smaller the hole is the greater the amount of diffraction.

If the beam is cast on a screen, the image will consist of a set of concentric circles, as shown below in Fig. 8.5.

While the boundary fades away asymptotically to zero at infinity, the *essential* image does have a size, with a diameter given by

$$d_{\text{im}} \approx d + 2\lambda \frac{L}{d}, \tag{8.4}$$

where L is the distance from the hole to the screen.

[1] It is interesting to note that the above relation is similar to the expression for the angle for first order constructive interference for two sources of waves. (See (7.6).) As long as the angle θ is much less than one radian,

$$\sin\theta \sim \theta, \tag{8.2}$$

where θ on the right-hand side is expressed in radians. Then the equation for the angle for first order interference becomes

$$\frac{\lambda}{d} = \sin\theta_1 \sim \theta_1. \tag{8.3}$$

In this last equation, the parameter d refers to the distance between two sources, whereas in this chapter d refers to the diameter of a hole.

A sharp image, free of diffraction, would have $d_{\text{im}} = d$. The increase of the diameter by an amount $2\pi L/d$ increases with increasing wavelength and decreasing hole diameter.

The diffraction profile described above also holds for the very beam of light emitted by a laser. Such a beam is typically regarded as being straight and of essentially constant cross-section as it propagates. In fact, it fans out as in Fig. 8.4.

Suppose that the laser beam has a wavelength of 600 nm ($= 600 \times 10^{-9}$ m; "n" = nano $= \times 10^{-9}$) and a diameter d when it leaves the laser of 1 mm. In the table below, we list the calculated beam diameter for various distances from the laser along the beam.

When $d = 1$ mm

L (cm)	$2L/d$ (mm)	$d_{\text{im}} = d + 2L/d$ (mm)
1	0.012	1.012
100 (= 1 m)	1.2	2.2

With an initial beam diameter of 1 mm, there is a significant broadening of the beam, with the beam more than doubling in diameter at a distance of 1 m from the laser.

Consider now a more common sized beam, with an initial diameter of 10 cm and the same wavelength.

When $d = 10$ cm

L (cm)	$2L/d$ (mm)	$d_{\text{im}} = d + 2L/d$ (mm)
1	0.00012	1.00012
100 (= 1 m)	0.012	1.012

Since **diffraction broadening** is only a few percent or less of the total image size, it is no wonder that we are not aware of diffraction effects of light.

We now turn to an interesting question: Suppose we have a source of light with the above frequency and we wish to cast an image with the smallest possible diameter on a screen that is a given distance of 2 m away, by varying the diameter d of the source. Without detailed thought, one might think that there is no limit to how small an image we can make. We merely have to shrink d down to as small a value as we want. However, diffraction broadening increases with decreasing source diameter, and at some point this broadening dominates the image diameter, as we can see from the table below:

d (mm)	$2L/d$ (mm)	$d_{\text{im}} = d + 2L/d$ (mm)
3	0.8	3.8
2	1.2	3.2
1	2.4	3.4
0.5	4.8	5.3

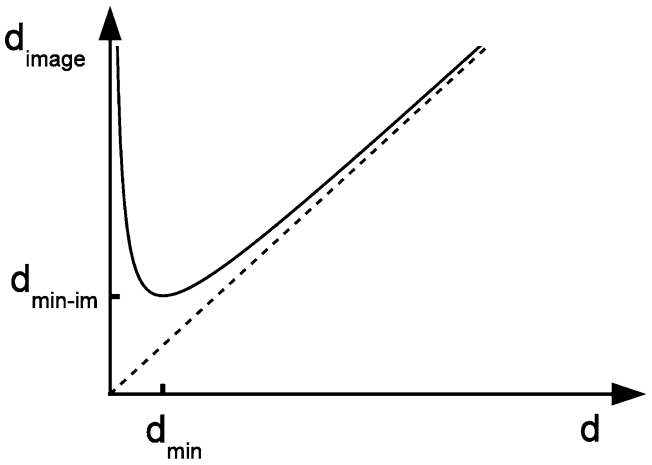


Fig. 8.6 The diameter of the image, d_{im}, vs. the diameter of the source, d

We see that as the source diameter is decreased, the image diameter first decreases and then increases. In Fig. 8.6, we present a graph of d_{im} vs. d.

Notice that d_{im} has a minimum value for a certain value d_{min} of the *source*. It can be shown that d_{min} corresponds to the source diameter and diffraction broadening being equal; that is, $d = 2L/d$. Solving this equation for d, we obtain

$$\begin{aligned} d_{\text{min}}^2 &= 2\lambda L \\ d_{\text{min}} &= \sqrt{2\lambda L} \end{aligned} \tag{8.5}$$

for a minimum image diameter.

The corresponding **minimum image diameter** is given by twice d_{min}:

$$\text{minimum image diameter} \equiv \min d_{\text{im}} = 2\sqrt{2\lambda L}. \tag{8.6}$$

Sample Problem 8-1

Find the minimum possible image size for the values $\lambda = 600\ \text{nm} = 6 \times 10^{-7}\ \text{m}$ and $L = 2\ \text{m}$.

Solution

We obtain

$$\min d_{\text{im}} = 2\sqrt{2 \cdot 6 \times 10^{-7} \cdot 2} = 3.10\ \text{mm}$$

which corresponds to a source diameter of 1.55 mm.

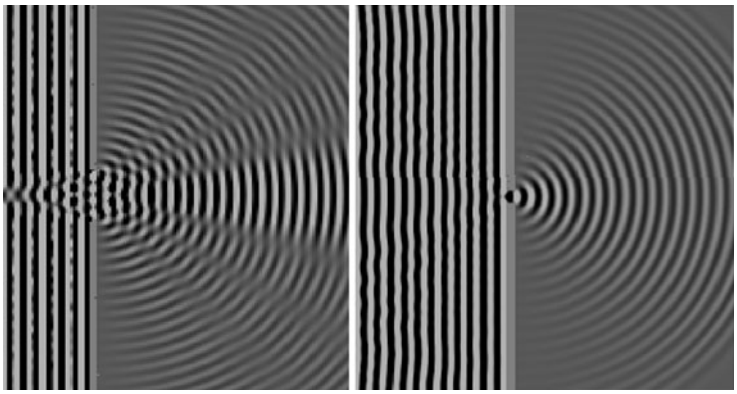


Fig. 8.7 Diffraction by a small slit

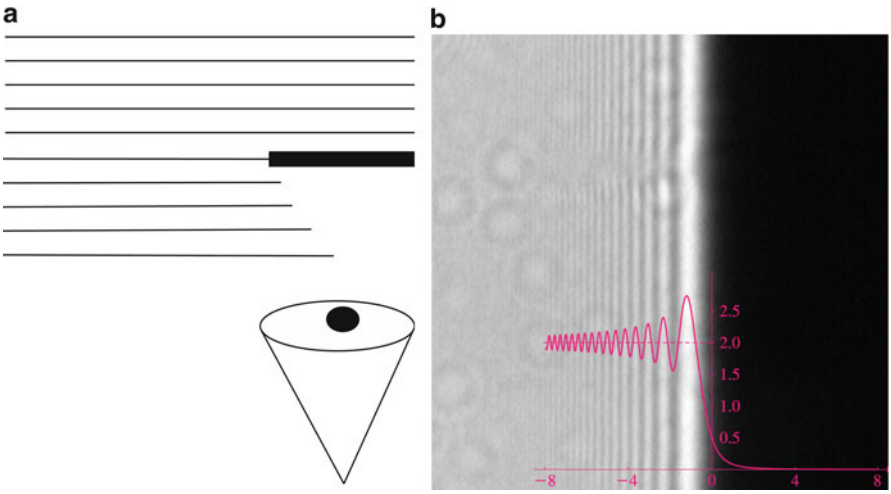


Fig. 8.8 Diffraction around an edge (source of photo: http://dlmf.nist.gov/7.SB1)

When the source diameter is less than the wavelength, the wave beyond the hole is "fanned out," producing the spherical waves of a point source. In Fig. 8.7, we see two images: to the left a small slit, to the right a minute slit with slit width much less than the wavelength.[2] Oscillations with respect to the direction from the slit are apparent in the first case, but not in the second case.

Diffraction effects are also relevant when one half of space is blocked off by a wall or a mountain, as seen in Fig. 8.8.

In the schematic drawing, we see a light wave incident downwards towards the viewer below. A screen (the black rectangle in the schematic drawing) is set up to

[2]The figures were produced with applet on the website (2-11-2011): http://www.falstad.com/ripple/.

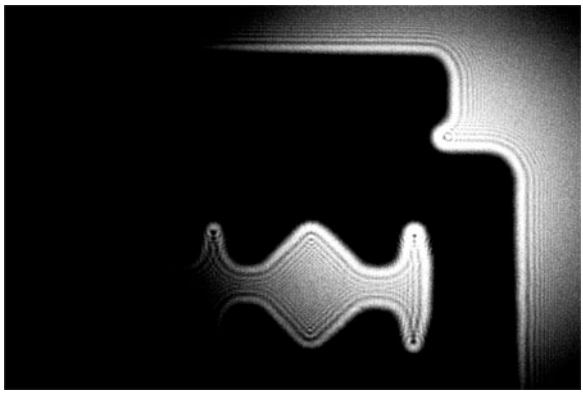


Fig. 8.9 Diffraction by a razor blade (source: Photo courtesy of Harvard Natural Sciences Lecture Demonstrations. Copyright 2011, President and Fellows Harvard College, All Rights Reserved)

block the light behind the screen. In the absence of diffraction, there would be a sharp shadow. The schematic figure shows how the wave proceeds into the shadow region behind the screen. The image to the right in Fig. 8.8 is taken at a distance of about one wavelength from the screen. A sharp boundary is depicted by the dotted line, with a value $y = 2.0$. Instead, light extends a bit into the shadow region. In addition, along the negative x-axis, the intensity oscillates, with a maximum intensity above $y = 2$ at the maxima of the oscillations. As indicated in the schematic drawing, the further away we are from the screen, the more will the light extend into the shadow region.

In Fig. 8.9, we see an actual photograph of the shadow of a razor blade. The blurriness due to diffraction is quite evident.

Questions to ponder:

1. Which voices will be heard better through a crack in a door, high-pitched or low-pitched ones?
2. Which radio waves will be more easily picked up at large distances over hilltops, AM or FM waves?

8.1.1 Scattering of Waves and Diffraction

Consider a plane wave incident upon an object as shown in Fig. 8.10. The object acts as an **obstacle** to the wave. There are the extreme cases. Let d be the diameter of the object. Then,

1. $\lambda \ll d$: The obstacle produces a sharp shadow, with mild diffraction effect that increases with increasing distance. We can see that the wave is "collapsing" around the sphere.
2. $\lambda \gg d$: The incident wave is barely affected by the obstacle. The obstacle produces a weak "**scattered wave**." This case is exhibited in Fig. 8.11.

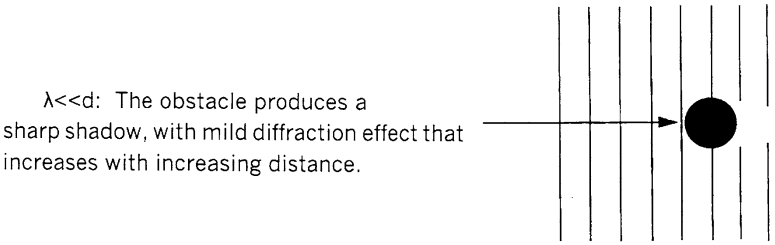


Fig. 8.10 Shadow created by a spherical object – wavelength comparable to the diameter of the object

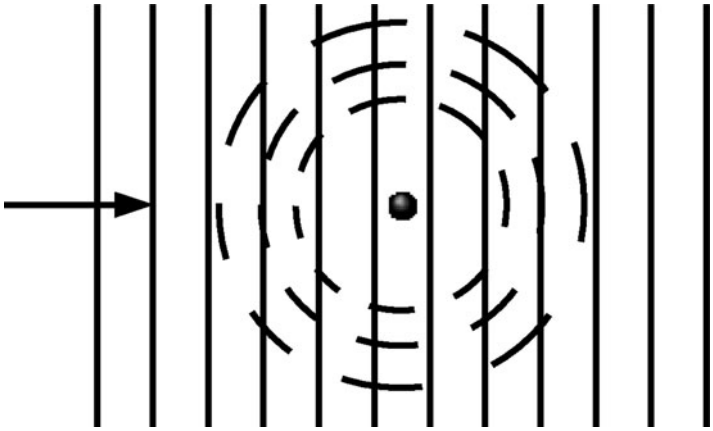


Fig. 8.11 Schematic of a wave scattered by an object with diameter much smaller than the wavelength

The two cases above that are represented schematically in the above figures are shown in Fig. 8.12 produced by a simulation of a surface wave in water.[3]

Note that in (a) there is strong blockage yet we can easily see the effects of diffraction; in (b) there is strong diffraction, so that the wave is barely affected by the presence of the object. Still, if you look carefully you should see the weak spherical scattered wave.

Home exercise: Observe the effect your body has on large ocean waves.

The above discussion allows us to understand why it is that ordinary laboratory microscopes that use light for illumination cannot allow us to clearly see objects that are on the order of the wavelength of light ($\sim 5 \times 10^{-7}$ m or less). Diffraction produces blurry boundaries. We can also appreciate why AM radio waves can "cross" mountains more easily than FM radio waves, which have a higher frequency and hence a smaller wavelength.

[3]The figure was made from the applet on the website (2-5-2011): http://www.falstad.com/ripple/.

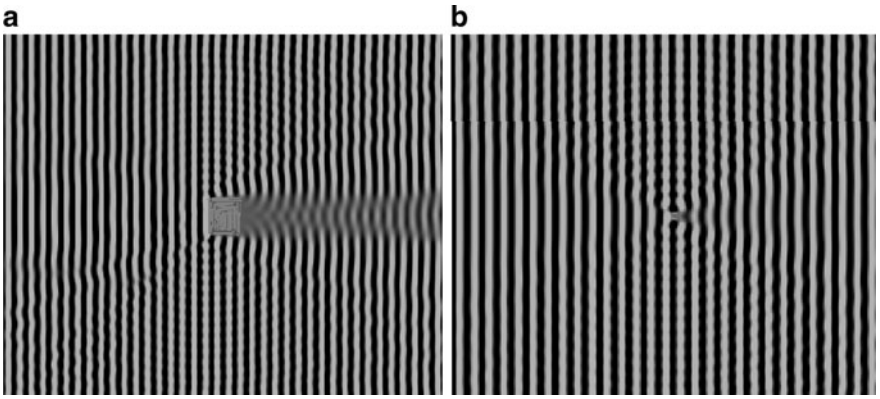


Fig. 8.12 The scattering of a wave: (**a**) object diameter a few times the wavelength; (**b**) object diameter much smaller than the wavelength

8.1.2 Why Is the Sky Blue?

Our sky is blue because of the scattering of sunlight by the molecules of air of the atmosphere. How can we understand this phenomenon? After all, the diameter of air molecules is on the order of a few Ångstroms, while the wavelength of light is much greater $\lambda \sim 5{,}000\,\mathring{A}$. That amounts to a ratio of about 1,000 to 1. As a result, diffraction effects are very strong so that *one would expect that very little scattering of light by a single molecule would take place*. (See Fig. 8.11.) The sky is bright on account of the sum of scattering by a vast number of molecules.[4]

We can now understand why the sky is blue. A detailed analysis was provided about 100 years ago, when Lord Rayleigh (alias John William Strutt) showed that the intensity of scattered light behaves like:

$$I_{\text{scatt}} \propto \frac{1}{\lambda^4} \tag{8.7}$$

which gives the relationship between the intensity of scattered light and the wavelength of the light.

We should expect the degree of scattering to increase with decreasing wavelength since there will be decreasing diffraction. As a consequence, the violet end of the visible spectrum, which has the shortest wavelength, is scattered most. For

[4]Interestingly, one can show that if the density of air molecules was to be perfectly uniform, this sum would result in no net scattering; the sky would be perfectly transparent! It is the modest degree of nonuniformity of the density that is responsible for the scattering. In fibre optics communication, the glass is so pure, that is, free from impurities and inhomogeneities, that it is the small degree of nonuniformity in the molecular density associated with the random thermal motion of the molecules, that is responsible for the small attenuation in the fibres.

a numerical comparison, we will calculate the ratio of the scattered light for a wavelength $\lambda = 4 \times 10^{-7}$ m (violet) to that for $\lambda = 7 \times 10^{-7}$ m (red):

Now sunlight is whitish, consisting of a mixture of wavelengths which extends from the infrared to the ultraviolet. Since the violet end of the visible spectrum is scattered most, scattered sunlight looks bluish. On the other hand, when we look directly at the sun through the atmosphere or observe the horizon in the west at sunset, we are seeing light which has started out white (from the sun) and has had the violet end of the spectrum removed most by scattering. Such light looks reddish. (See Chapter 14, THEORY OF COLOR VISION, for details on color perception.)

8.2 Reflection

The characteristics of a wave that is reflected off an object can be, surprisingly, quite complicated, whether it be a light wave, sound wave, ocean wave, or any other wave. We will discuss three relatively simple cases; they can be described in terms of the appearance of a surface under the reflection of light: **shiny surface** (like a mirror), **dull surface**, and **sparkling surface**.

Let us first consider the reflection of *light* off a painted wall. Some walls are dull; others, with a concentrated enamel paint are shiny. What is the physical difference between the two surfaces? We know that rubbing a surface often polishes the surface, meaning that the surface is made shiny. We recognize that polishing involves making a surface smoother. But how smooth must a surface be to be shiny? How do we characterize smoothness?

The central factor is what we will refer to as the **length scale of roughness**, with a symbol ℓ_r. The smaller this length is, the smoother a surface is. The degree of shininess is obtained by comparing this length to the wavelength λ of the light. (Recall that the range of wavelengths of light λ is about $4{,}000 - 7{,}000\,\mathring{A}$.)

We can describe the surface of the wall with varying degrees of detail. The surface might be smooth to the touch, so that without close scrutiny, we would simply describe the surface as being smooth. However, if we examine the surface with a microscope, we might be able to see bumps on the order of 0.1 mm in size. We would say that the length scale of roughness ℓ_r is about 0.1 mm. In this case, $\lambda \ll \ell_r$ and the surface will appear *dull*. Examination will reveal that a beam of light is reflected from the wall in many random directions. We have what is referred to as **diffuse reflection** (Fig. 8.13).

> **Note**
>
> If a laser beam casts a spot on a *dull wall*, everyone in the room can see the spot, since there exists a ray that reaches the observer's eye.

In the case of a wall painted with *enamel* paint, the length scale of roughness is on the order of a wavelength, so that the surface is shiny. When the length scale of

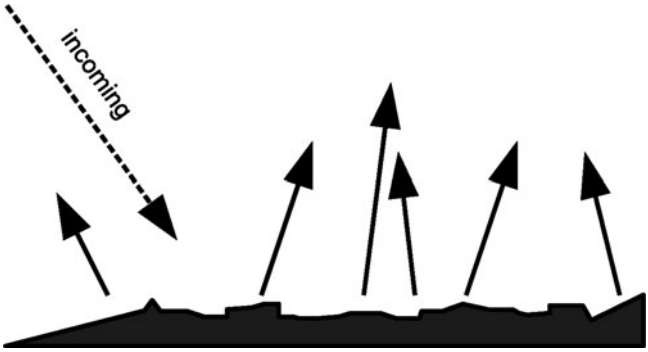


Fig. 8.13 Diffuse reflection off a rough surface

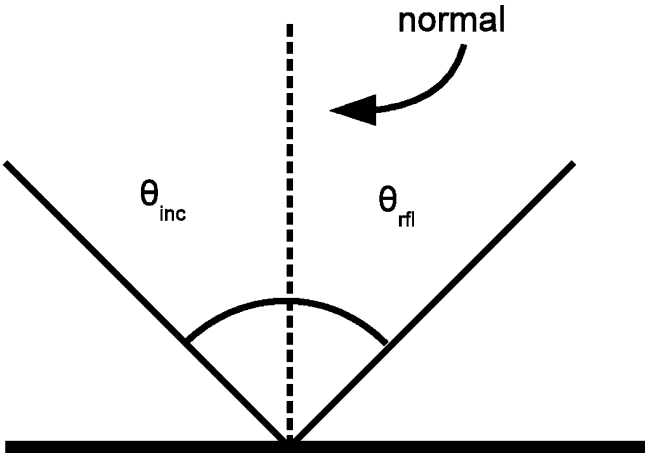


Fig. 8.14 Reflection by a smooth surface

roughness is much smaller than a wavelength, the surface acts like a mirror and we have what is referred to as **specular reflection** or **mirror reflection**. A given ray of light produces a single reflected ray as exhibited in Fig. 8.14.

We see that an incident ray of light produces a reflected ray, with the **angle of incidence** θ_{i} equal to the **angle of reflection** θ_{rfl}. Notice that these angles are measured relative to a line perpendicular to the interface. This line is called the **normal**, since the word "normal" means perpendicular.

Suppose that two viewers, labeled V_1 and V_2, respectively, look at a point source through a mirror. The point source emits rays in all directions, but each viewer receives through their own eye only a small *set* of rays concentrated around the respective rays in Fig. 8.15:

In the figure, S is the point source while I is the apparent position of S – as seen by V_1 and V_2. I is said to be the **mirror image** of S. Note that $\overline{SM} = \overline{IM}$ and

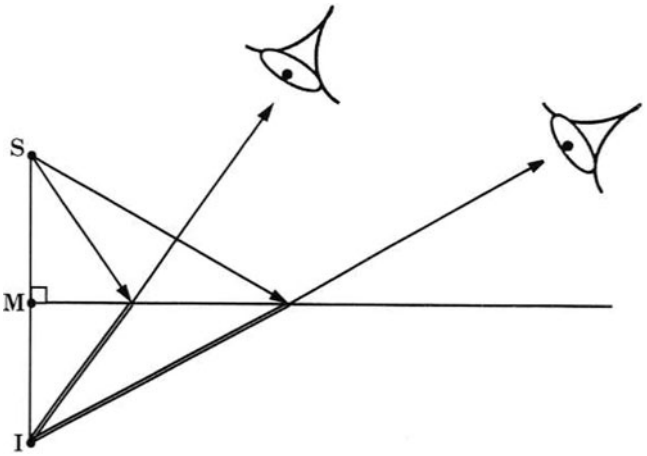


Fig. 8.15 Two observers, each receiving its own respective set of reflected waves from the source S

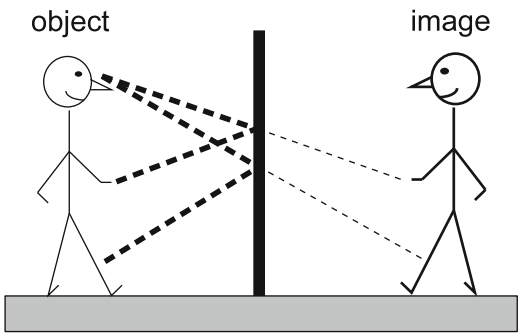


Fig. 8.16 The rays of light from a hand and from a leg that reflect from a mirror into the eye

that $\overline{SMI}$ is perpendicular to the interface. The arrowed line segments represent the actual rays traveling toward the viewers. The key result is that the viewers' brains will interpret the incoming rays as coming from S and perceive the source as actually being at I. Since the source is not really at I, the image is said to be a **virtual image**.

In Fig. 8.16, we see a person standing in front of a mirror. The particular rays that reflect off a hand and off a leg and which strike the mirror and reach the eye (as dashed lines) are indicated. The visual system, whose main components are the eye and the brain – the **eye-brain system** – assumes that the ray left a point on the leg of the image (as a fine-dashed line).

Note that if a laser beam is reflected off the wall under specular reflection, only an eye – regarded approximately as having an iris the diameter of one ray of light – that is in line with the single reflected ray will be able to see the light.

Consider now a light beam incident on a dull painted wall. Its dullness reflects the fact that you can see an image of the light beam on the wall no matter where you

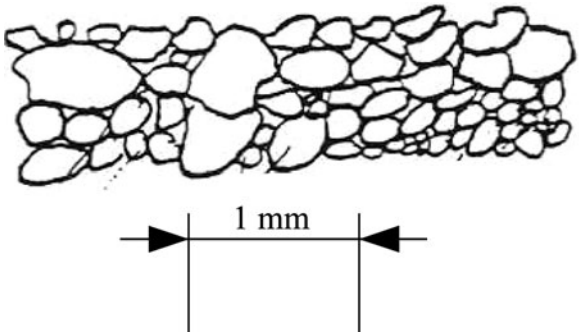


Fig. 8.17 Magnified surface of sand

stand. This situation indicates that the reflection is diffuse, as shown in Fig. 8.13. Since the light has a wavelength lying between 4×10^{-7} and 7×10^{-7} m, to have specular reflection the length scale of roughness must be on the order of 10^{-7} m or more.

In sum, if $\lambda \ll \ell_r$, we have **diffuse reflection**; while if $\lambda \gg \ell_r$, we have **specular reflection**.

These results hold for the reflection of any type of wave off a surface. Examples are sound waves off walls, a concert audience, or wall tapestry, as well as radio waves off a forest of trees or suburban houses. Since the smallest wavelength of audible sound (corresponding to the largest audible frequency) is $\lambda = v/f = (340\,\text{m/s})/(20{,}000\,\text{Hz}) = 0.17\,\text{m} = 170\,\text{mm}$, we see that *audible sound waves reflect specularly off walls*. (See also the problems at the end of the chapter.)

8.2.1 A Complex Surface: A Sand Particle

Let us consider how we would regard a *flat* surface of **sand**, flat on a scale of centimeters, as depicted below. We know that the sand consists of a multitude of sand particles, of varied shapes and sizes. The range of particle diameters certainly does not exceed a value on the order of 1 mm. Depending upon the sample of sand, it may not fall below 0.1 mm.

For our purposes, a grain of sand has at least three important **length scales**:

First, there is the *size* (average diameter) ℓ_g of the *grain of sand*. Next, if one looks closely at the grain of sand, one would find the surface rough and bumpy, as seen in Fig. 8.17. The second length scale is the average size of the bumps on the grain's surface, ℓ_r in the above figure. This is the length scale of roughness. Finally, there is the size ℓ_a of the individual *atoms*, at the Ångstrom level. Certainly, $\ell_g > \ell_r \gg \ell_a$. We will assume, for simplicity, that $\ell_g \gg \ell_r$; that is, the bumps are much smaller than the grain size.

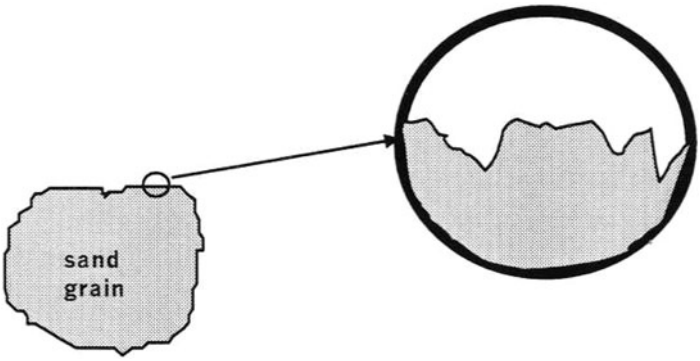


Fig. 8.18 Surface of a grain of sand – magnified

Let us now discuss the reflection of light off the surface of the sand depicted in Fig. 8.18. Certainly, λ is much smaller than the size of a grain. As a consequence, the surface of the sand appears rough. Furthermore, the wavelength λ of the light is much greater than the average atomic diameter ℓ_a. [$\ell_a \ll 5000\,\mathring{A}$.] The only remaining question is how λ compares with the size of the bumps, ℓ_r. If $\lambda \gg \ell_r$, the particle surface will appear shiny and the sand as a whole will sparkle. Otherwise, the sand will appear dull at all levels.

Questions:
How would sound waves reflect off an audience of people?
How would AM or FM radio waves reflect off a forest of trees or a suburban neighborhood of homes?

8.3 Reflection and Reflectance

Suppose that a fisherman is standing at the edge of a lake and is concerned that if he talks the fish will hear him. To answer this question, we need to determine what fraction of sound energy that is incident on the surface of the water is transmitted into the water. Or, suppose that a light beam is propagating in a transparent medium and is incident upon a second transparent medium, how does the intensity of the transmitted beam compare to the intensity of the incident beam?

In Chapter 4, ENERGY, we discussed absorption and attenuation of waves as they propagate in a medium. Often one talks about the fraction of an incident wave that is **reflected**, **transmitted**, and **absorbed**. When a sound wave traveling in air is incident upon a dense material such as acoustic tile, we focus on absorption as a process rather than transmission. For our purposes, we will disregard the details of what happens to a transmitted wave once it enters the second medium.

The ratio of the intensity of the reflected wave to the intensity of the incident wave is referred to as the **reflectance** with a symbol R; the ratio of intensities of the

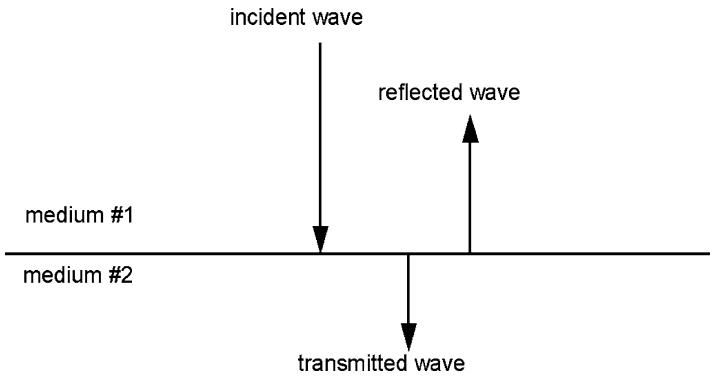


Fig. 8.19 Reflectance and transmittance of an incident wave

transmitted wave and the incident wave is referred to as the **transmittance**, with a symbol T. The incident wave is replaced by the reflected and the transmitted waves. Thus,

$$R = \frac{I_{\text{rfl}}}{I_{\text{inc}}}$$
$$T = \frac{I_{\text{trans}}}{I_{\text{inc}}}. \tag{8.8}$$

Since the total intensity must equal the total resulting intensity. That is,

$$I_{\text{rfl}} + I_{\text{trans}} = I_{\text{inc}}. \tag{8.9}$$

Alternatively,

$$R + T = I. \tag{8.10}$$

In answering these questions, we will restrict ourselves, for simplicity, to a plane wave that is incident **perpendicularly** on a plane surface. See Fig. 8.19. (When the angle of incidence is not 0°, the results are too complicated for us to discuss here.)

8.3.1 The Reflectance for a Light Wave

The reflectance of light at an interface between two transparent materials depends upon what is called the **index of refraction**, which is given the symbol n. In vacuum, the speed of light is $c = 3.00 \times 10^8$ m/s. In a transparent medium it is given by

$$v = \frac{c}{n}. \tag{8.11}$$

Clearly, $n = 1$ in a vacuum. In a medium, the speed of light is less than c, so that $n > 1$.

Below, we list the values of the index of refraction for some materials:

Material	Index of refraction n
Air (STP)	1.0003 (The denser the air, the greater the index of refraction)
Water	$1.33 \sim 4/3$
Diamond	2.42
Crown glass	1.52
Flint glass	1.66

Unless otherwise stated, we will set $n = 1$ for air as an excellent approximation.

The reflectance R of a light wave is determined solely by the respective indices of refraction, n_1 and n_2 of the two media. It is given by:

$$R = \left[\frac{n_1 - n_2}{n_1 + n_2}\right]^2. \tag{8.12}$$

Note

If $n_1 = n_2$, then $R = 0$. There is no reflected wave if the indices are equal – even if the two materials are otherwise quite different! Only the indices of refraction determine the reflectance.

Note

The reflectance is the same whether the incident ray is in medium #1 or in medium #2.

The basis for this result can be seen as follows:

The result of interchanging media can be determined by interchanging the symbols n_1 and n_2 in (8.12). The value for the reflectance does not change since

$$\frac{(n_1 - n_2)^2}{(n_1 + n_2)^2} = \frac{(n_2 - n_1)^2}{(n_2 + n_1)^2}.$$

This result will be shown to hold for the reflection of sound waves, too. (See below.)

Sample Problem 8-2

Find the reflectance for light at an air–water interface, given that $n_1 = 1$ and $n_2 = 4/3$.

Solution

$$R = \frac{\left(1 - \frac{4}{3}\right)^2}{\left(1 + \frac{4}{3}\right)^2}$$
$$= \left[\frac{-\frac{1}{3}}{\frac{7}{3}}\right]^2 = \left(-\frac{1}{7}\right)^2 = \frac{1}{49} \approx 0.02.$$

Thus only 2% of the energy is reflected. Most of the energy is transmitted.

Note

Without being quantitative, we will say something about the case when the incident wave is *not* perpendicular to the surface. **The reflectance increases as the angle of incidence increases.** Check this by examining reflection off a surface. Note that the shininess is dramatic if you view a surface at a grazing angle even if the surface is dull!

Problem for the reader: Using the relation $v = c/n$, show that the reflectance for an EM wave is given by

$$R = \left[\frac{v_2 - v_1}{v_2 + v_1}\right]^2, \tag{8.13}$$

where v_1 and v_2 are the wave velocities of the respective media.

8.3.2 The Reflectance for a Sound Wave

Here the different mass densities ρ_1 and ρ_2 must be taken into account:

$$R = \left[\frac{\rho_2 v_2 - \rho_1 v_1}{\rho_2 v_2 + \rho_1 v_1}\right]^2. \tag{8.14}$$

Note that the interchange subscripts in the expression leaves the reflectance unchanged as in the case of light waves. Note too that (8.13) for light waves can be obtained from (8.14) by assuming that the media through which light propagates all have the same mass density.

In (8.14), the product "ρv" is called the **impedance** of the medium. The typical symbol for the impedance is Z, so that

$$Z \equiv \rho v. \tag{8.15}$$

Then we can write:

$$R = \left[\frac{Z_2 - Z_1}{Z_2 + Z_1}\right]^2. \tag{8.16}$$

Sample Problem 8-3

Find the reflectance for an air–water interface at STP, given that

Air: $Z = \rho v = 1.3\,\text{kg/m}^3 \times 345\,\text{m/s} = 450\,\text{kg/m}^2\,\text{s}$
Water: $Z = \rho v = 1{,}000 \times 1{,}500 = 1.5 \times 10^6\,\text{kg/m}^2\,\text{s}$.

Solution

Substitution into (8.14) leads to $R = 0.999$. Thus, only 0.1% of the sound energy is transmitted! On the basis of this result, a fish should have difficulty hearing a person who is on the shore talking.

Sample Problem 8-4

How many dB corresponds to 0.1%?

Solution

$\Delta SL = 10\log(0.001) = -30\,\text{dB}$.

8.4 Refraction

Suppose that a plane wave is traveling in one medium and is incident upon an interface of this medium with a second medium through which the wave can propagate. Examples include light in air that is incident upon an air/glass interface, or a sound wave in air, incident upon an air/water interface.[5] We have both a reflected wave and a **transmitted wave**. We focus here on the transmitted wave.

In Fig. 8.20, we exhibit a wave incident on an interface. Notice that the transmitted wave is not in the same direction as the incident wave. This change in direction is called **refraction**. Of greatest interest is the relationship between the **angle of incidence** θ_i and the **angle of refraction** θ_r.[6]

[5] We are used to referring to the interface between air and water as the **surface** of water. However, how should we refer to the boundary between water and oil? The word **interface** is a neutral term, clearly superior to the term **surface**.

[6] Not to be confused with the angle of reflection θ_rfl.

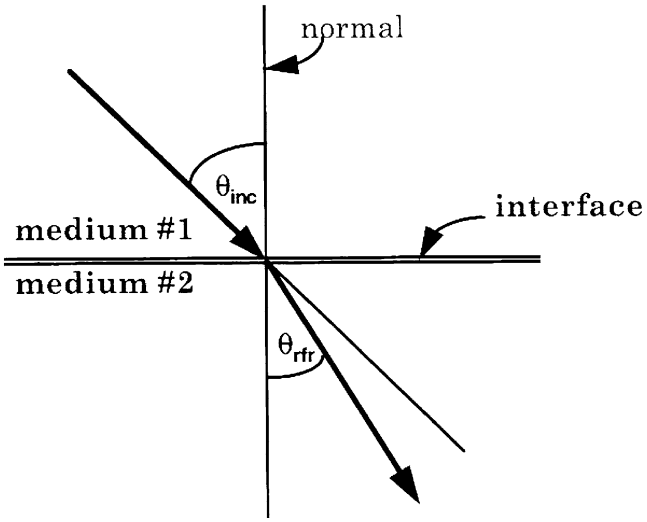


Fig. 8.20 Refraction at the interface of two media

In 1621, Willebrord Snell discovered a relationship between these two angles for light beams. They are related by the indices of refraction of the two media.

$$n_{\mathrm{i}} \sin\theta_{\mathrm{i}} = n_{\mathrm{r}} \sin\theta_{\mathrm{r}}. \tag{8.17}$$

Applications of this law follow below:

Sample Problem 8-5

Given $n_{\mathrm{i}} = 1$, $\theta_{\mathrm{i}} = 60°$, and $n_{\mathrm{r}} = 4/3$, what is θ_{r}?

Solution

We have $1 \times \sin 60° = (4/3) \sin\theta_{\mathrm{r}}$, so that $\sin\theta_{\mathrm{r}} = (3/4) \sin 60° = 0.650$. We then obtain

$$\theta_{\mathrm{r}} = \arcsin(0.650) = 40.5°.$$

Sample Problem 8-6

Given $n_{\mathrm{i}} = 4/3$, $\theta_{\mathrm{i}} = 40.5°$, and $n_{\mathrm{r}} = 1$, what is θ_{r}?

Solution

It should be clear from Snell's Law that $\theta_{\mathrm{r}} = 60°$. This example describes the path of the ray of the previous problem when reversed in direction! See Fig. 8.21.

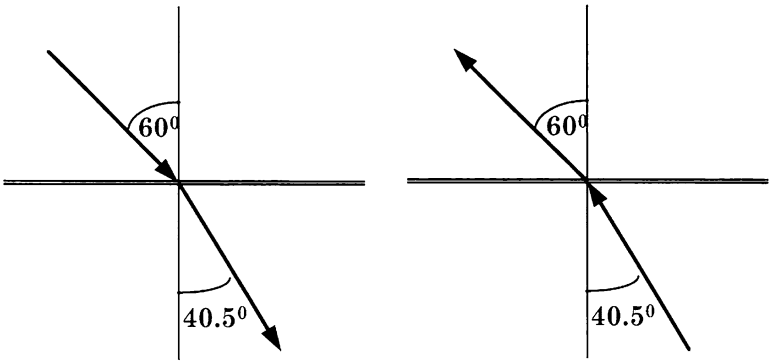


Fig. 8.21 Refraction for two opposite directions of a ray of light

Note

It follows generally from Snell's Law that if an incident ray were to have the direction of the original refracted ray, the resulting refracted ray would be in the direction of the original incident ray.

Comments

1. Suppose $\theta_i = 0°$, then $\theta_r = 0°$: Thus, an incident ray which is perpendicular to the interface is not refracted.
2. If $n_i > n_r$, the beam is refracted *away from* the normal. If $n_i < n_r$, the beam is refracted *toward* the normal.

8.5 Total Internal Reflection

Consider the following problem.

Sample Problem 8-7

Given that $n_i = 4/3$, $n_r = 1$, and $\theta_i = 60°$, find θ_r.

Solution

From Snell's Law, $n_r \sin\theta_r = n_i \sin\theta_i$ so that

$$1 \times \sin\theta_r = (4/3)\sin 60° = 1.15.$$

This last equation leads to $\sin\theta_r$ having to be equal to 1.15, which is *impossible*; *this equation has no solution* since the sine of an angle cannot be greater than unity. Then, what does it mean to have an equation (Snell's Law) that has no solution?

What happens is that we have *no transmitted refracted beam*. We have what is called **total internal reflection**. Generally, refraction is always accompanied by reflection; the fraction of the intensity that is reflected vs. the fraction that is refracted depends upon the two indices of refraction and the angle of incidence. In this case, there is no transmitted, refracted ray.

Total internal reflection can happen only if the index of refraction of the medium of the incident ray is greater than the index of refraction of the medium of the transmitted ray. Thus, a ray incident from air onto an air–water interface cannot be totally reflected. This conclusion should be evident from Snell's Law:

As θ_i is increased, so is θ_r. But $\theta_r > \theta_i$. Therefore, θ_r will reach 90° before θ_i does. And, if θ_i were to be further increased, there is no solution to Snell's equation. Then we would obtain

$$\sin\theta_r = (n_i/n_r)\sin\theta_i > 1$$
$$\sin\theta_i > n_r/n_i.$$

The angle for which this last inequality is replaced by an equality is called the **critical angle**, which we give symbol θ_c. It satisfies the equation

$$\sin\theta_c = \frac{n_r}{n_i}. \tag{8.18}$$

In sum, we have *total internal reflection when the angle of incidence exceeds the critical angle*. For this situation to be possible, we must have $n_i > n_r$.

For a water–air interface (with the incident ray in the water), we obtain a critical angle of 48.6°:

$$\sin\theta_c = 1/(4/3) = 3/4$$
$$\theta_c = \arcsin(3/4) = 48.6°.$$

NOTE: Even when there is a refracted ray, there is a reflected ray. In the case of total internal reflection, there is no refracted ray.

8.6 The Wave Theory of Refraction

There were two theories of light presented to account for refraction: Newton's was based on a particle theory, while **Christian Huyghens**' theory of refraction was based on his wave theory of propagation (Fig. 8.22).

Fig. 8.22 Christian Huyghens (source: http://upload.wikimedia.org/wikipedia/commons/thumb/0/03/Christiaan-huygens4.jpg/170px-Christiaan-huygens4.jpg)

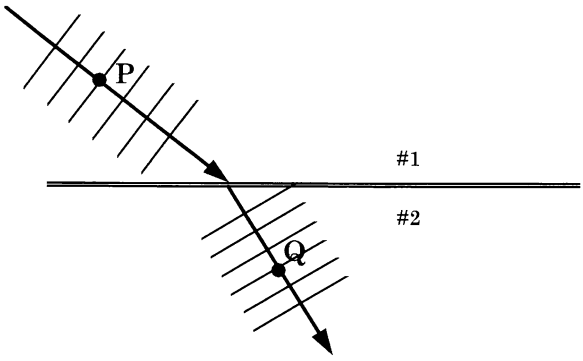


Fig. 8.23 Refraction of waves

Here is Huyghens theory: We assume that a wave is traveling in medium #1 and is then transmitted into medium #2, in which the wave velocity is smaller. The most important thing to note is that the *frequency of a wave is unchanged upon transmission*.[7] As a result the wavelength is smaller in medium #2, as seen in Fig. 8.23. The result also follows from the equation $\lambda = v/f$, with reduced wave velocity and constant frequency.

[7]To see this, suppose that we have observers at points P and Q, in the two respective media. The wave proceeds in a continuous manner. Thus, the rate f_1 at which crests pass point P must equal the rate f_2 at which crests pass point Q. Thus, we will replace the two symbols f_1 and f_2 by the common symbol f.

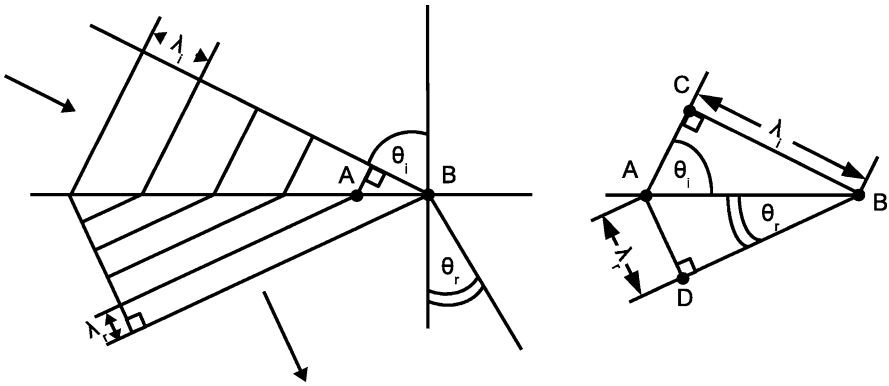


Fig. 8.24 Detailed schematic of the Huyghen's wave theory of refraction

Next, we see that the two triangles, ABC and ABD, in the offset are *similar* (Fig. 8.24).

Finally, recall that $\lambda_1 = v_1/f$ and $\lambda_2 = v_2/f$, where v_1 and v_2 are the wave velocities in the two respective media. Then,

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{\frac{v_1}{f}}{\frac{v_2}{f}}. \tag{8.19}$$

or

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}. \tag{8.20}$$

This last equation is a general expression for the refraction of waves. Note that it makes no reference to the index of refraction. In fact, **it can be applied to sound waves** as well as light waves. For example, consider the following problem:

Sample Problem 8-8

A sound wave in air is incident at an angle of 10° on a water–air interface. Find the angle of refraction.

Solution

We have $v_\mathrm{i} \cong 340\,\mathrm{m/s}$ and $v_r = 1{,}400\,\mathrm{m/s}$. Thus,

$$\sin\theta_\mathrm{r} = \frac{v_r}{v_\mathrm{i}}\sin\theta_\mathrm{i} = (1{,}400/340)\times\sin 10^\circ = 0.72.$$

so that

$$\theta_\mathrm{r} = 46^\circ.$$

We now return to light waves and **derive Snell's Law** from (8.20) above: We have

$$v_{\mathrm{i}} = \frac{c}{n_{\mathrm{i}}} \qquad \text{and} \qquad v_r = \frac{c}{n_{\mathrm{r}}}. \tag{8.21}$$

so that

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{\frac{c}{n_1}}{\frac{c}{n_2}} = \frac{n_2}{n_1}. \tag{8.22}$$

This is Snell's Law in an algebraically rearranged form.

> **Note**
>
> For "refraction" of a beam of **particles**, it can be shown that (cf. (8.20))
>
> $$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_2}{v_1}. \tag{8.23}$$

Then, if $\theta_2 < \theta_1$, we also have $v_2 > v_1$. That is, a beam should be refracted **away** from the normal on passing from air into glass. The experimental finding about 150 years ago, that for *light waves* the opposite is true (i.e., $v_2 < v_1$), gave further confirmation (after the interference experiments described in Chap. 7) that light propagates as a wave.

8.7 Application to Mirages

On a very hot day, the ground will be heated up to temperatures greatly exceeding the temperature of the air above. As a consequence, the air close to the ground will be hotter than the air above. The hotter air has a lower density than the cooler air above and hence has a lower index of refraction. Now imagine a ray of light that originates from a region where the index of refraction is larger than that at ground level and that propagates downward toward the ground. We can have an occurrence of **total internal reflection**. The light ray can strike someone's eye, thus producing an effect that is referred to as a **mirage**. This phenomenon is exhibited in Fig. 8.25.

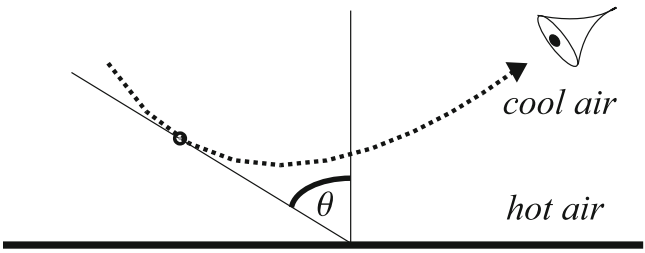


Fig. 8.25 A mirage

According to Snell's Law applied to refraction at the boundary between two homogeneous media, the product $n \cdot \sin\theta$ is the same for the two media in which a ray travels. In the case of the hot air, the index of refraction varies continuously. It can be shown that Snell's Law applies too in the same form: **The product $n \cdot \sin\theta$ is constant all along the path of the ray**, where θ is measured from the vertical. The angle θ corresponding to a point along the dotted ray in Fig. 8.25 is measured relative to the vertical.

The ray will continuously become ever more horizontal as it propagates and the index of refraction decreases. If it becomes absolutely horizontal (i.e., $\theta = 90°$), thereafter it will propagate upward. The ray experiences total internal reflection.[8]

8.8 The Prism

In Fig. 8.26, we depict a transparent **prism**, with a ray of light incident on one face and a refracted ray leaving a second face.

Note in the figure how the directions of the two refracted beams, one inside the prism and the other outside the prism, are each determined with respect to their respective normals, being toward the normal #1 for the first one, since the ray is going from air into the prism and away from the normal #2 for the second one in going from the prism into the air.

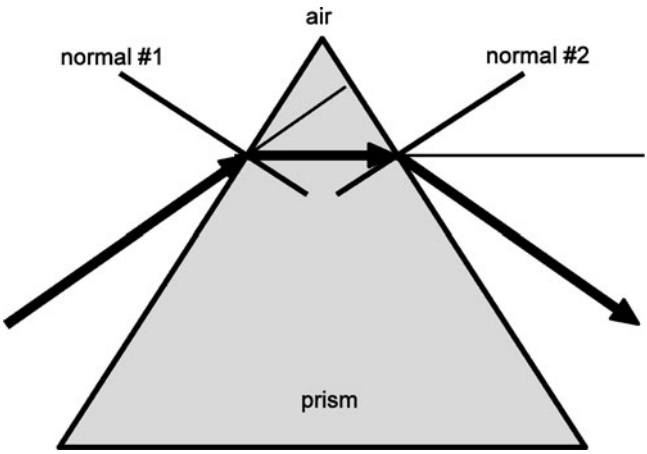


Fig. 8.26 Schematic of a prism

[8]It is often thought that the ray described above originates from the sun. A bit of analysis shows that this is impossible: In outer space $n = 1$, so that $n \cdot \sin\theta = sin\,\theta \leq 1$. At the turning point, $\theta = 90°$, so that $n \cdot \sin\theta > 1$. The only way we can have total internal reflection is for the ray to originate from light scattered by the atmosphere.

8.9 Dispersion

In the case of waves along a taut string, without stiffness, any wave pattern will propagate along the string without any change in shape, to the extent that attenuation can be neglected. Furthermore, all patterns propagate at the same velocity (for a given tension). The same holds true for sound waves and light waves in vacuum.

However, in the case of a light wave in a medium such as glass or plastic, or even air to a very small extent, **only sine waves propagate without a change in shape.** This fact is directly connected to another characteristic of such waves: The **wave velocity of a sine wave depends upon the wavelength and hence the frequency.**

For example, in the case of light propagating through *light flint glass*[9]

$$v = 1.88 \times 10^8 \,\mathrm{m/s} \;@\; \lambda = 434\,\mathrm{nm},$$

while

$$v = 1.91 \times 10^8 \,\mathrm{m/s} \;@\; \lambda = 768\,\mathrm{nm}.$$

The variation in wave velocity with wavelength is referred to as **dispersion**. A wave having this property is said to be **dispersive**. Generally the degree of variation is relatively small. Nevertheless, it is significant enough as to be responsible for the colors of the rainbow, which is produced when sunlight passes through microscopically small drops of water in the sky.

Light waves traveling in a medium and waves propagating along the surface of a liquid, such as ocean waves, are dispersive. Waves traveling along a string without stiffness, sound waves, and light waves in vacuum are **nondispersive**. Waves along a string that has stiffness are dispersive. The very significant importance of dispersion in fiber optics is discussed in Sect. 8.9.2.

8.9.1 Effect of Dispersion on a Prism

The speed of light in a medium depends upon the wavelength. Therefore, the index of refraction and the degree of refraction depend upon frequency.

In the figure, we see that for this prism, violet light is refracted more than red light. Alternatively stated, refraction decreases with increasing wavelength.

Question: How does the speed of light depend upon wavelength for this prism?

Compare this behavior with *diffraction*, which *increases* with increasing wavelength. However, we must note that the behavior of the prism depicted here is

[9]Reference: Handbook of Chemistry and Physics, 65th edition, (Chemical Rubber Comp., Boca Raton, FL, 1984).

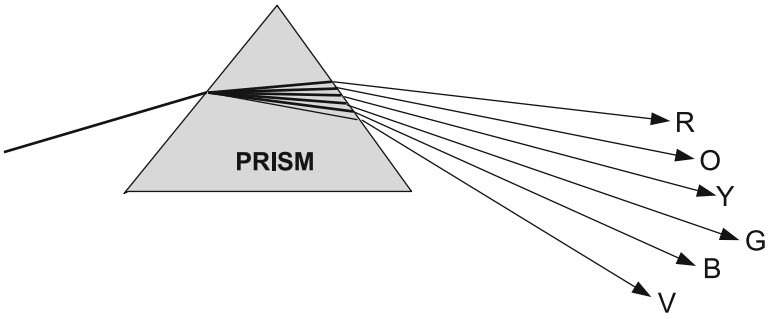


Fig. 8.27 Prism splitting colors as a result of dispersion

dependent upon the particular material that the prism is made of: I know of no physical principle that does not allow for an *increase* of refraction with increasing wavelength.

Here is a useful application of dispersion. As was pointed out in Chapter 6, THE ATOM AS A SOURCE OF LIGHT, the prism can be used to carry out a spectral analysis of a beam of light. (See Fig. 6.1.) If the incident beam is not monochromatic, each monochromatic component in the mixture will leave the prism at a different angle, as shown schematically in Fig. 8.27.

8.9.2 Effect of Dispersion on Fiber Optics Communication

Recall that the attenuation of a sine wave depends upon its frequency. Therefore, as a wave propagates in the presence of attenuation, its Fourier components attenuate at different rates. The tone quality of a sound wave depends upon the ratio of its Fourier amplitudes. As a result, not only does the intensity and overall loudness decrease as the sound wave propagates, its tone quality changes too.

The shape of a wave is determined only by the ratio of the amplitudes and the phase relations among the Fourier components. (If attenuation rates were the same for all frequencies, the *ratio* of the amplitudes would remain constant and the shape would not change.) From the above, we conclude that attenuation will change the shape of the waves.

What about the effect of dispersion? The amplitudes of the Fourier components do not change as a result of dispersion. However, since each Fourier component travels at a different speed, the peaks of each component will shift one relative to another. That is, the relative phases of the Fourier components will change as the wave propagates. As a result, ***dispersion*** *causes the shape of a wave to change.*

We are now in a position to understand why dispersion can cause a serious problem in fiber optics communication: *Analog* communication of sound converts the sound wave into an electric voltage whose variation in time is the same as the pattern of the sound wave. Fiber optics communication, on the other hand, transmits

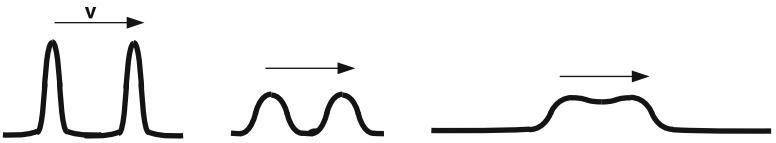


Fig. 8.28 Spreading of a pair of pulses due to dispersion

information by sending a sequence of light pulses down a glass fiber. The pattern of the sound wave is represented by a corresponding sequence of time intervals between the pulses. This system is an example of *digital* communication.

As a sequence of pulses travels down the glass fiber, the amplitudes of the pulses are attenuated due to the intrinsic properties of the glass. This is a problem that has to be dealt with in fiber optics communication. Fortunately, in Chap. 4 we noted that current attenuation is at a very low level of about 0.1 dB/km. There is a more serious problem. The glass fiber has a significant degree of dispersion. As a consequence, the shape of any non-sinusoidal wave will change its shape as it propagates along the fiber. In particular, a pulse will either become broader or become narrower. It happens that the pulses produced in fiber optics communication become broader. In time, pulses proceeding along a long fiber will become so broad as to overlap with their neighbors to such an extent that they cannot be distinguished as individual pulses. The information contained therein will become non-discernible. See Fig. 8.28.

In order to deal with this problem, special devices are inserted in sequence along the fiber. They read the broadened pulses before they are indistinguishable and replace each pulse with a re-emitted narrower pulse. The original is "reconstituted." Thus the information contained therein is preserved.

8.10 Lenses

There are two major types of **lenses**, the **converging lens** and the **diverging lens**. They are commonly used to correct a person's vision or serve as the major component in cameras, telescopes, microscopes, and film and slide projectors. The human eye itself has a lens. (See Chap. 12 for its unique characteristics.)

8.10.1 The Converging Lens

Consider what would happen if a plane wave was incident on a pair of prisms as arranged below (Fig. 8.29):

We have here represented an incident plane wave by a series of parallel rays because different parts of the wave strike different parts of the system. *Diffraction*

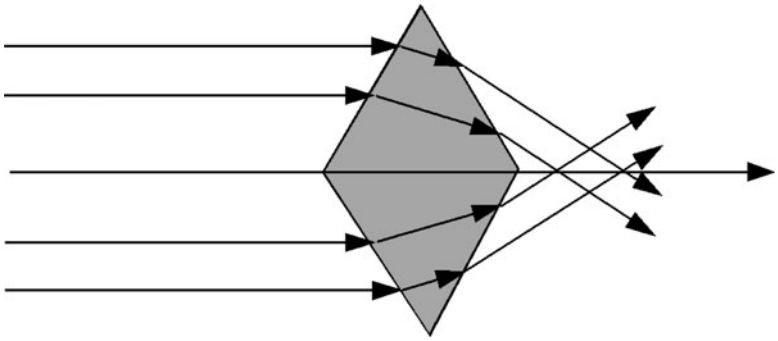


Fig. 8.29 A pseudo-converging lens from two prisms

effects are completely neglected in this section. Note how the corresponding rays exit parallel to each other.

A converging lens is a refined version of the above system, consisting of two faces which are sections of spheres of relatively large radius. This large radius corresponds to the lens being rather thin.[10]

A converging lens has the remarkable property that all incoming rays parallel to the **axis of the lens** *essentially* converge at a single point on the axis, call the **focal point**, labeled F. We see this property in Fig. 8.30.

The distance between the center of the lens and the focal point is called the **focal length**, with the symbol f. The greater the index of refraction, the more the convergence and the *smaller* the focal length.

8.10.2 Lens Aberrations

- In the previous paragraph, I stated that all rays "essentially converge to a single point." The reason is that convergence to a single point is never perfect for an actual lens. The lack of perfect convergence is called **spherical aberration**, and decreases with decreasing lens thickness. This unavoidable property of a lens is shown in Fig. 8.31, where we see what actually happens.

  The degree of aberration is much reduced for rays that are incident close to the axis, as seen in Fig. 8.32.
- Since the index of refraction depends upon the wavelength of the light – the phenomenon referred to as **dispersion** – a beam of white light cannot be focused

[10] Another term for a converging lens is a **convex lens**, since both sides of the lens are convex. The diverging lens, discussed later, is also called a **concave lens**. There also exist lenses that are concave on one side and convex on the other. If these are possibilities for consideration, one must remove any unambiguity by referring to a **biconvex lens**, or a **biconcave lens**, or a **convex-concave lens**.

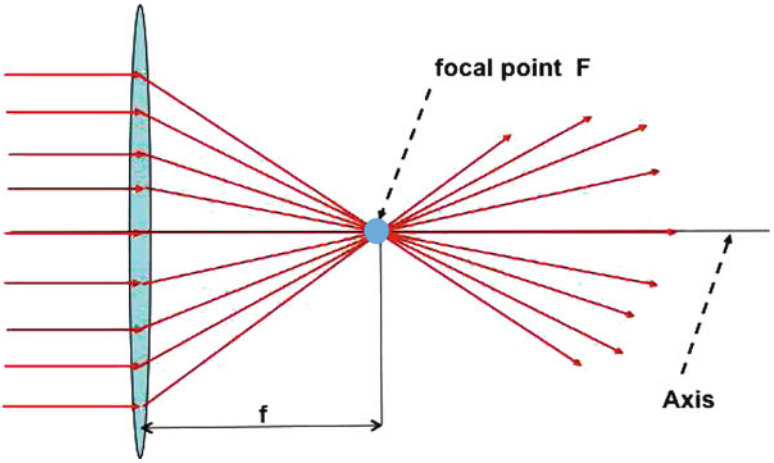


Fig. 8.30 Rays parallel to the axis meet at the focal point in an ideal lens

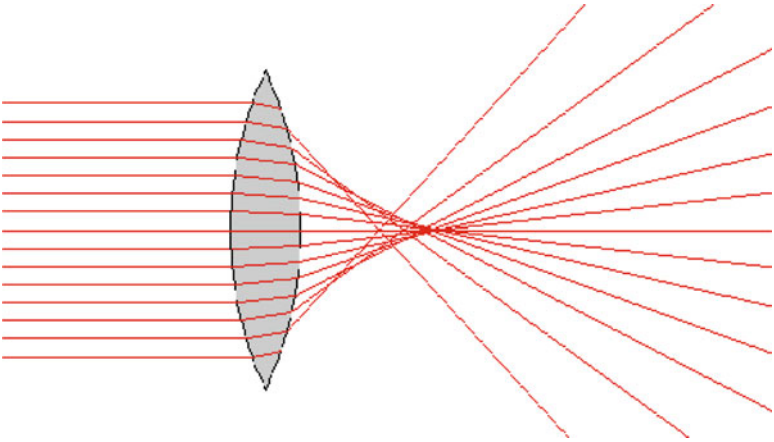


Fig. 8.31 Spherical aberration of a converging lens

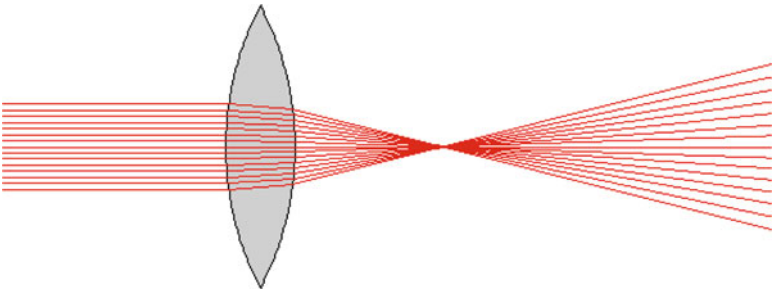


Fig. 8.32 Spherical aberration – weak for rays near the axis

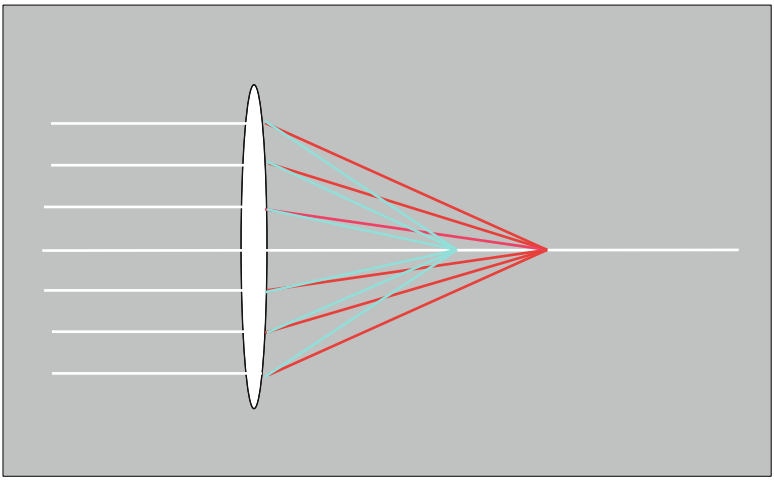


Fig. 8.33 Chromatic aberration

at a point even if **spherical aberration** was absent: The various monochromatic components of the beam will be focused at different points, as we observed in the context of the prism. We can simply say that the focal length of a monochromatic beam is dependent upon the wavelength. This defect in the lens is referred to as **chromatic aberration**. We see chromatic aberration exhibited in Fig. 8.33. The incoming beam is a mixture of two monochromatic components, one red and the other cyan. They happen to be such as to produce a white beam when mixed. They are said to be **complements**. (See Chap. 14, where color vision is discussed in detail.)[11]

> **Note**
>
> We will neglect aberration effects in the following discussion. The neglect of spherical aberration is referred to as the **thin lens approximation**.

Now suppose we have a point source S of light located to the left of a lens. The wave crests from a point source are spheres so that the wave is referred to as

[11] Polycarbonate is a material often used for eye lenses because of its strong shatter resistance and light weight. It has a drawback in having stronger chromatic aberration than glass. The so-called *Abbé* number is used as a material's level of dispersion. The larger the number, the lower the level of dispersion. Thus, while crown glass has an *Abbé* number of about 55, polycarbonate has a value of about 32. See Wikipedia (1-6-2011): http://en.wikipedia.org/wiki/Abbe_number.

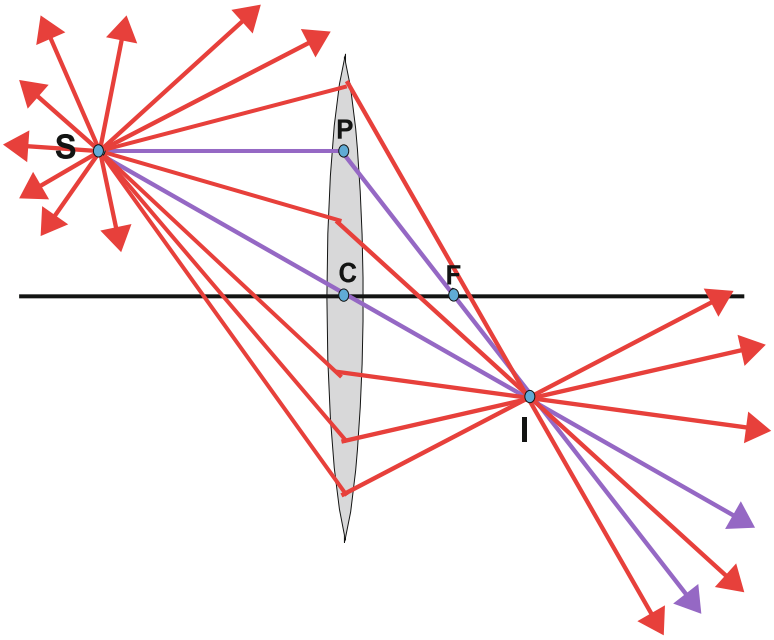


Fig. 8.34 Image of an object using a converging lens

a **spherical wave**. The wave will therefore be represented by many rays directed radially outward, as seen in Fig. 8.34. Most rays never strike the lens and go their merry way. However, a fraction pass through the lens.

The remarkable property of an ideal (aberrations neglected) lens is that all rays from S, *which pass through the lens*, cross at the same point I – called the **image point**. I is said to be the **image** of S.

Spherical aberration will blur the image. Chromatic aberration will produce a different image point for each component of monochromatic light from the source point. White light from the source point that has a continuum of all visible wavelengths will produce a line segment with a rainbow of colors.

Here is a simple way to determine the position of the image: A very important ray in lens analysis is the **parallel ray**, SP. It will be refracted so that it passes through the focal point at F. Another important ray in lens analysis is the **central ray** SC, which goes through the center of the lens, unrefracted. These two rays – shown in blue – cross at the image point I. Thus, knowing the paths of these two rays determines the position of the image point.

Note

Every source point has its own *unique* image point. Source points and image points are said to be in one-to-one correspondence. This is the fundamental property of a **thin lens.**

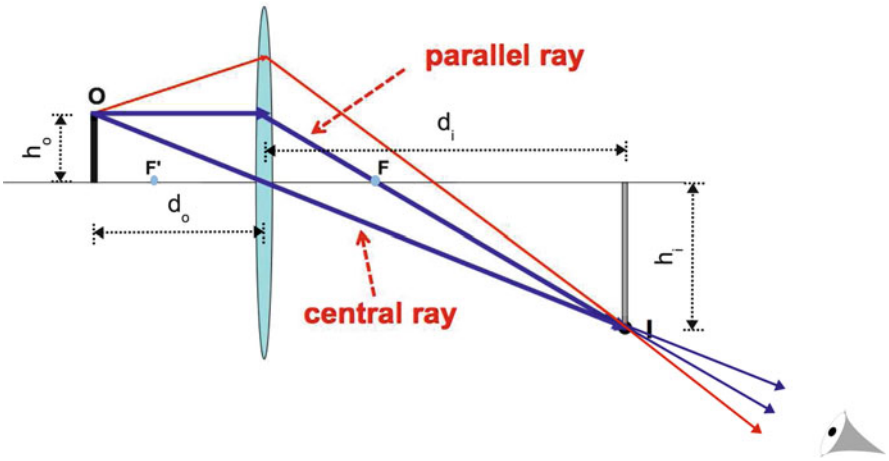


Fig. 8.35 Geometric analysis: real image of a converging lens

8.10.3 Image Produced by a Converging Lens

Now suppose that an object is located at some distance from the lens, d_0. You will note a blue point labeled F' to the left of the lens. This point is a focal length distance from the lens. It is significant that I placed the object more than a focal length distance from the lens; you will soon understand the significance of this placement. See Fig. 8.35. We call this distance the **object distance**. We assume that the object is very thin and is lined up perpendicular to the axis:

We can locate the image of the "head" of the object at O using two special rays. They are the **central ray** and the **parallel ray**. The remaining image points of the extended object produce a vertical image as shown, at the **image distance** d_i – all pairs of points being in a one-to-one correspondence. Also indicated in the figure are the height of the object h_o and the height of the image h_i.

> **Note**
>
> We assume that the lens is so thin that its thickness is negligible compared to the object distance or image distance. This unfortunately is not reflected in the figures of the text!

Using elementary trigonometry, we can derive a relation among the object distance d_o, the image distance d_i, and the focal length f. The relation is called the **thin-lens equation**:

$$\frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}} = \frac{1}{f}. \tag{8.24}$$

Note

1. If a screen is placed at the position of the "image," a clear image will appear on the screen. The image is said to be a **real image**. Any other placement of the screen will produce a blurred image.
2. Note that if someone were to be looking at the object through the lens, it would appear to be as the image shown in Fig. 8.35.
3. As the object approaches the focal point of the lens, that is, as d_o approaches f, it can easily be shown from the thin lens formula that d_i approaches infinity.

Sample Problem 8-9

Given $f = 2\,\text{cm}$ and $d_0 = 4\,\text{cm}$, find d_i.

Solution

We have $1/4 + 1/d_0 = 1/2$ so that

$$\frac{1}{d_i} = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$d_i = 4\,\text{cm}.$$

Sample Problem 8-10

Given $f = 2\,\text{cm}$ and $d_0 = 2.1\,\text{cm}$, find d_i.

Solution

We have $1/d_i = 1/2 - 1/2.1 = 0.024$ or $d_i = 1/0.024 = 42\,\text{cm}$.

Sample Problem 8-11

We next consider an object that is located less than a focal length distance from the lens: $d_o < f$. Given $f = 2\,\text{cm}$ and $d_0 = 1\,\text{cm}$, find d_i.

Solution

We have $1/d_i = 1/2 - 1/1 = -1/2$, so that $d_i = -2\,\text{cm}$.

The **negative** value for d_i means that the image is on the same side of the lens as the object.

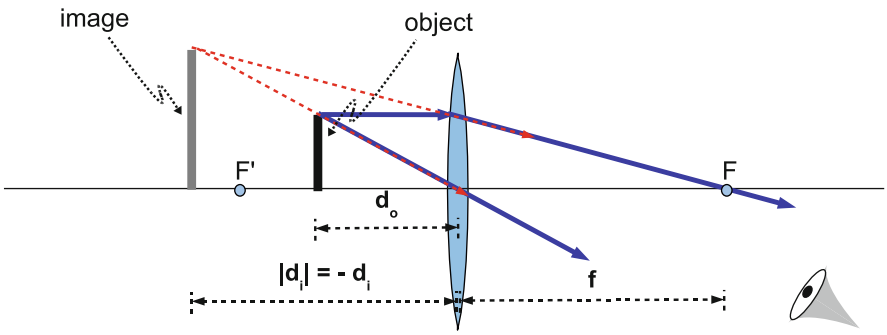


Fig. 8.36 Geometric analysis of the converging lens producing a virtual image

What is the meaning of this result? In this situation, all rays from an object point which pass through the lens *appear to be coming* from a single point behind the lens. See Fig. 8.36. The dashing of two of the line segments indicates that in fact there is no light ray along these segments. The perception by the visual system (consisting of the eye and brain) is based on the existence of these virtual rays.

The image is said to be a **virtual image**: No image will appear on a screen placed at the image position. In this arrangement, the image is larger than the object and the converging lens is serving as a **magnifying glass**.

In the first two problems, the object distance was to the left of the focal point, which results in the image being an **inverted and real image**. In the third problem, the object was between the focal point and the lens, which results in the image being an **upright and virtual image**.

8.10.4 Magnification

How does the size of the image compare to the size of the object?[12] Let us redraw Figs. 8.35 and 8.36 as simple figures to highlight the plane geometry. For both the real image [Fig. 8.37] and the virtual image [Fig. 8.38] we find the following:

Let h_o be the height of the object and h_i be the height of the image. The ratio of the two is called the **magnification**, with a symbol M. Since the two triangles $\triangle CAO$ and $\triangle CBI$ are *similar*, we have

[12]In Appendix G, we discuss **magnifying power**, which is a related though distinct property of a lens and instruments such as the telescope and microscope that consist of a series of lenses. Magnifying power represents the ability of an optical instrument to increase the image size on the retina that is produced by an object. In order to appreciate this material, it is necessary to understand how the eye works, as discussed in Chap. 12.

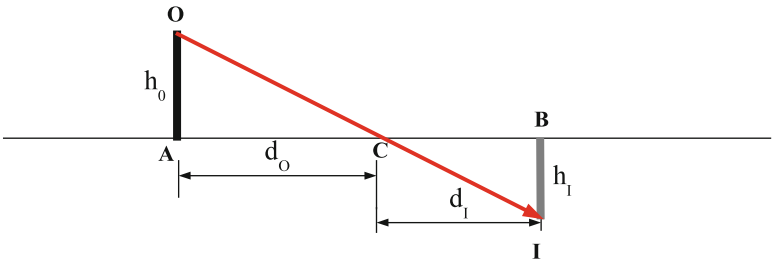


Fig. 8.37 Simplified geometric analysis: real image of a converging lens

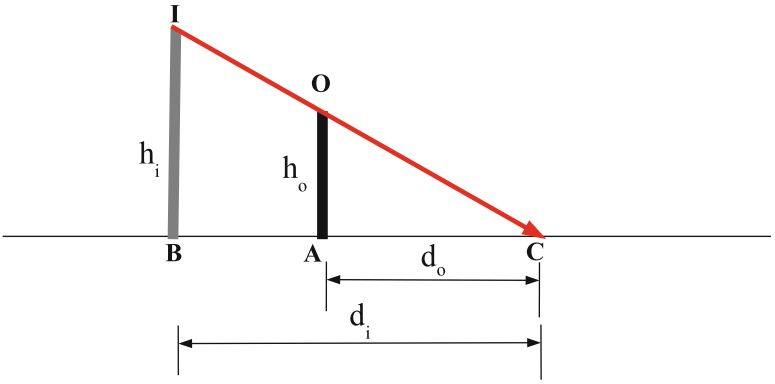


Fig. 8.38 Simplified geometric analysis: virtual image of a converging lens

$$M \equiv \frac{h_\mathrm{i}}{h_\mathrm{o}} = \left| \frac{d_\mathrm{i}}{d_\mathrm{o}} \right| . \tag{8.25}$$

The absolute value takes into account the case when the image distance d_i is negative.

Numerical examples:

For problem 1, $M = |4/4| = 1$.
For problem 2, $M = |42/2.1| = 20$.
For problem 3, $M = |-2/1| = 2$.

The Real Image of a Converging Lens as a Secondary Object

To an observer who is to the right of the real image of a converging lens, *all rays* from an object, which have gone through the lens, *appear to be coming from the real image*. For this reason, if you look through a converging lens at an object that is at a distance from the lens greater than the focal length so as to produce a real image, you will see an inverted image that appears closer to you than the lens. (There is a requirement that the rays of the image reach your eye.) This property

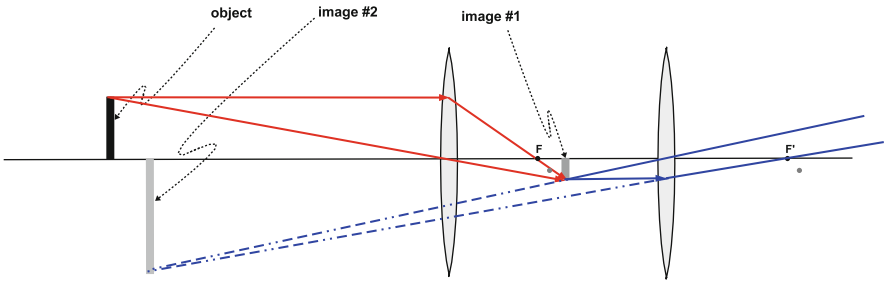


Fig. 8.39 Schematic representation of a telescope (not to scale)

of lenses is manifest for any series of lenses in sequence. Together, the set of lenses form what is referred to as a single **compound lens**. The eye itself has a number of interfaces between two media, as we will see in Chap. 12. At each interface, we have refraction. The image produced by each interface becomes the object of the next interface. Other examples of compound lenses are **microscopes** and **telescopes**.[13] The microscope is discussed in some detail in the Appendix on the **Magnifying Power of an Optical System**.

The use of a compound system of lenses in a **telescope** reveals a bit of ingenuity: We saw above that a magnifying glass, consisting of a single lens, requires that the object distance of the object be very close to the focal length. For a distant object, the magnifying glass therefore fails to magnify. However, we can place a second lens in front of the magnifying glass in a position such that the image of this second lens is at an appropriate close position to the magnifying glass. Then the magnifying glass can serve its original function and magnify the distant object. The two lenses together constitute a telescope! See Fig. 8.39. We see the original black object, upright at the far left. The first lens produces a small inverted gray image between the two lenses. The position of this image is determined by the pair of red rays. The blue rays show us how to determine the final inverted image that is produced by the second lens to the far right. The dashed segments are the continuations of these rays into a region where there are actually no light rays at all. An eye-brain situated to the right of the second lens will believe that the object is located at the final inverted image that is located a bit to the right of the object. For an actual telescope, the object is very far away from the telescope – perhaps light years away – while the images are in the vicinity of the telescope.

[13]To determine the ultimate position of the image produced by a compound lens, one must apply the thin lens equation sequentially. For the effect of each lens, one must make sure to use the distance from that lens of the image produced by the previous lens as the object distance of that current lens. In the case of eyeglasses, the distance of the eyeglasses from the eyes is so small that one usually can assume that the eyeglasses are coincident with the center of the compound lenses of the eyes.

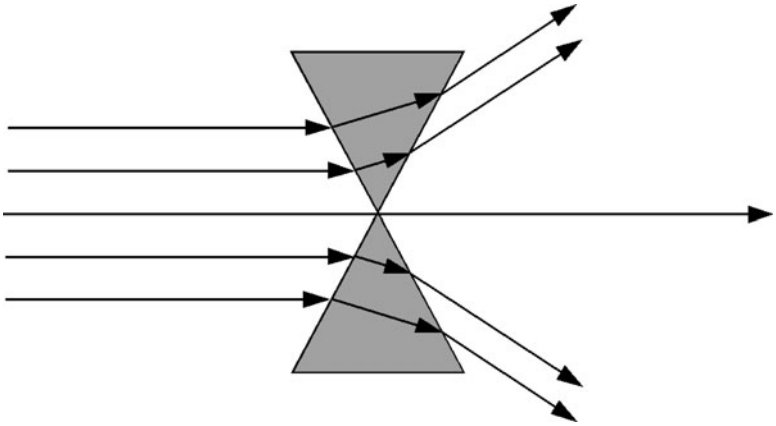


Fig. 8.40 A crude diverging lens using prisms

8.10.5 Reversibility of Rays: Interchange of Object and Image

If we examine the thin lens formula (8.24), we note a **symmetry** in the two distances, the object distance and the image distance: For a given focal length, if the object has the position of the image, the image must have the original position of the object. Alternatively, if a ray leaves the object and passes through the lens, it must pass through the image position. Conversely, if this ray is reversed in direction at the image position, it will pass through the lens and then pass through the object position. This behavior is referred to as **reversibility**. Note how reversibility is manifest for the single refraction of a ray at an interface between two materials owing to the symmetry in Snell's Law.[14]

8.10.6 The Diverging Lens

The diverging lens is a refinement of the following arrangement of two prisms (Fig. 8.40):

[14]Reversibility is manifest in the orbit of a planet about the Sun: If a planet was stopped dead in its tracks and its path reversed so that at that point the original direction is reversed while the speed is the same, the planet would retrace its path into the past, where it came from. What we would observe could be seen by taking a movie of the planet's motion and then running the movie backward. Reversibility is manifest in the basic laws of physics. The consequence is that every sequence of events has a possibility of occurring. Yet, there are movie scenes that are hilarious if they are run backward. Why? Because the reversed sequence is regarded as impossible. (Imagine someone shown jumping off a ladder onto the ground.... Now reverse the sequence.) Such sequences are referred to as being **irreversible**. One of the challenges of physics is to understand how such extremely unlikely, irreversible sequences are never seen and yet have a possibility of occurring.

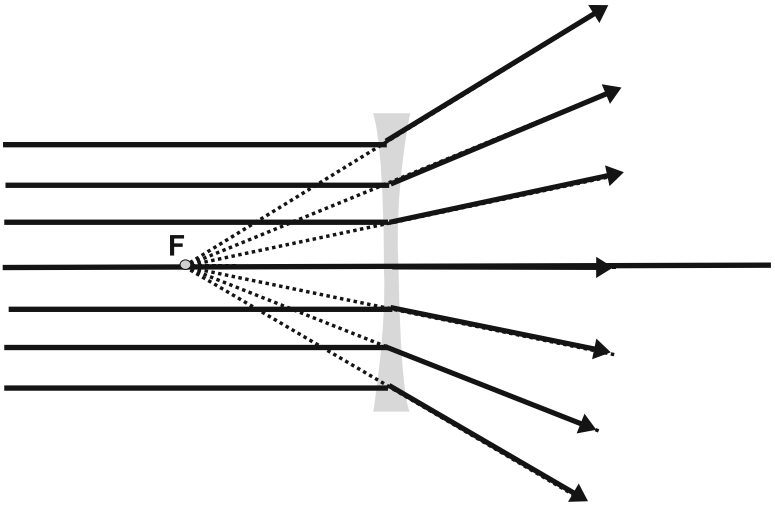


Fig. 8.41 The focal point of a diverging lens

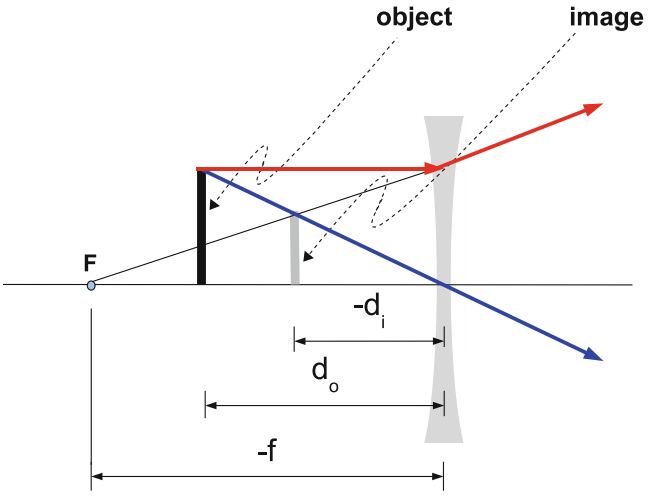


Fig. 8.42 Geometric analysis: the virtual image of a diverging lens

Look at Fig. 8.41. Note that incident *parallel* rays emerge so as to appear to be coming from a common point, the focal point F of the lens. Note that now the focal point is on the same side of the lens as the source of light – to the left of the lens.

Next, in Fig. 8.42, we display the image of an object as produced by a diverging lens. Images are **always erect and virtual** and appear on the same side as the object, both between the lens and the focal point.

The thin lens equation can still be used, with the focal length f in the equation set equal to a **negative** number! The image distance parameter d_i obtained is always negative, indicating that the image is on the same side as the object, to the left of the lens. Furthermore, $|d_\mathrm{i}| < d_\mathrm{o}$ always. This indicates that the magnification M is always less than unity. **Thus, the image is always smaller than the object**.

Sample Problem 8-12

Given a diverging lens with a focal length of $-2\,\mathrm{cm}$, and an object distance, $d_0 = 3\,\mathrm{cm}$, find the image distance and the magnification.

Solution

We have

$$\begin{aligned}\frac{1}{d_\mathrm{i}} &= \frac{1}{f} - \frac{1}{d_\mathrm{o}} \\ &= \frac{1}{-2} - \frac{1}{3} = -\frac{1}{2} - \frac{1}{3} = -\frac{5}{6}.\end{aligned}$$

so that $d_\mathrm{i} = -6/5\,\mathrm{cm}$.

The magnification is given by $M = |d_\mathrm{i}/d_\mathrm{o}| = (6/5)/3 = 2/5 = 0.4$.

8.10.7 Determining the Focal Length of a Diverging Lens

We can determine the focal length of a converging lens by measuring the distance from the lens to the image of a far off source or by comparing the image distance with the object distance. Unfortunately, a diverging lens produces no **real image**, so that it is not immediately clear how its focal length can be determined. We will outline a method of doing so that uses a second, converging lens in tandem with the diverging lens, as shown in the complex Fig. 8.42:

The position of virtual image (#1) is obtained as follows: We draw the red ray from the top of the object parallel to the axis. The second red ray is determined by connecting its origin at the top of the diverging lens back toward the focal point at F. The first blue ray passes through the center of the diverging lens; its intersection with the second red ray continued to the focal point F determines the position of the top of image (#1).

This first, virtual, image serves as an effective object for the converging lens. We find the position of the ultimate image (#2) as follows: We draw the black parallel ray from the top of image (#1) to the top of the converging lens. This ray is continued through the focal point F' of the converging lens. We draw a second central ray from the top of image (#2); its intersection with the previous ray determines the position of the bottom of the final image (#2). The figure shows a third red ray from the

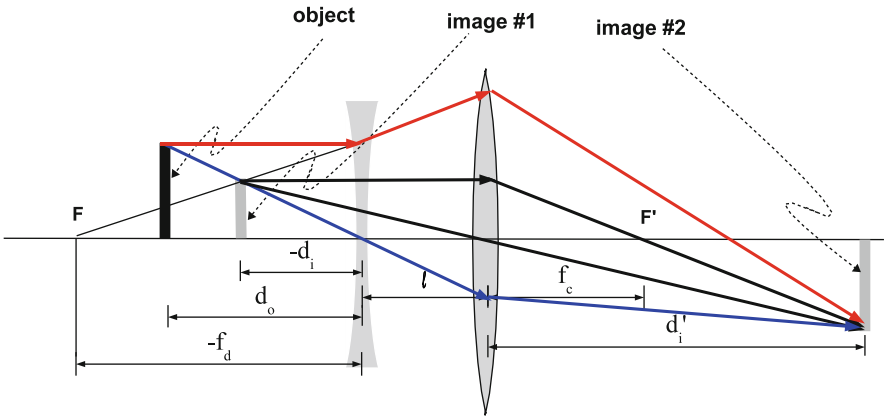


Fig. 8.43 Setup for determining the focal length of a diverging lens

converging lens to image (#2). The three red rays together form one actual path that starts from the object and ends up at the ultimate image. Another actual path is represented by the two blue rays. *Can you determine a third actual path?*

These facts are reflected in Fig. 8.43. We use $(-f_\mathrm{d})$ and $(-d_\mathrm{i})$ for the distances in the figure and (8.26) below because both are negative.

We can measure the parameters d_0, ℓ, f_c, and d_i' that are shown in the figure. The object distance of the converging lens is given by $(-d_\mathrm{I} + \ell)$, so that

$$\frac{1}{-d_\mathrm{i} + \ell} = \frac{1}{f_\mathrm{c}} - \frac{1}{d_\mathrm{i}'}. \tag{8.26}$$

We must solve this equation for the parameter d_i. Next, since we now know d_o and d_i, f_d can be determined from the lens equation applied to the diverging lens:

$$\frac{1}{f_\mathrm{d}} = \frac{1}{d_\mathrm{o}} + \frac{1}{d_\mathrm{i}}. \tag{8.27}$$

8.11 The Doppler Effect

Have you ever paid attention to the sound of a car racing past you while you stand at the side of a highway? The whirring sound is noisy; nevertheless one can discern a distinct pitch. This pitch steadily decreases as the car approaches and then recedes away. This effect is due to the **Doppler effect**. The Doppler effect occurs whenever there is relative motion between a source of waves and an observer of the waves, whatever the type of wave.

If the source emits a sine wave of frequency f, the observed frequency f' differs from the source frequency f. The ratio f'/f depends upon the motion of both the source and the observer as well as their relative positions. We will discuss two simple cases in detail:

1. Source and observer approaching head-on
2. Source and observer receding directly away from each other

The Doppler effect of waves of a disturbed medium, such as sound waves or waves on the surface of a liquid, is distinctly different from the Doppler effect of EM waves, which can propagate through a vacuum, without the presence of a medium.

It happens that in case the relative velocity u of the source and observer is small compared to the wave velocity v, the equations for the Doppler effect are approximately the same for both sound waves and EM waves in vacuum. If we let $\Delta f \equiv f' - f$ = change in frequency, we have

$$\frac{\Delta f}{f} \approx \frac{u}{v}. \tag{8.28}$$

8.11.1 Doppler Effect for Waves in a Medium

Consider the Doppler effect of sound waves in air. The source and/or the observer could be moving with respect to the air. We will see below that the observed frequency depends upon the motion of the source and the observer with respect to the medium. Since there is no medium for the propagation of light, the formulas below are not relevant for light. We will discuss two simple cases.

Case (i): The source Is at Rest, While the Observer Is Moving with Respect to the Medium

A point source of sine waves will emit a wave with wave crests that are concentric spheres, as shown in Fig. 8.44. The distance between neighboring spheres is equal to the wavelength .

The wave crests are traveling at a speed v with respect to the air. They therefore approach the observer with a speed

$$v' = v + u. \tag{8.29}$$

Therefore, the observed frequency is given by

$$\begin{aligned} f' &= \frac{v'}{\lambda} = \frac{v+u}{\lambda} \\ &= \frac{v}{\lambda}\left(1 + \frac{u}{v}\right). \end{aligned} \tag{8.30}$$

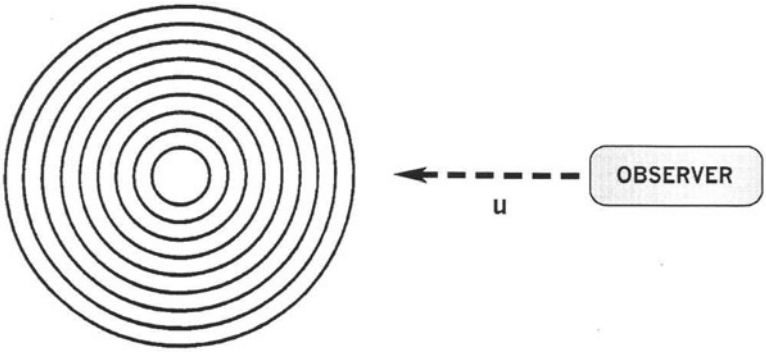


Fig. 8.44 A moving observer in the presence of a stationary source of sound waves

Since $v/\lambda = f$, we have

$$f' = f\left(1 + \frac{u}{v}\right). \tag{8.31}$$

If the observer is moving away from the source, the resulting formula can be obtained simply by replacing $+u$ by $-u$ above:

$$v' = v - u$$

$$f' = f\left(1 - \frac{u}{v}\right). \tag{8.32}$$

Example 8-1

$$f = 1{,}000\,\text{Hz}, \qquad u = 34.0\,\text{m/s}, \qquad v = 340\,\text{m/s}.$$

We have $u/v = 0.1$.

Then if motion is toward:

$$f' = 1{,}000(1 + 0.1) = 1{,}100\,\text{Hz}.$$

If motion is away:

$$f' = 1{,}000(1 - 0.1) = 900\,\text{Hz}.$$

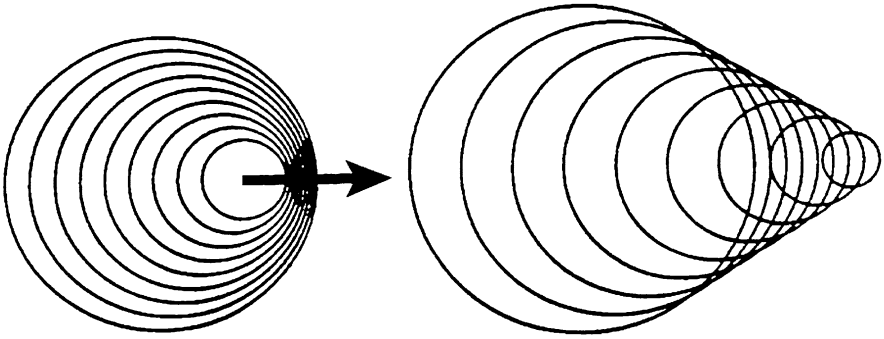


Fig. 8.45 The waves from a source moving at two different speeds

Case (ii): The Source Is Moving, While the Observer Is at Rest

Crests travel at the wave velocity with respect to the medium. However, because of the motion of the source, the wave crests are not concentric spheres. In the direction of motion of the source, the wavelength is decreased, while behind the source, the wavelength is increased. In Fig. 8.45, we see two sources, one (a) moving at a speed $u < v$, the other (b) at a speed $u > v$.

Since the observer is at rest with respect to the medium, the crests move at a speed v with respect to the observer. As a result, we can show that

$$\lambda' = \lambda\left(1 - \frac{u}{v}\right) \text{ and } f' = \frac{v}{\lambda'}$$

$$\text{so that } f' = \frac{f}{1 - \frac{u}{v}}. \tag{8.33}$$

By letting $u \Rightarrow -u$, we obtain

$$f' = \frac{f}{1 + \frac{u}{v}}. \tag{8.34}$$

Example 8-2

Given $f = 1{,}000\,\text{Hz}$, $u = 34.0\,\text{m/s}$, and $v = 340.0\text{m/s}$, find f' for the source moving away and coming toward you.

We have $u/v = 0.100$, so that

- If toward: $f' = 1{,}000/(1 - 0.1) = 111\,\text{Hz}$
- If away: $f' = 1{,}000/(1 + 0.1) = 909\,\text{Hz}$

Notes:

1. It is interesting to compare the above numerical results. We have $u = 0.1v$, so that speed $u \sim v$. We can see that the frequencies f' do depend upon which of the two, the source or the observer, are moving with respect to the medium.
2. There is a **mathematical catastrophe** in the formula (8.33) for the case when the source is moving toward the observer at a speed which exceeds the wave velocity. In this case, we see that the frequency is negative and so is meaningless. (Moreover, when the source is moving at the wave velocity itself, the expression for the frequency diverges ($f \Rightarrow \infty$)!) What happens, in fact, is that there is a **shock wave** and no wave is perceived! Clearly, there cannot be a wave in front of a source if the source is moving faster than the wave would move. This situation is shown in Fig. 8.45b.

There are numerous interesting phenomena connected with shock waves. We will discuss a few below.

1. When a jet plane accelerates through the speed of sound (known as **Mach one**), thus breaking the sound barrier, the production of the shock wave is accompanied by a loud sound. This sound was regarded as a great nuisance when commercial jet planes were first introduced.
2. The triangular trailing pattern of a boat moving along the surface of a lake is a beautiful example of a shock wave. One can sometimes observe swarms of minuscule, miniature insects swimming in a jagged manner along the surface of a pond leaving a feathery, impressionistic pattern of waves. Without these miniature shock waves, the insects would barely be noticeable!
3. The production of a shock wave is the principle behind the **whip**. A whip is made with a long piece of leather whose diameter is tapered. In cracking a whip, a pulse is sent down the length of the whip from the thicker end. Now recall that the wave velocity along a taut string increases with decreasing linear mass density. (The process is complicated by a change in the tension too.) As the pulse proceeds, the wave velocity increases. By the time the pulse reached the tip of the whip, the wave velocity exceeds the speed of sound in the air. The resulting shock wave is responsible for the "crack" of the whip![15]
4. When a charged particle, such as an electron, is moving through a medium at a speed faster than the speed of light in that medium, light is given off within the conical trail of a shock wave of electromagnetic radiation. This light is called **Čerenkov radiation**.[16] See Fig. 8.46, wherein we see the blue Čerenkov radiation from charged fundamental particles moving in a nuclear reactor.

[15] See http://www.hypography.com/article.cfm?id=32479 for a summary of recent research on the cracking of a whip. Also, see the following website for a discussion of how shock waves are the clue behind the trick for cracking a piece of wood with one's bare hand: http://www.worldkungfu.com/whip.html#WHIPS.

[16] The letter Č is pronounced like "ch" in "cheer."

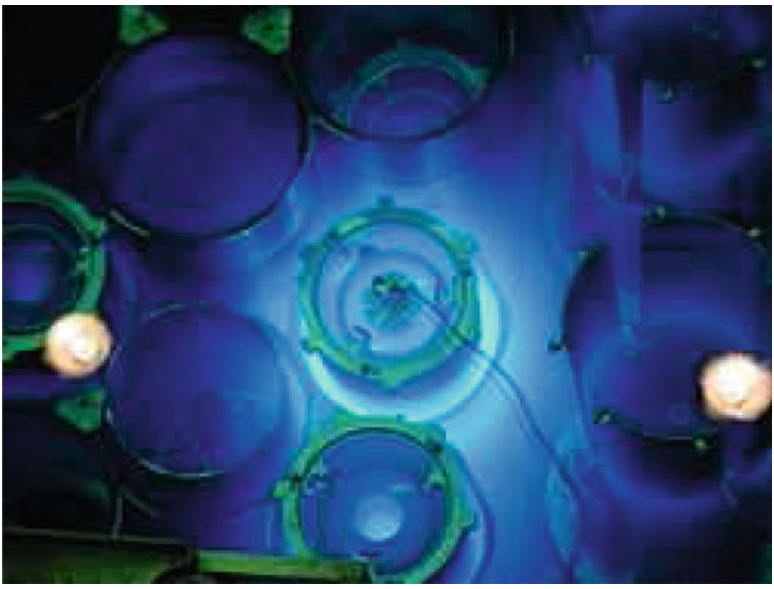


Fig. 8.46 *Blue* Čerenkov light from a high flux isotope reactor (source: http://en.wikipedia.org/wiki/Cherenkov_radiation)

8.11.2 Doppler Effect for Electromagnetic Waves in Vacuum

Cases (i) and (ii) above cannot have meaning here because there is no medium. Only the **relative velocity** of the source and observer can matter. The correct formulas are obtained using **Einstein's Theory of Special Relativity**. They are (remember that $v = c$):

$$f' = \sqrt{\frac{1 \pm \frac{u}{c}}{1 \mp \frac{u}{c}}}. \tag{8.35}$$

We use the upper signs if the source and observer are moving toward each other; we use the lower signs if the two are moving away from each other. The examples will make this point clear.

Example 8-3

$f = 1{,}000\,\text{Hz}$, $u = 3.00 \times 10^7\,\text{m/s}$, so that $u/c = 0.100$.

If away:

$$f' = 1{,}000\sqrt{\frac{1 + 0.100}{1 - 0.100}} = 1{,}106\,\text{Hz}.$$

If toward:

$$f' = 1{,}000\sqrt{\frac{1-0.100}{1+0.100}} = 905\,\text{Hz}.$$

Note

Suppose that we compare the results for EM waves with those for sound waves that neglect the effects of special relativity. We see that (8.35) is the square root of the product (hence the *geometric mean*) of the corresponding two expressions that hold for a medium, one when the observer alone is moving (8.31) and one when the source alone is moving (8.33). A similar relation holds when the source and observer are moving away from each other. The geometric mean of two numbers always lies between the two numbers.

Note

For small relative velocities ($u/v \ll 1$), the expression for f' is approximately the same whether we have a moving source or moving observer with sound waves or EM waves in a vacuum.

8.11.3 Applications of the Doppler Effect

In the problems at the end of the chapter, the reader will be shown how the Doppler effect can be used:

1. Have you ever wondered how police can determine the speed of vehicles using radar? See Problem 8.20 to learn how the Doppler effect and beats are involved.
2. Recall that we mentioned that astronomers and cosmologists have discovered that the Universe is expanding. How can they know this? They make use of the well-known spectra of atoms, whose wavelengths are known upward of eight significant figures. The determination of the velocity of a star or galaxy with respect to the earth can made by measuring the shift in its atomic spectra. If a star is moving away from the earth, the frequency is lowered, so that colors change toward the red end of the visible spectrum. We have what is referred to as a **redshift**. An example is shown in the figures below from Palomar Observatory. In Fig. 8.47, we see the quasi-stellar radio source (quasar) 3C273.[17] It appears as a large bright star. In Fig. 8.48, the light spectrum from the quasar is shown above the spectrum of a stationary source of hydrogen and helium. The spectral

[17]This quasar is estimated as having a mass equal to about one-billion solar masses.

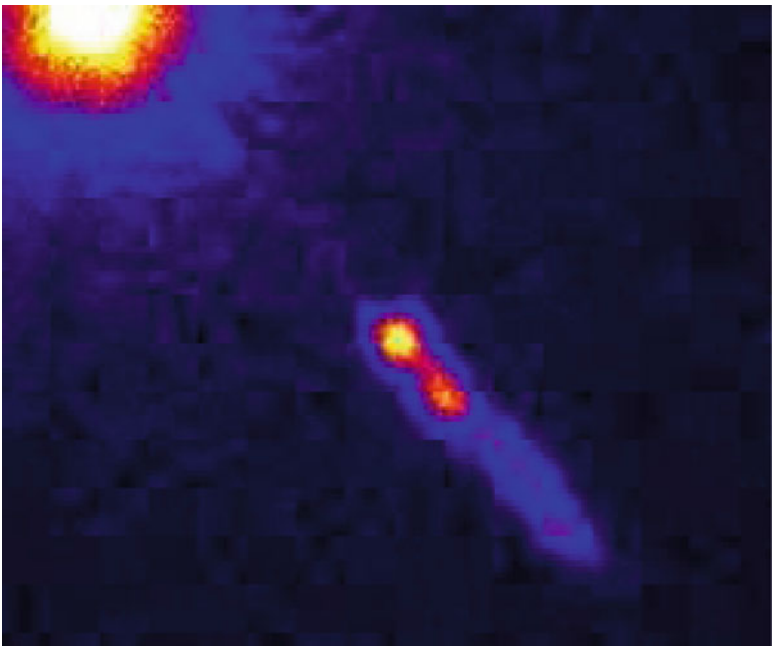


Fig. 8.47 Quasar 3C273 (source: http://en.wikipedia.org/wiki/3C_273)

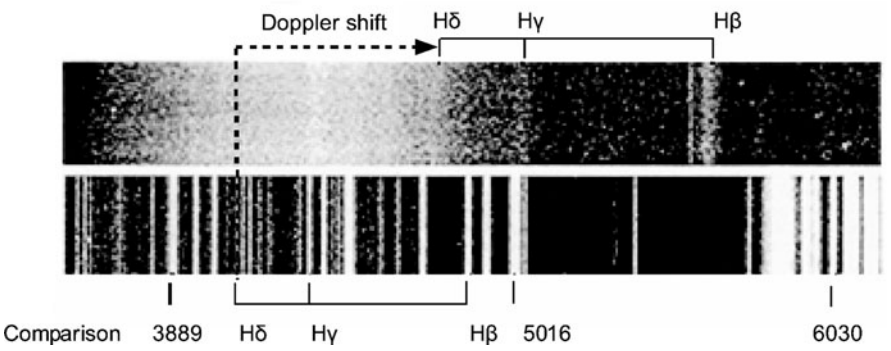


Fig. 8.48 Spectrum of the quasar (source: http://chandra-ed.harvard.edu/3c273/quasars.html)

lines labeled H_α, H_β, and H_γ are seen to be shifted to the right in the figure, corresponding to larger wavelengths. Calculations indicate that the quasar is moving away from the earth at about one-seventh the speed of light.

Note that in addition to a redshift due to a Doppler effect, there is also a **gravitational redshift**, which is akin to the slowing down of an object when we throw it up into the air.

8.12 Polarized Light**

In this section, we will be discussing the nature of **polarized light**. It is of importance not only for fundamental physics reasons and practical applications but also because some artists make use of polarized light in their art creations. We have mentioned that a simple EM wave is *transverse*, with a direction of displacement that is determined by the direction of oscillation of the electric field. In Fig. 8.49, we see both the electric field and magnetic field as they vary in space at some moment in time. The wave is propagating in the direction **k**.

Light that is emitted from typical light sources consists of a mixture of waves having a random distribution of electric field orientations. We say that the light is **unpolarized** and will indicate this state by the symbol shown in Fig. 8.50.

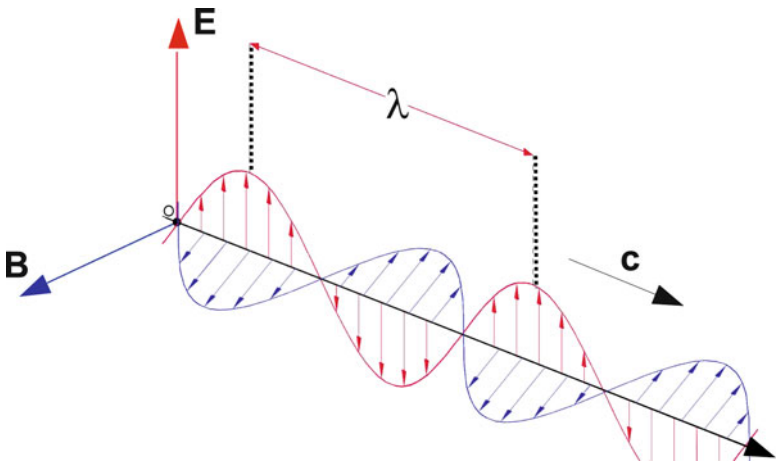


Fig. 8.49 Light as an electromagnetic wave (source: http://en.wikipedia.org/wiki/Wave#Electromagnetic_waves)

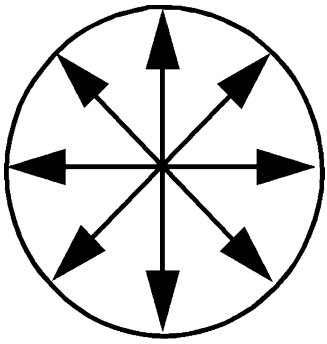


Fig. 8.50 Symbol for unpolarized light

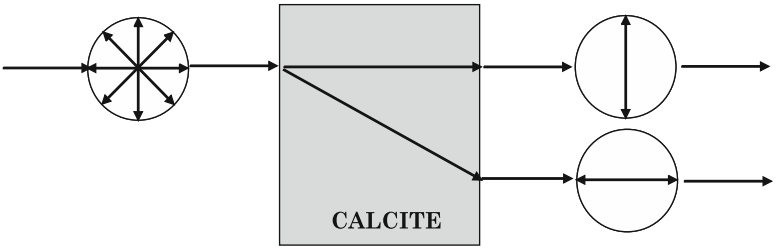


Fig. 8.51 Calcite-based polarizer

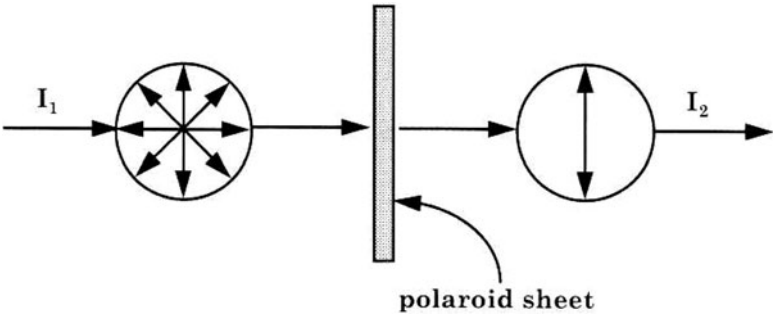


Fig. 8.52 Polaroid-based polarizer

8.12.1 How Can We Obtain a Beam of Polarized Light?

There are numerous ways to obtain a beam of polarized light. Laser light is polarized. So is the light from excited atoms whose quantum states have been properly selected. Below are two ways that are easily produced in a classroom.

The first material is **Calcite**. If an unpolarized beam of light is passed through a single crystal of calcite which is properly cut, we get two, physically separated, polarized beams, polarized perpendicular to each other (see Fig. 8.51).

Each component in the unpolarized incoming beam is decomposed by the calcite into two specific *mutually perpendicular* polarized components. The directions of these components depend upon the orientation of the crystal; they are indicated by the circle symbols on each of the two outgoing beams. As a consequence, if we view a single dot through the above calcite crystal, we will see two dots, with the light from one dot polarized horizontally and the other vertically. The property of a material that enables it to split a beam into two polarized components is called **birefringence**.

Second, **Polaroid** is a plastic material, discovered by Edwin Land, which can produce a polarized beam from an unpolarized one by absorbing a component of polarization, as shown in Fig. 8.52:

In contrast to the calcite, there is a single outgoing beam whose direction of polarization depends upon the orientation of the sheet of Polaroid. **Ideally, the**

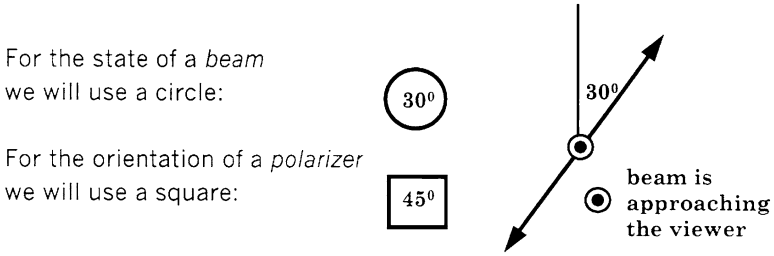


Fig. 8.53 30° Polarizer

outgoing beam would have one-half the intensity of an incoming unpolarized beam: We would write $I_2 = I_1/2$. Polaroid is *not* ideal. However,

in the discussion that follows,
we will assume that the polarizers are ideal.

8.12.2 Series of Polarizers

A light beam can pass through a series of polarizers. Below, we will consider a number of examples that indicate how polarizers affect light beams.

It will be useful to use the following notation:

The **direction of the axis of polarization** of both the polarizers and the beams will be specified by its angle with respect to the vertical, *as observed with the beam coming **toward** the viewer*. The angle will be indicated using the following symbols:

For the state of a *beam* we will use a circle; for the orientation of a *polarizer* we will use a square. The choice of how to specify the angle is indicated in Fig. 8.53.

> **Note**
>
> If the angle was specified with respect to a viewer that observes the beam in the direction in which the beam is traveling, the sign of the angle must be reversed. Thus 30° must be replaced by −30°.[18]

We will let I_0 be the incident intensity and I_1, I_2, ... be the intensities at various subsequent stages.

In Fig. 8.54, we depict what happens to an unpolarized beam of intensity I_0 after passing through two sequential polarizers set at angles 0° and 45°, respectively:

We note that an angle of 45° between the beam and the polarizer axes leads to a reduction in beam intensity of one-half. What about other angles?

[18]To avoid errors, it is essential that you compare the sign convention used by a book or an article on polarization to the one we use in this text.

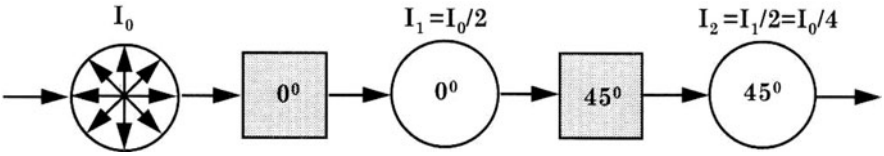


Fig. 8.54 Two polarizers in series

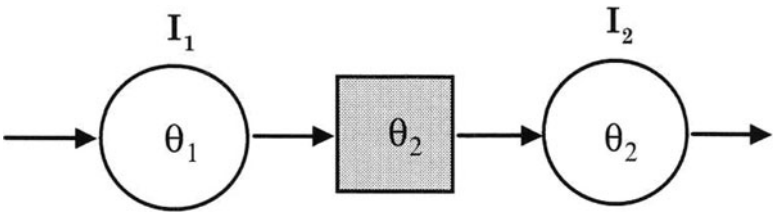


Fig. 8.55 General series of two polarizers

Here are two rules you need to know about the action of **ideal polarizers**:

1. **Whatever the state of the incoming beam, the outgoing beam has the State of Polarization of the Polarizer** (Fig. 8.55).
2. The following general relation between the incoming intensity I_1 and the outgoing intensity I_2:

$$I_2 = I_1 \cos^2(\theta_2 - \theta_1). \tag{8.36}$$

This equation is referred to as **Malus' Law**.

In the applications that follow, we will use the following *exact* values for the cosine so that, for example, we do not end up with a figure like 0.49 when the exact answer, $1/2$, is significant.

$$\cos 45^\circ = \frac{1}{\sqrt{2}}, \qquad \text{so that} \cos^2 45^\circ = \frac{1}{2}$$

$$\cos 30^\circ = \frac{\sqrt{3}}{2}, \qquad \text{so that} \cos^2 30^\circ = \frac{3}{4}$$

$$\cos 60^\circ = \frac{1}{2}, \qquad \text{so that} \cos^2 60^\circ = \frac{1}{4}.$$

8.12.3 Ideal vs. Real Polarizers

Polaroid does not behave as an ideal polarizer. Some light is reflected off a piece of real polarizer material. In addition, real polarizers absorb some light. In Fig. 8.56, we depict the **transmittance** (the fraction of the incident intensity that is transmitted)

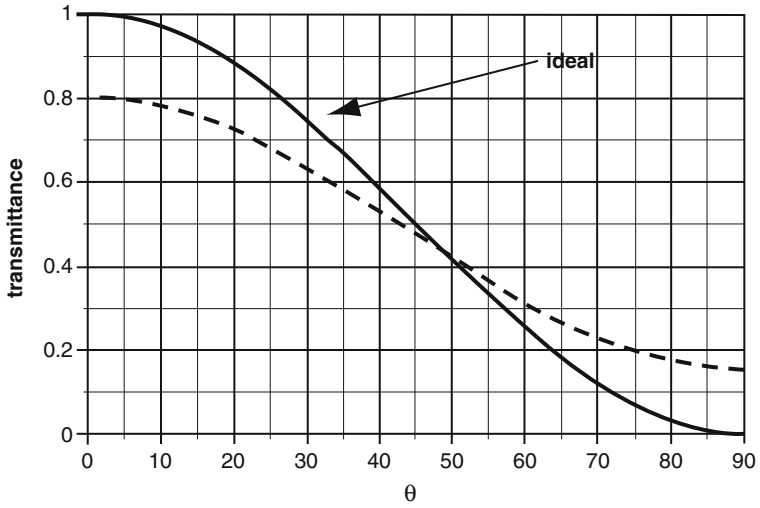


Fig. 8.56 Real vs. ideal polarizer

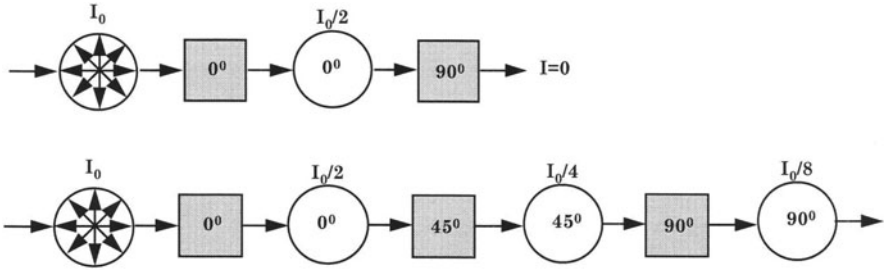


Fig. 8.57 Polarizer Example 1

vs. the relative angle θ, for both the ideal polarizer and a typical real polarizer. We note that **transmittance** is not unity when the angle is zero. Thus, the polarizer absorbs energy even for the component of the incoming beam that is parallel to the axis of polarization. Also, the outgoing beam is not zero for an angle of 90°. This fact indicates that the outgoing beam is generally **incompletely polarized**.

8.12.4 Sample Problems

We now turn to some examples to illustrate how we apply the two basic rules.

Example 1 In the situations shown in Fig. 8.57, the intensities of the beams are indicated in the figures themselves since they are relatively simple to determine from our discussion above.

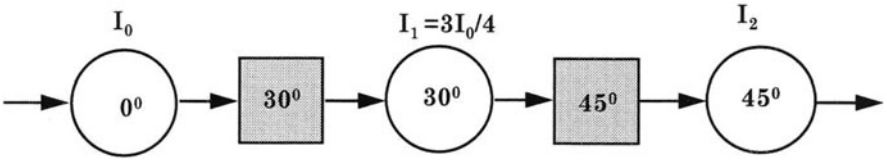


Fig. 8.58 Series of polarizers for Example 2

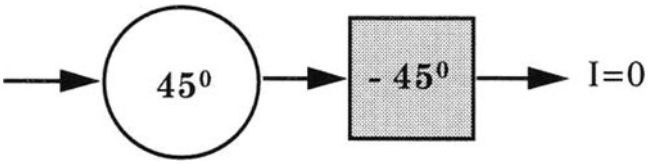


Fig. 8.59 Setup for Example 3

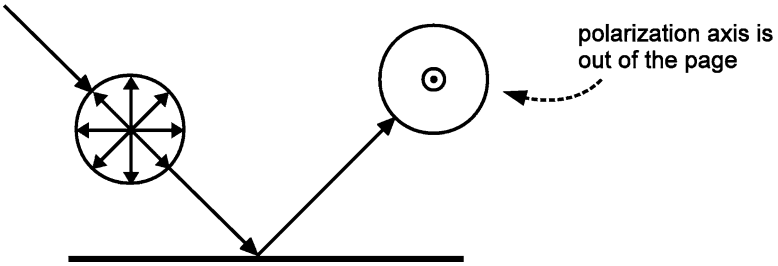


Fig. 8.60 Polarization of reflected light

Example 2 In Fig. 8.58

$$I_1 = I_0 \cos^2 30^\circ = \frac{3}{4} I_0$$

and

$$I_2 = I_1 \cos^2(45^\circ - 15^\circ) = \frac{3}{4} I_0 \times 0.93 = 0.70 I_0.$$

Example 3 In Fig. 8.59, $\theta_1 - \theta_2 = 90^\circ$, so that the outgoing intensity is zero.

8.12.5 Partial Polarization of Reflected Light

Light that is reflected off a surface is **partially polarized** along an axis that is *perpendicular to the plane determined by the incident and reflected rays*, as shown in the Fig. 8.60.

This phenomenon can be used to determine the axis of polarization of an isolated polarizer, as follows: Look at an unpolarized beam of light that is reflected off a

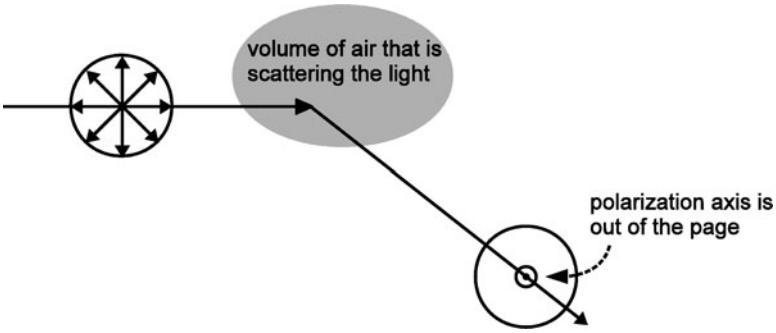


Fig. 8.61 Polarization of scattered light

flat surface (*not* a mirror!) through the polarizer, as in the above figure. Rotate the polarizer until the image on the surface is brightest. Then the axis of polarization of the polarizer is currently horizontal. The phenomenon also demonstrates that the polarizer material is not isotropic as one would at first assume.

We have here an explanation for why Polaroid sunglasses are so useful for cutting down the glare of sunlight reflecting off lake and ocean surfaces. By looking at a shiny floor through a piece of Polaroid, one can see the varying degrees of shininess as one rotates the Polaroid.

8.12.6 The Polarization of Scattering Light

We pointed out in Sect. 8.1 that the atmosphere scatters light preferentially toward higher frequencies; as a result the sky is blue. Another interesting property of scattered light is that *it is partially polarized along an axis which is perpendicular to the plane determined by the incoming and scattered rays*. The polarization is a result of the variation of the intensity of the scattered waves with incident polarization, being a maximum for a polarization that is perpendicular to the page. The geometry is depicted in the Fig. 8.61. (Compare the geometry here with that of partially polarized reflected light.)

Can you see why Polaroid sunglasses are not useful for cutting down the glare of a bright sky in all situations?

8.12.7 The Polarizer Eyes of Bees

The eyes of bees each have a circular array of eight polarizers whose axes are oriented at angles $360°/8 = 45°$ apart, as shown in the figure below. As a result, there are varied degrees of transmission of the polarized sunlight from the sky. From the intensity pattern of the polarizers, the bee is able to determine the orientation of its body with respect to the sun and hence its beehive! A bee has a built-in analog for

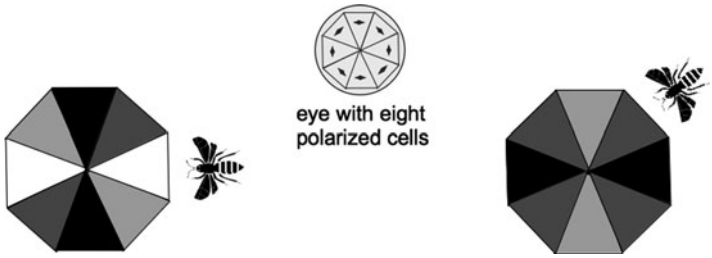


Fig. 8.62 The Frisch Experiment of the polarizing eyes of bees

a magnetic compass. This deduction is based on experiments pioneered by Karl von Frisch, wherein he changed the polarization of the light seen by bees and observed the bees changing their direction of flight in response. Figure 8.62 is a schematic of what the bees eyes see in response to a change in their orientation. The figure is based upon the discussion of Karl von Frisch in his book, *The Dance Language and Orientation of Bees*, Belknap Press of Harvard University Press, Cambridge, MA, 1967.

8.12.8 Using Polarization of EM Radiation in the Study of the Big Bang

According to cosmologists, our Universe evolved with a **Big Bang** from an extremely dense concentration of energy about ten billion years ago. It has been expanding ever since. A mere 400,000 years after the start of the Big Bang, the radiation that filled the entire Universe became decoupled from the matter in the Universe as Black Body radiation It thus contains a record of the situation at the time of decoupling. Currently, the radiation is at a temperature of 3 K. This "3 K" radiation has experienced much scattering from one region to another and is therefore polarized. The variation of temperature and direction of polarization of the universe is exhibited in Fig. 8.63, taken from the website http://map.gsfc.nasa.gov/m_mm.html. Colors indicate "warmer" (red) and "cooler" (blue) spots. The white bars show the "polarization" direction of the oldest light.

Will the currently expanding Universe expand forever or will it eventually reach a maximum expansion and then collapse? This is probably the most important question not yet resolved (in the year 2008). Detailed information such as is provided by the above figure will help answer this question.

8.12.9 Optical Activity

Certain materials have the remarkable property that when a polarized light beam passes through the material, the axis of polarization of the beam is *rotated*. Such

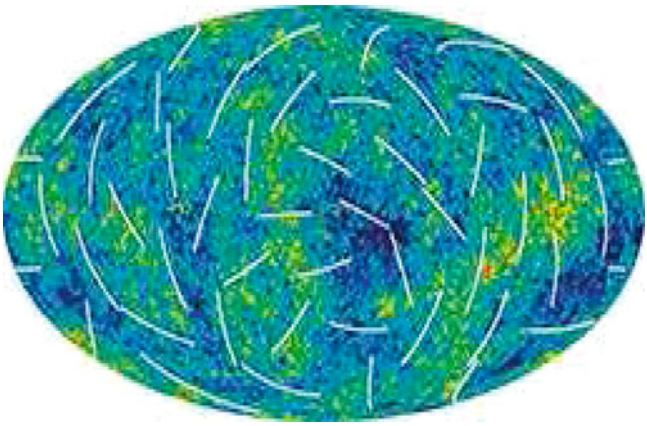


Fig. 8.63 Variation of temperature and polarization in various directions from the earth (source: http://en.wikipedia.org/wiki/Big_Bang)

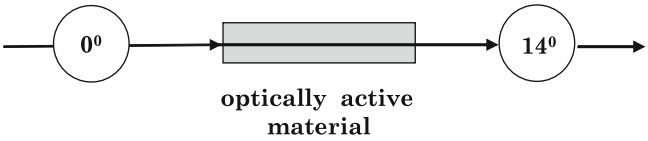


Fig. 8.64 Rotation of the axis of polarization by an optically active material

materials are said to be **optically active**. Thus, in Fig. 8.64 we have an example of a rotation by 14°:

Generally, the angle of rotation, $\Delta\theta$, is proportional to the distance ℓ of material through which the beam passes: $\Delta\theta \propto \ell$.

If $\Delta\theta$ is positive, corresponding to a **clockwise rotation**, the material is said to be **right handed**. Examples of such materials are: solutions of dextrose sugar, quartz, and camphor.

If $\Delta\theta$ is negative, corresponding to a **counterclockwise rotation**, the material is said to be **left handed**. Examples of such materials are: solutions of levulose sugar (also called "fruit sugar" or **fructose**), nicotine, menthol, and turpentine.

> **Note**
>
> Corn syrup, sold as KARO syrup, is a mixture of dextrose and levulose; it is a right-handed material.

It can be shown that for a material to be optically active, its **molecular structure must be such as to differ from its mirror image**. Generally, a system that differs from its mirror image is said to be **chiral**. Systems whose mirror images differ from each other are said to possess **chirality**. The molecule and its mirror image are called **enantiomorphs**.

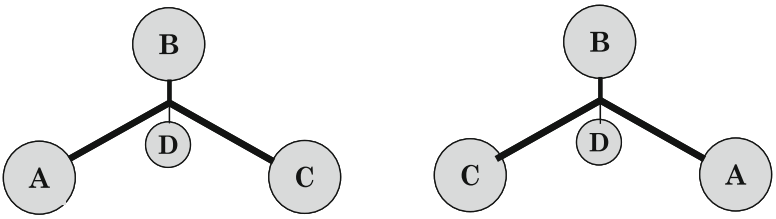


Fig. 8.65 Enantiomorphs

In the figure below, we exhibit two molecules that have the same set of four atoms, A,B,C, and D. Atoms A and C are in the foreground, while atom D is in the rear. They are mirror images of each other. Note that they are different in that one cannot be rotated into the other, as is the case with our right and left hands (Fig. 8.65).

Notes:

1. The mirror image of a material is left handed and vice versa. As an example, consider glucose: dextrose is simply D-glucose. The "D", for "dextrorotatory", means that dextrose rotates the polarization clockwise as you look at the polarized beam; dextrose is also said to be "right handed." L-glucose is the corresponding mirror image, "L" referring to its being "levorotatory," that is "left handed" or counterclockwise rotating.
2. The constant of proportionality in the relation $\Delta\theta \propto \ell$ is called the **specific rotatory power**, for which we will use the symbol σ. Thus

$$\Delta\theta = \sigma\ell. \tag{8.37}$$

3. **Rotatory power** depends upon the wavelength of the light. Also, for solutions, it depends upon the concentration of the solution, being proportional to the concentration. Consequently, the concentration of a solution can be determined from the angle of rotation that the solution produces on a polarized beam that is passed through it.

In Table 8.1 we present the specific rotary power of quartz and of a water solution of sucrose (commonly known as cane sugar). In the first case, the specific rotary power is expressed in degrees rotation/cm. In the case of a solution, optical rotation depends upon how much sugar is in the solution so that specific rotatory power is expressed in degrees rotation/cm *per* unit concentration of 1 g/cm^3. Thus,

$$\sigma = \sigma' \times c, \tag{8.38}$$

where *sigma*$'$ is the specific rotation per unit concentration and c is the concentration.

Notice the *negative* specific rotary power of levulose, corresponding to a *counterclockwise* optical rotation.

Table 8.1 Table of specific rotatory powers – optical activity and the asymmetry of biological systems

Material	Wavelength of light (nm)	Specific rotary power σ or σ'
Quartz	410	$\sigma = 475°$ per cm
Quartz	589	$\sigma = 217°$ per cm
Sucrose in water solution	410	$\sigma' = 15°$ per cm per g/cm^3 concentration
Sucrose in water solution	589	$\sigma' = 6.5°$ per cm per g/cm^3 concentration
Levulose in water solution	546	$\sigma' = -10.5°$ per cm per g/cm^3 concentration

Sample Problem 8-13

Suppose that 589 nm polarized light is passed through 5 mm of quartz. Through what angle will the axis of polarization be rotated?

Solution

$$\Delta\theta = \sigma\ell = 217 \times 0.5 = 109°.$$

Sample Problem 8-14

Suppose that 410 nm light is passed through 30 cm of a solution of sucrose in water having a concentration of 50 g/l. Through what angle will the axis of polarization be rotated?

Solution

One liter $=$1,000cm^3, so that 1 g/l$=$1 g/10^3 cm^3 $=$ 10^{-3} g/cm^3. The specific rotatory power is given by

$$\sigma = 150 \cdot 50 \times 10^{-3} = 0.750 \text{ per cm.} \tag{8.39}$$

Then the angle of optical rotation is given by

$$\Delta\theta = \sigma\ell = 0.75 \times 30 = 23°. \tag{8.40}$$

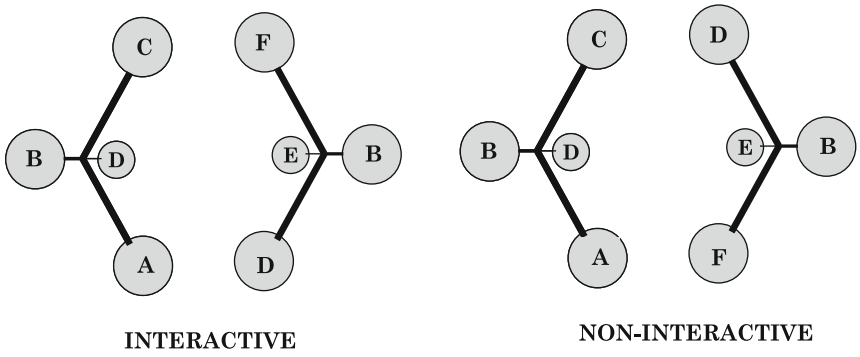


Fig. 8.66 An interactive pair of appropriate enantiomorphs vs. a noninteractive pair

8.12.10 Our Chiral Biosphere

Our entire **biosphere** relies on the chemical interactions within a huge system of optically active molecules. Only one member of each enantiomorphic pair is represented in this system. Some are dextrorotatory while others are levorotatory. A mirror image of representative members of our biosphere will not react at all or not react correctly with the system of molecules of our biosphere.

In Fig. 8.66 we see two schematics of a pair of molecules approaching each other, an ABCD molecule and a BDEF molecule. The configuration of the ABCD molecule is the same in both schematics. However, the BDEF molecule is represented by the mirror image enantiomers. Consider the left schematic. We see the following pairs of atoms lined up next to each other such that an interaction can take place: AD, CF, and DE. On the other hand, in the right schematic this pairing cannot take place simultaneously: we see the pairing AF, CD and DE. As a result there is no chemical interaction.

For example, L-glucose is not digestible. Levulose, which is L-fructose (commonly referred to as "fructose" or "fruit sugar") is digestible, while its mirror image, D-Fructose, is not. It is interesting to note that the asymmetry of the molecules of our biosphere is matched and may be connected with the asymmetry of our bodies (e.g., the heart is on the left side).

[19]In principle, one could have an entire biosphere of animals and plants looking outwardly absolutely identical to ours, except that the chiral molecules are mirror images of ours. Yet, any attempted mating between a member of our biosphere and a member of the opposite sex of the mirror image biosphere would be unsuccessful.

[19]The strikingly different chemical properties of enantiomorphs has played an extremely important role in pharmaceuticals. I will present two different scenarios, one which led to wonderful pharmaceutical benefits and the other to more expensive and probably unnecessary medications.

A fascinating question arises: **Why is it that only one of the two systems of mirror image biospheres are found on earth?** Why is life on earth **homochiral**? If life arose in many places on earth *independently*, one would expect equal probability for the two types of biospheres to have developed. Are there external factors – e.g., polarization of electromagnetic radiation from outer space (there is such a thing as "circularly" polarized electromagnetic waves which might have been effective) or other cosmic radiation with a handedness (distinguishing left from right) – which might have favored our type of biosphere over its mirror image?

A number of explanations have been put forward. One goes like this: An analysis of the way two systems of mirror image biospheres would evolve indicates that certain sorts of competition between the two systems will result in an *instability* of the ratio of the populations of the two systems: A small inequality in the two populations – e.g., population of system A greater than population of system B leads to the eventual extinction of the system with an initially lower population. Thus, it may well be that long ago the earth had our present biosphere along with its mirror image biosphere. The mirror image biosphere became extinct and we were left with ours alone. You can see how this idea might raise more questions than it answers.

Recently, it has been shown[20] that a combination of unpolarized light and a magnetic field can produce an excess of one enantiomorph over another starting with a nonchiral medium. This phenomenon is believed to be a possible explanation for the homochirality of life on earth. Much further research remains to be done to give this possibility strong support.

---

(1) The first scenario is represented by **thalidomide**. Among its early uses was its treatment of morning sickness in pregnant women. Unfortunately, the drug was prone to producing severe birth defects and was withdrawn from the market. Subsequently, it was found that one of its enantiomers is responsible for the birth defects, while the other provides the desired pharmaceutical effects. Separating out the desired enantiomer made the drug available for numerous diseases and medical issues. (2) The second scenario is the role of patents in the pharmaceutical industry. Consider, **Prilosec**, which is a drug used to treat heartburn. When the patent owned by the pharmaceutical company AstraZeneca ran out, the company produced a form that had the pure drug-effective enantiomer and called this new drug **Nexium**. The company was able to obtain a new patent on the drug and sell it at a much higher price than Prilosec. I have researched the web for many studies that compared the two drugs and have yet (1-16-2011) to find one that reported a significant difference in their effectiveness – that is, more than a few percent in whatever way improvement can be measured. You can read information on this subject on the following websites (1-29-2011):

http://en.wikipedia.org/wiki/AstraZeneca#Nexium;
http://en.wikipedia.org/wiki/Esomeprazole#cite_note-12;
http://www.medscape.com/viewarticle/481198_8.

In spite of the negative responses toward AstraZeneca's actions, you should always be on the lookout for contrary opinions about the Nexium-Prilosec controversy. Beware about the significance of a claim that drug A is more effective than drug B. The comparative effectiveness might mean, in simple terms, that A is 5% more effective than B. If so, A might be 90% effective while B might be 85% effective. The ratio is a mere 1.05. On the other hand, the respective effectivenesses might be 10$ and 5%, in which case A is twice as effective as B!

[20] See the two articles in the June 22, 2000 issue of the science journal NATURE.

8.13 Terms

- Angle of incidence
- Angle of reflection
- Angle of refraction
- Birefringence
- Central ray
- Chromatic aberration
- Concave lens
- Converging lens
- Convex lens
- Diffraction grating
- Diffraction
- Diffraction angle
- Diffuse reflection
- Dispersion
- Diverging lens
- Doppler effect
- Fiber optics communication
- Focal length
- Handedness of our biosphere
- Image
- Image distance
- Image point
- Impedance
- Index of refraction
- Length scale of roughness – l_r
- Lens
- Lens axis
- Magnification
- Magnifying glass
- Minimum image diameter
- Mirage
- Mirror image
- Object distance
- Object point
- One-to-one correspondence
- Parallel ray
- Plane wave
- Prism
- Real image
- Reflectance
- Reflection of a wave
- Refraction
- Scattered wave
- Scattering of a wave
- Snell's Law
- Specular reflection
- Spherical aberration
- Thin lens approximation
- Thin lens equation
- Transmittance
- Virtual image
- Wave crests

8.14 Important Equations

Condition for specular reflection:

$$\lambda \gg \ell_{\mathrm{r}}. \tag{8.41}$$

Condition for diffuse reflection:

$$\lambda \ll \ell_{\mathrm{r}}. \tag{8.42}$$

Speed of light in a medium of index of refraction n:

$$v = \frac{c}{n}. \tag{8.43}$$

Snell's Law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \tag{8.44}$$

Generalized law for refraction:

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}. \tag{8.45}$$

Equation for the critical angle for total internal reflection:

$$\sin\theta_c = \frac{n_1}{n_2}. \tag{8.46}$$

Thin lens equation:

$$\frac{1}{d_O} + \frac{1}{d_I} = \frac{1}{f}. \tag{8.47}$$

Magnification vs. object/image distances:

$$M \equiv \frac{h_I}{h_O} = \left|\frac{d_I}{d_O}\right|. \tag{8.48}$$

Reflectance vs. indices of refraction:

$$R = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2. \tag{8.49}$$

Reflectance in terms of the wave velocities:

$$R = \left(\frac{v_2 - v_1}{v_2 + v_1}\right)^2. \tag{8.50}$$

Approximate equation for the Doppler effect:

$$\frac{\Delta f}{f} \sim \frac{u}{v}. \tag{8.51}$$

8.15 Questions and Problems for Chap. 8

1. When the dimensions of scattering particles are smaller than the wavelength of blue light, (CHOOSE ONE)

   (a) Red light is scattered more effectively than blue light.
   (b) Both red and blue light are scattered about equally and better than green light.

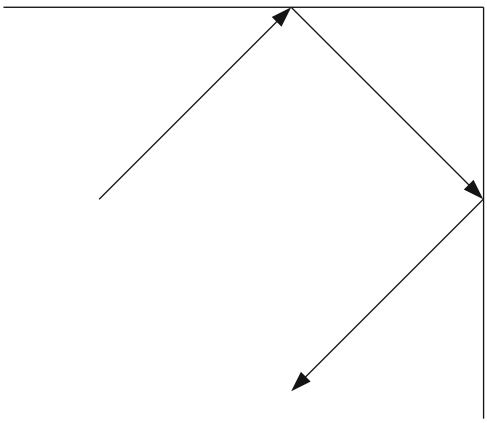


Fig. 8.67 Light ray reflecting off two perpendicular mirrors

(c) Both red and blue light are scattered about equally but not as well as green light.
(d) Blue light is scattered more effectively than red light.

2. (a) Red light diffracts:more/less: than blue light.
   (b) A soprano's voice diffracts more/less: than a bass's voice.
3. The animal called the **bat** is not blind.[21] However, it hunts at night and makes use of **echo-location**, wherein an ultrasonic series of pulses are emitted and their reflection used to locate a prey.
   (a) Why must the wavelength of the sound be much smaller than the object size for the bat to have a clear "image" of the object?
   (b) Estimate the minimum frequency of sound necessary for the bat to discriminate objects with a resolution on the order of 1 mm.
4. What two physical phenomena account for the ability of a prism to analyze light?
5. Will an FM radio wave with frequency 89.3 MHz be reflected **specularly** or **diffusely** off a field of corn?
6. A laser beam starts out on earth with a diameter of 2 mm. Find the diameter of the image of the laser beam on the moon, a distance 400,000 km away. The laser's wavelength is 6,300 $\mathring{A}$.
7. A laser beam is to produce the smallest possible image on a screen 1/,km away. Its wavelength is 5,500 $\mathring{A}$. What should the size of the aperture be? What will be size of the image?
8. A light ray is reflected off two mirrors that are at right angles with respect to each other. Prove that the reflected ray is parallel to the incident ray. See Fig. 8.67.

[21] See Wikipedia (1-11-2011): http://en.wikipedia.org/wiki/Bat.

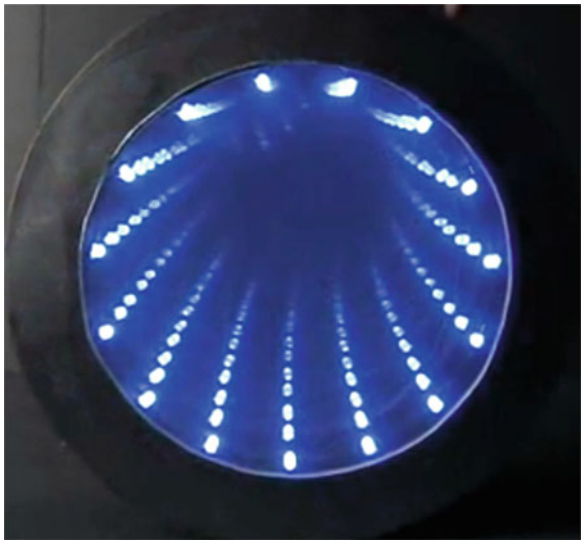


Fig. 8.68 A device that produces an infinite series of images (source: http://www.youtube.com/watch?v=VTONKZkaVX4&NR=1&feature=fvwp)

9. A person is standing at a distance of 5 ft from a mirror. What is the apparent distance of the person's image from the person?
10. A popular device for producing an image of an infinite series of replications of actual objects uses a mirror plus a sheet of glass that is coated so as to have a very low transmission. We see one of these devices in Fig. 8.68. The only lights that are actually present are those on a circle near the perimeter. All the other concentric circles of lights are virtual.

    Use physical principles to account for the infinite series of images as well as the fact that the images appear to be increasingly distant.
11. When light passes through a color filter, there is both reflection at the interface between air and the filter when the light strikes the filter from the air as well as reflection at the second interface where the light leaves from the filter and enters the air.

    Suppose that the index of refraction of the filter material is 1.5.

    (a) Show that the reflectance R at the first interface is 1/25 or 4%.
    (b) Find the reflectance at the second interface.[22]

[22] In fact, the light that strikes the second interface is partially transmitted into the air. The part that reflects back from the second interface toward the first interface is also partially reflected back from the first interface toward the second interface. The light can be represented as making an infinite number of attempts to "escape" from the filter into the air through the second interface. The resultant ultimate transmission out into the air is a bit more than the square of the transmittance through one interface, in this case a bit more than 92%.

12. Consider a steel rod that is struck at one end so as to produce a sound wave that travels down the rod. When the sound meets the other end, which is an interface between the steel rod and the air, there is reflection and transmission. Calculate the reflectance and transmittance at the steel-air interface. The sound impedance Z of air is given by $Z = \rho v = (1.3\,\mathrm{kg/m^3})(340\,\mathrm{m/s}) = 440\,\mathrm{kg/m^2 - s}$. The sound impedance of steel is $Z = \rho v = (8{,}000\,\mathrm{kg/m^3})(5{,}000\,\mathrm{m/s}) = 4 \times 10^7\,\mathrm{kg/m^2 - s}$.
13. Let I_{trans} = transmitted intensity and I_{inc} = incident intensity. Then the transmittance T and reflectance R are given by:

$$T = \frac{I_{\mathrm{trans}}}{I_{\mathrm{inc}}} = 1 - R. \tag{8.52}$$

(a) **Express** the transmittance of a light wave in terms of the **indices of refraction**. See Sect. 8.3.1.
(b) **Express** the transmittance of a sound wave in terms of the **impedances**.
(c) Suppose that my voice produces a sound level of 40 dB at the surface of a lake.

i. What is the reflectance? Assume a normally incident sound wave – so that you can use the results of (a) and (b) above.
ii. Find the sound level of the transmitted sound.

14. (a) Prove that the reflectance R of light can be expressed as:

$$R = \left[\frac{v_2 - v_1}{v_2 + v_1}\right]^2. \tag{8.53}$$

Note that this expression can be rewritten in terms of the *ratio* v_2/v *alone*:

$$R = \left[\frac{v_2/v_1 - 1}{v_2/v_1 + 1}\right]^2. \tag{8.54}$$

(b) **Prove** that the reflectance for **sound** can be written as:

$$R = \left[\frac{\frac{Z_2}{Z_1} - 1}{\frac{Z_2}{Z_1} + 1}\right]^2. \tag{8.55}$$

Only the ratio of the impedances, $Z \equiv \rho v$, appears in the relation.

15. For flint glass, the critical angle is 37°. Thus,

(a) Light incident on the glass from the air with an incident angle larger than 37° will be totally **refracted**.

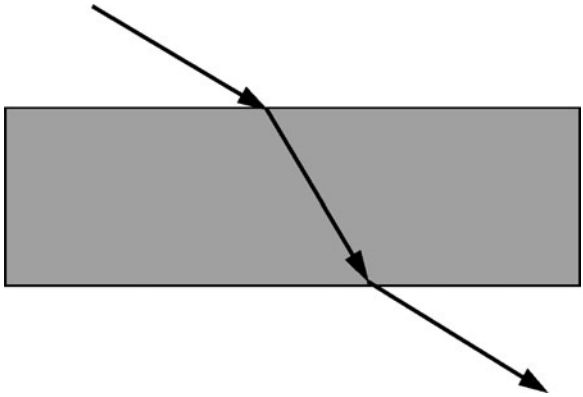


Fig. 8.69 Refraction through a slab of glass

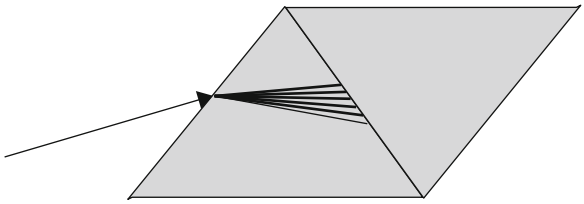


Fig. 8.70 Dispersion with two prisms in tandem

(b) Light incident on the glass from the air with an incident angle larger than 37° will be totally **reflected**.
(c) Light incident on the air from the glass with an incident angle larger than 37° will be totally **reflected**.
(d) Light incident on the air from the glass with an incident angle smaller than 37° will be totally **reflected**.
(e) Light incident on the glass from the air with an incident angle smaller than 37° will be totally **reflected**.

16. Prove that when a light ray is refracted twice by a block of glass with parallel faces (see Fig. 8.69), that the outgoing ray is parallel to the incoming ray.
17. We know when a ray of white light is incident upon a prism, dispersion will lead to an outgoing beam with a rainbow of colors, each component wavelength traveling in a different direction. Suppose that a ray of white light is incident upon a pair of prisms in tandem, as shown in Fig. 8.70. Describe the outgoing beam of light.
18. A light ray, traveling in water, is incident on a water-glass interface at an angle of incidence of 30°. Find the angle of refraction if the index of refraction of the glass is 1.5.
19. Find the critical angle for total reflection of a light wave in a glass having an incident of refraction of 1.7 at an interface with air.
20. Find the **critical angle** for total reflection of a sound wave in air incident on a water surface.

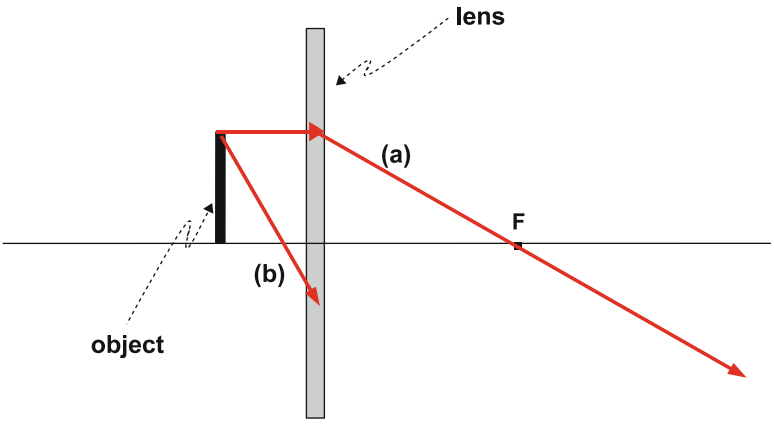
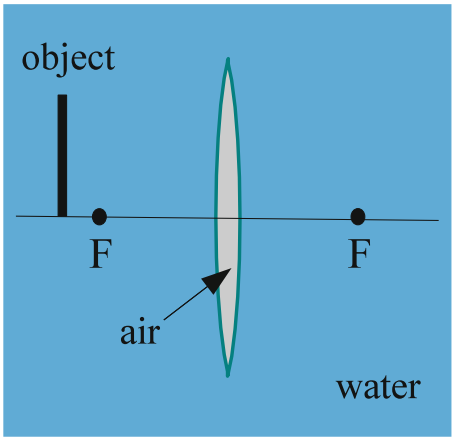


Fig. 8.71 Ray diagram of a lens

Fig. 8.72 Air lens in water



21. Fig. 8.71 is a diagram of an object (the vertical black rectangle) that is to the left of a lens (the vertical gray rectangle).

    We see two rays that leave the top of the object and strike the lens at two positions along the lens. The parallel ray (a) is shown passing through the focal point. The second ray (b) is not completely shown.

    (a) Is the lens converging or diverging?
    (b) Complete the diagram by showing where the image is formed and then showing where ray (b) progresses to the right of the lens.
    (c) Is the image real or virtual?
    (d) Suppose that the object distance is $0.5\,|f|$, where f is the focal length. Determine the image distance in terms of f.

22. Fig. 8.72 shows a **biconvex** air lens under water. It consists of a balloon in the shape of a convex lens and filled with air.

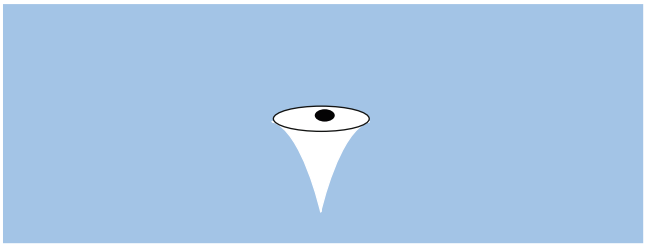


Fig. 8.73 Eyeball underwater looking up toward the surface of a swimming pool

Where would the image of the object arrow be located? To answer this question, think of the way a light beam will be refracted in passing from water into air and from air into water.

(a) Between the object arrow and the focal point, F
(b) Between the focal point, F, and the lens
(c) On the right side of the lens
(d) To the left of the object arrow
(e) There would be no image for this type of lens

23. Suppose that a piece of opaque material is placed just in front of the lens of a slide projector, thus blocking the light emanating from the *top half* of the lens. Describe the effect on the image on the screen. Choose from the following:

    (a) The top half of the image be removed.
    (b) The bottom half of the image be removed.
    (c) Neither of the above be true. (Describe very qualitatively what one should observe).

    HINT: Use the fact that there is a one-to-one correspondence between an object point and an image point. In fact, the image point is produced by the set of all rays that leave an object point and pass through the lens.

24. Suppose that you look at an object through a diverging lens having a focal length of -4 cm. The object is placed at a distance of 7 cm from the lens. Find the image distance. Is the image erect or inverted? Real or virtual?
25. Suppose that a person of 5 ft height is at a distance of 25 ft from a convex lens having a focal length of 0.4 ft. Find the image distance and the image height. Is the image real or virtual? Erect or inverted?
26. Suppose that you are submerged in a pool of water with an absolutely calm surface. Above the surface the room is full of light. The walls and floor are perfectly black, so that they absorb all light completely, and do not reflect any light back into the water. What will you see when you look up at the surface? (You will *not* see a fully lit surface.)

    HINT: Use reversibility of rays that are refracted. Consider the paths of all rays of light that emanate from the eye and can emerge from the water into the air above (Fig. 8.73).

27. On a road, John drives toward Marsha tooting his horn. Marsha immediately notices that

    (a) John looks slightly bluer than normal.
    (b) John looks slightly redder than normal.
    (c) The horn is pitched a little higher than normal.
    (d) The horn is pitched a little lower than normal.

28. The atomic spectrum of hydrogen is observed in the light coming from a star. If the star was *not* moving (relative to the earth), the *red* line would be observed to have a wavelength of 6,560Å. Instead, the *observed wavelength* is 6,562Å.

    (a) Is the star moving toward or away from the earth?
    (b) Find the speed u of the star relative to the earth, assuming it is moving along a line joining it to the earth. You may assume that $u \ll c$ and therefore use (8.28).

29. A train is moving *toward* you at a speed of 100 km/h while sounding a whistle having an intrinsic frequency of 600 Hz.

    (a) Using a sound velocity of 340 m/s and the small speed approximation (8.28), calculate the frequency of sound that you will perceive.
    (b) Repeat the above if the train is moving away from you.

30. Here is how a radar device is used to determine the speed of a car: The device sends a radar signal of frequency 10 GHz ($= 10^{10}$ Hz) toward the car. Given that the car is moving directly toward the source, the car receives a Doppler shifted frequency f' with respect to the car. In turn, the car becomes a source in sending back radar waves to the device. Since the car is moving toward the device (which has a radio transmitter and a radar receiver), the frequency observed (measured) by the receiver will be a Doppler shifted frequency f'' of f'. We have **two** Doppler shifts. Given that $u \ll c$, we can use (8.28) except that we must double the shift. Thus, we obtain

$$\Delta f = f'' - f = (f'' - f') + (f' - f) = \frac{2u}{c} f. \tag{8.56}$$

    Now suppose that f'' and f produce a beat frequency of 2,000 Hz.
    Find the speed of the car.

31. The Doppler effect of ultrasound is used to determine the flow velocity of blood in a blood vessel; this is especially useful in determining whether there is a blood clot impeding flow in an artery of a leg.

    (a) Assuming that the sound velocity in the body is about 1,500 m/s and that one needs a resolution of an image to be about 1mm, what should the minimum wavelength of the ultrasound be and the corresponding frequency?

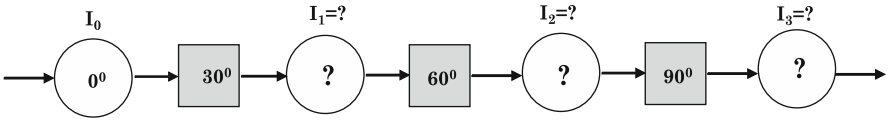


Fig. 8.74 Sequence of three polarizers

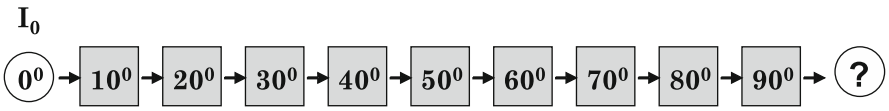


Fig. 8.75 Sequence of nine polarizers

(b) Suppose that the frequency used is that obtained in the previous problem. Assume also that the beat frequency is 1 Hz between the frequency of the incident sound and the frequency of the reflected sound. Determine the velocity of the blood in the artery if the blood is flowing directly toward the source of sound.

32. *(This is a complex algebraic problem.)* A train passes you by while sounding a whistle whose frequency varies from 550 to 500 Hz. That is

$$f'_{\text{toward}} = 550\,\text{Hz}$$
$$f'_{\text{away}} = 500\,\text{Hz}.$$

Assume that the speed of sound in air is 340 m/s.

(a) Find the speed of the train – assumed constant.
(b) Find the frequency f of the whistle.

Hint: Show that

$$\frac{f'_{\text{towards}}}{f'_{\text{away}}} = \frac{1 + u/v}{1 - u/v}. \tag{8.57}$$

Then show that

$$\frac{u}{v} = \frac{\frac{f'_{\text{towards}}}{f'_{\text{away}}} - 1}{\frac{f'_{\text{towards}}}{f'_{\text{away}}} + 1}. \tag{8.58}$$

33. (a) Fill in the circles and determine the intensities for the sequence of polarizers in Fig. 8.74.

(b) Repeat for the following sequence of nine polarizers, determining only the final state of the beam (Fig. 8.75).
(c) Can you see how one could use ideal polarizers to rotate the axis of polarization by 90° without any loss in intensity? Explain.

34. Suppose you are walking due West down a street at sunset (Fig. 8.76).

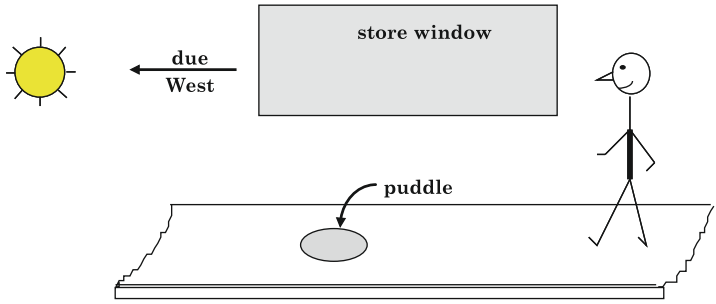


Fig. 8.76 Polarizing of light reflected off a puddle of water

Determine the direction of the axis of partial polarization of sunlight which you see

(a) Reflected off the window of a store
(b) Reflected off the water puddle on the sidewalk
(c) Scattered by the atmosphere directly above you

35. Polarized light of wavelength 410 nm is passed through a distance of 7 mm of quartz. Find the angle through which the axis of polarization is rotated.
36. This problem shows us how optical activity can be used to quickly determine the concentration of a solution of chiral molecules.

    Polarized light with a wavelength of 589 nm is passed through a distance 10 cm of a solution of sucrose in water. The axis of polarization is found to be rotated by an angle of 19°. Determine the concentration of sucrose.
37. The following problem is not a trivial one at all. You are presented with this problem mainly to get you thinking about it so that you can appreciate how difficult the problem is:

    Suppose that you were establishing communication with an extraterrestrial being. You describe yourself in broad terms and succeed to establish that you have an organ (your heart) that pumps fluid (blood) through your body. You now want to indicate that the heart is essentially on the ***left*** *side of your body. How might you establish the difference between* ***left*** *and* ***right*** *using only radio communication and thus without identification of and reference to various celestial bodies? Might your ability to refer to celestial bodies help?*

Chapter 9
The Ear

SOUND

Insects one hears
and one hears the talk of men –
with different ears

***Haiku** by **Masaoka Shiki** (1867–1902)*

We have studied the nature of sound and how sound waves are produced and propagate through media. However, the focus of this text is sound as experienced by people. Sounds of insects, of men, of a multitude of sources, reach our ears, perhaps providing us with our principle means of communication with the outside world. How can we hear these sounds "differently"? What happens to the sound that enters our ears? What is the essence of hearing? What is the source and explanation for the pleasure we have in hearing beautiful music or for the annoyance at hearing loud or dissonant noise? Many a reader might hope that science can arrive at answers to such questions. Unfortunately, science is severely limited in this domain.

Hearing begins with the ears and ends with the **brain**. The ear is the organ that is used to gather a sound wave and convert the waveform of the sound wave, as faithfully as possible, into **nerve signals** that travel to the brain to be analyzed and interpreted. Our mode of hearing is determined in most instances by the manner in which the brain analyzes the auditory nerve signals it receives. In addition, the brain has been shown to have the remarkable capacity to alter the physical state of the ear and hence its manner of converting sound into nerve signals. In any case, the physical distinction between the brain and the ears exists: The poetical "ears" in Shiki's Haiku include the brain, while the physicist's "ears" do not.

This chapter is concerned with the physics of the **human ear** – that is, the physical processes whereby the ear converts a sound wave into nerve signals. We do not discuss how the brain analyzes these nerve signals. This subject is beyond the scope of this text. Briefly, the ear can be compared to a highly sensitive microphone, capable of responding to a range of frequencies from 20 to 20,000 Hz and to a range of intensities spanning 12 orders of magnitude, with a high efficiency and extremely low level of distortion. A knowledge of how the ear functions enables one to partially understand how what we hear is related to the sound incident upon our ears. This is the subject of Chap. 10. In particular, we will be able to qualitatively account for the response characteristics of the ear and our ability to discriminate pitch. Most significantly, in Chap. 10, we will be able to provide a

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_9,

qualitative basis for the existence of consonant and dissonant musical intervals and the related perception of **combination tones** which are certain sounds that a person can hear even though they are not present in the sound wave incident upon the ear.

9.1 Broad Outline of the Conversion Process

Figure 9.1 is a drawing of the human ear. The **pinna**, or outer ear, serves to funnel a sound wave into the auditory canal. The wave travels down the canal, at the end of which it sets the **tympanic membrane** (also referred to as the **eardrum**) into motion. The eardrum in turn sets a system of three bones – the **ossicles** – into motion. At the other end of the ossicles is the **footplate**, which is the base of the **stapes** (also referred to as the **stirrup**). (See Fig. 9.5 for more details.) In fact, the footplate covers the hole shown in the snail-shaped **cochlea**. This hole, the **oval window**, leads into the inner chamber of the cochlea that is filled with a fluid – the **cochlear fluid** – that is set into motion by the vibrations of the footplate. The cochlea contains the neural sensors that transmit the information about the sound to the brain.

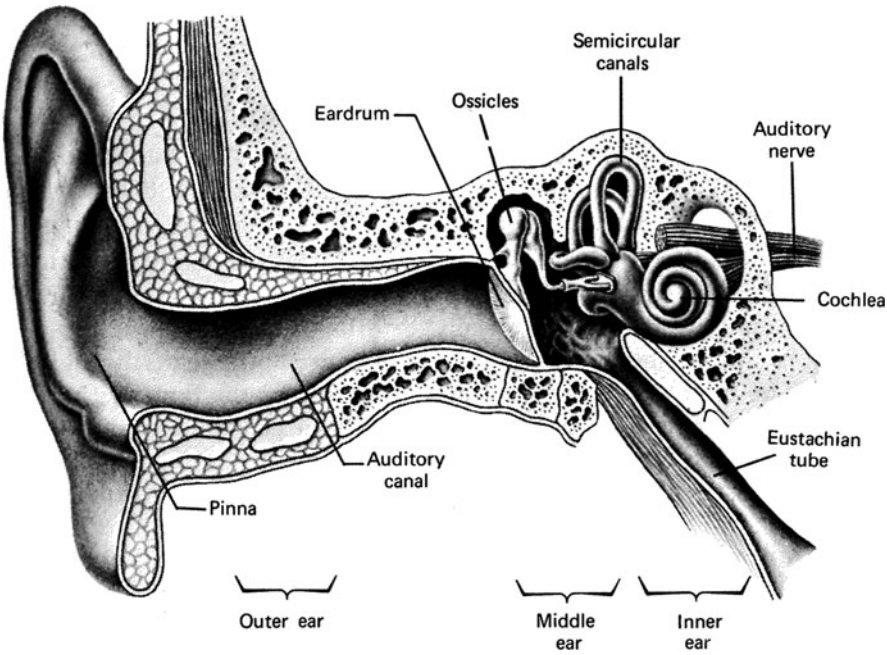


Fig. 9.1 The human ear (source: John B. Palmer, *Anatomy for Speech and Hearing*, 2nd ed., (Harper & Row, 1972))

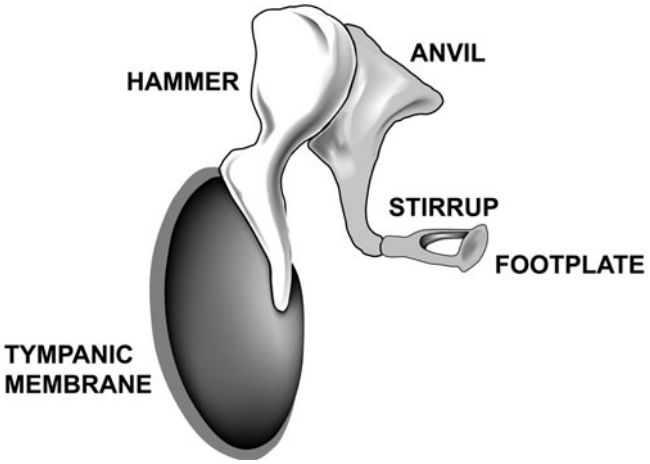


Fig. 9.2 Ossicles – details (source: courtesy of Russ Dewey)

Fig. 9.3 Photo of replica of the Ossicles (photo: Leon Gunther)

In Fig. 9.2 we see a drawing of the ossicles, while in Fig. 9.3 we see a closeup photograph of a plastic replica of the actual ossicles alongside a centimeter ruler.

To view the replica directly with your eyes can be quite breathtaking: These bones, which are an essential instrument for transmitting the wonderful sounds we hear, over an incredible range of intensities and with great fidelity, are puny and

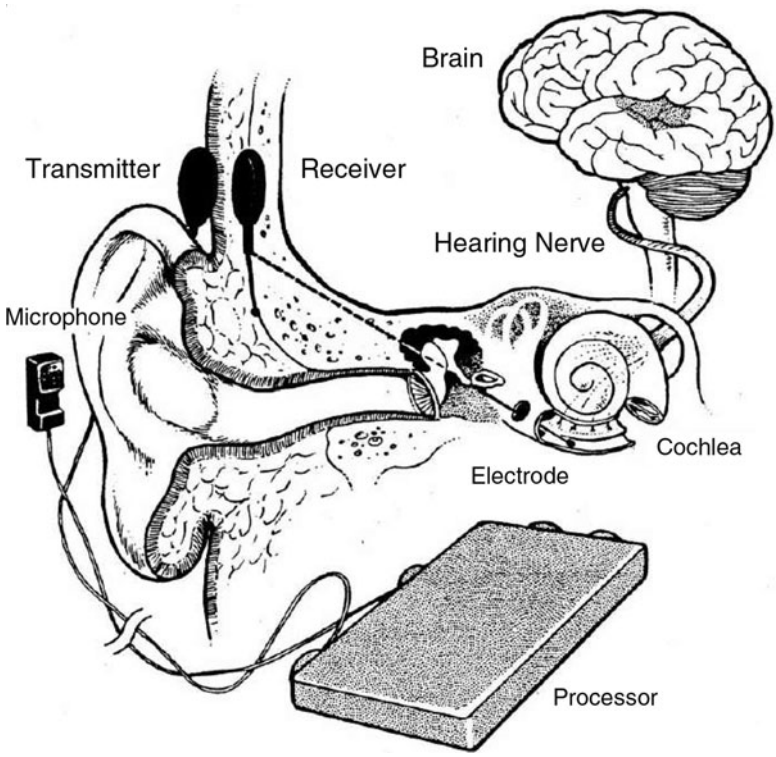


Fig. 9.4 Cochlear implant (source: Lahey Medical Center Journal, Burlington, MA)

delicate. Equally amazing is the strong evidence that the three bones evolved from a combination of parts of the gills of a fish and the jawbones of reptiles.[1]

In Fig. 9.4 we see a schematic of the entire ear, along with a cut away to see inside the cochlea. This figure was produced for the purpose of showing how a cochlear implant works. See later in this chapter for more information.

[1]Here is an excerpt from the website http://museum.utep.edu/archive/biology/DDossicles.htm: "Hearing is a wonderful thing, able to translate vibrations of air into sound. Numerous desert creatures rely on sound more than on sight, for many are nocturnal, only active during the dark hours. Part of the great sensitivity in mammals is due to the three small bones in our middle ears, the auditory ossicles. These transmit and amplify the vibrations of the ear drum, conveying them to the inner ear. What's fascinating from an evolutionary viewpoint is their origins. Studies of embryos and fossils trace their origin far back in time. The ossicle next to the inner ear, the stapes, can be traced back to part of a gill arch in a very distant fish ancestor. The other two ossicles, the malleus and incus, are derived from bones that, in our reptilian ancestry, formed the joint between skull and lower jaw, as they do today in modern reptiles. After incipient mammals evolved a new jaw joint, those bones were in perfect position to be incorporated into a hearing device in evolution's favorite avenue–jury-rigging structures for new roles."

One may well wonder why so many steps are involved in the conversion process from sound to nerve signals. Why has not the ear evolved so that an incident sound wave will be incident directly upon the oval window? The primary reason is that of a very poor matching of the **impedances** of the air and the cochlear fluid.

The **impedance** Z is the product of the mass density ρ and the speed of sound v. Thus,

$$Z = \rho v. \tag{9.1}$$

It determines the resilience of a medium to being disturbed by a change in the external pressure. In Sect. 8.3.2, we discussed its relevance in the reflection of sound at an interface between two media: When a sound wave is traveling in one medium and is then incident upon an interface with a second medium, a certain fraction of sound energy will be reflected and a certain fraction will be transmitted. In (8.14) and (8.16) of Chap. 8, we have an expression for the reflectance R in terms of the impedances. With $Z = \rho v$, we have

$$R = \left(\frac{Z_2 - Z_1}{Z_2 + Z_1} \right)^2. \tag{9.2}$$

When the two impedances are equal, the reflectance vanishes and there is total transmission. Correspondingly, given that the transmittance T is given by

$$T = 1 - R \tag{9.3}$$

a bit of algebra leads to

$$T = \frac{4 Z_2 Z_1}{(Z_2 + Z_1)^2}. \tag{9.4}$$

This equation can be rewritten in a different algebraic form that is illuminating. If we divide both numerator and denominator by Z_1^2, we obtain

$$T = \frac{4 Z_2 / Z_1}{(Z_2/Z_1)^2 + 1}. \tag{9.5}$$

Thus we see that the **ratio of the two impedances determines the transmittance**. We already observed that the reflectance vanishes when the two impedances are equal, so that their ratio is one. For all other ratios, larger or smaller than one, the transmittance is less than one. The closer the ratio of the respective impedances of the media is to one, the closer will the transmittance be to unity. For example, a ratio of $\frac{1}{2}$ (or 2) leads to a fraction $8/9$ transmitted. The ratio for cochlear fluid to air is about 3,300 : 1, which results in only a fraction of about one part in 1,000 of the sound energy transmitted, corresponding to a decrease in sound level of 30 dB.

We will see how the eardrum and the middle ear serve to increase the fraction of sound transmitted to the cochlear fluid.

9.2 The Auditory Canal

The **auditory canal** is a crooked tunnel (not as straight as Fig. 9.1 would suggest) one of whose functions is to protect the delicate eardrum from injury. It also serves as a resonator which aids in reducing the fraction of sound reflected back into the air. For an approximate calculation of its resonance frequencies, it is adequately represented by an open-closed tube, with a length of about 2.7 cm and a diameter of about 7 mm. Suppose we assume that the temperature in the canal is 30°C (which lies between room temperature and a body temperature of 38°C). In Problem 9.1, it is shown that the speed of sound at this temperature is 346 m/s and that the fundamental frequency f_1 of the tube is 3,200 Hz. The overtone frequencies are $3f_1 = 9{,}600$ Hz, $5f_1 = 16{,}000$ Hz, ...

It has been shown that the fundamental resonance provides an amplification of the intensity by a factor of $3-10$ for frequencies between 2 and 5 kHz. It is therefore no mere coincidence that the ear as a whole is most sensitive to sound waves with a frequency of about 3 kHz (see Sect. 10.1).

9.3 The Eardrum

Before we go into detail, it is important to remember that it is the *difference* between the pressure on the outside, exposed side of the eardrum and the pressure within the ear, that leads to the net force on the eardrum. This difference is the **sound pressure**. The pressure within the ear is ideally maintained at the **ambient pressure** – that is, the pressure in the absence of the sound wave and therefore normally about one atmosphere. This is achieved by air inside of the ear being contiguous with the outside air via the **Eustachian tube**. (See Fig. 9.1.)

The eardrum is a very delicate membrane that serves to gather up sound energy and transmit it further on into the ear. It is oval in shape, having dimensions of approximately 9 mm by 12 mm. It also provides the major means of overcoming the **mismatch of impedances** by virtue of the fact that its area is about 15 times that of the oval window. As a result of the eardrum alone, the **sound pressure** acting on the oval window would be about 15 times that acting on the eardrum by the sound wave.

This effect can be understood qualitatively by considering the relative ease with which we can push a nail into the ground as opposed to pushing a block of wood into the ground. The force available is the same in both cases. However, we get a greater pressure by the nail on the ground because the area of contact between the nail and the ground is much smaller than the area of contact between a block of wood and the ground. (See Fig. 9.5.)

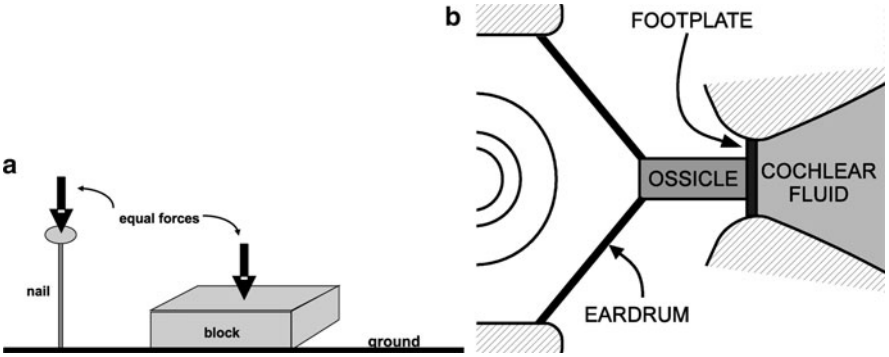


Fig. 9.5 Abstract diagram of the eardrum. (**a**) Although you have the same force, it is easier to push the nail into the ground, since it can assert more pressure. (**b**) The force of the sound wave is concentrated to a small area before entering the ossicle

9.4 The Ossicles

The ossicles, which are depicted in Fig. 9.2, consist of a set of three bones, known as the hammer, anvil, and stirrup (or, respectively, as the malleus, incus, and stapes, in medical terminology).

They serve as a further means of increasing the fraction of sound transmitted into the cochlear fluid. This increase is accomplished through the principle of **lever action**. Furthermore, they somewhat protect the inner ear from damage due to loud sounds. Obviously, they are inadequate in protecting the hearing of those who enjoy listening to loud rock bands. My own testing of students over a period of more than 30 years reveals a dramatic decrease in the highest frequency that can be heard. Around 1975, most students heard frequencies exceeding 20,000 Hz. By the year 2000, I have found that few students can hear frequencies above 18,000 Hz.

Consider the 7-ft seesaw illustrated in Fig. 9.6. A child of weight 40 lbs is seated to the left at end position, while a woman of weight 100 lbs is seated to the right at position B. The **fulcrum** (or pivot) at F is located at a position 2 ft from end B (i.e., $\overline{BF} = 2$ ft), so that the distance $\overline{AF}$ is 5 ft. The ratio $\overline{FB}/\overline{FA}$ is known as the **mechanical advantage**, which we will denote with the letter **r**. In the example above, the mechanical advantage is thus 2.5. In the case of the human ear, it has been reported to be only about 1.3. Generally, we have

$$\frac{F_B}{F_A} = \frac{\overline{AF}}{\overline{BF}} = \mathbf{r} \tag{9.6}$$

which is known as **Archimedes' Principle of Lever Action**.[2]

[2] **Archimedes**, who receives credit for this discovery, is purported to have stated, "Give me a long enough stick, a place to stand, and a pivot, and I'll move the earth!" The language here is quite

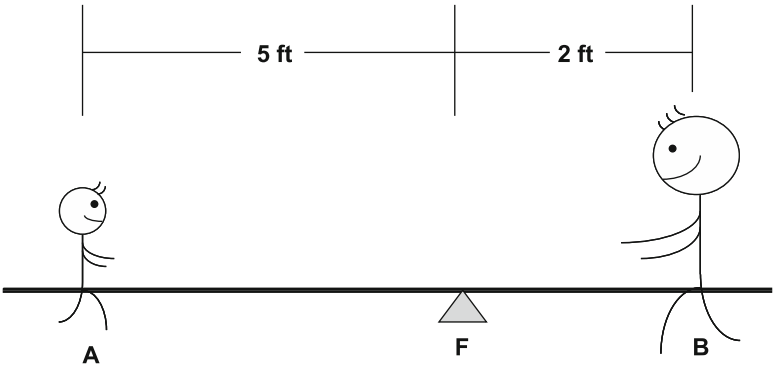
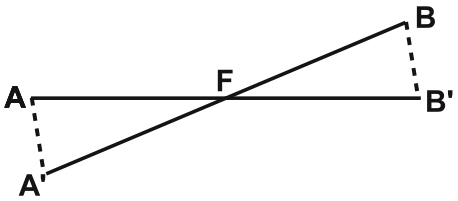


Fig. 9.6 A child and a woman on a seesaw

Fig. 9.7 Trigonometry of a lever



Lever action also leads to another essential means of increasing the fraction of sound energy transmitted to the cochlear fluid. A displacement of point A leads to a displacement of point B in ratio of the mechanical advantage **r**. (See Fig. 9.7.) This fact follows from simple trigonometry.

In the example of the seesaw, if end A is pushed down a distance of 5 in., the woman at B will move up a distance of only 2 in. In general, the ratio of the respective displacements, $\overline{AA'}$ and $\overline{BB'}$, at two ends of a lever is equal to the ratio **r** of the distances of the fulcrum from the two ends:

$$\frac{\overline{AA'}}{\overline{BB'}} = \frac{\overline{AF}}{\overline{BF}} = \mathbf{r}. \tag{9.7}$$

Since the ratio of the **forces**, F_B/F_A, is equal to 1.3 for the ear, the ratio $\overline{AA'}/\overline{BB'}$ is equal to 1.3 also, by virtue of (9.6). Thus, if the air pushes the point at which the hammer of the ossicles is connected to the eardrum by a distance of 1 nm (about three or four diameters of a single atom!), the footplate will move a distance of only about $1/1.3 \sim 0.77$ of a nanometer.

loose. In fact, we need no such fancy system to move the earth. We do so every time we jump up into the air or walk or run – albeit by a minuscule unobservable amount.

Note that the product of the force and displacement on one end is equal to the same product for the other end:

$$F_A \, \overline{AF} = F_B \, \overline{BF}. \tag{9.8}$$

In the example above, we have

$$40 \, \text{lbs} \times 5 \, \text{ft} = 100 \, \text{lbs} \times 2 \, \text{ft}. \tag{9.9}$$

9.5 Improving on the Impedance Mismatch: Details**

We will outline the basis for improving on the impedance mismatch. The expression for the transmittance in (9.5) in terms of air and fluid can be written as:

$$T = \frac{4Z_\text{a}/Z_\text{f}}{(Z_\text{a}/Z_\text{f})^2 + 1}. \tag{9.10}$$

The transmittance is low because $Z_\text{a} \ll Z_\text{f}$. We have pointed out that two factors play a role in increasing the transmittance – the ratio of the areas of the eardrum and the oval window, A_d/A_w, and the **mechanical advantage**, **r**, of the lever arm of the ossicles. An analysis to be outlined below leads to an effective increase in the ratio of the two impedances:

$$\frac{Z_\text{a}}{Z_\text{f}} \rightarrow \mathbf{r}^2 \frac{A_\text{d}}{A_\text{w}} \frac{Z_\text{a}}{Z_\text{f}}. \tag{9.11}$$

For the ear, the ratio of impedances is then increased by a factor of $(1.3)^2(15) \approx 25$, and therefore from 1/3,300 to 1/130. The corresponding transmittance is 0.03 with a reduction of the sound level by 15 dB instead of the 30 dB reduction without the eardrum and ossicles.

In Fig. 9.8, we see a simplified but concrete description of the process. An incident wave of intensity I_i strikes the eardrum with area A_d. A reflected wave has an intensity I_r. The eardrum has a displacement equal to that of the air at the eardrum, D_a, represented by the red arrow to the right. We see a schematic of the ossicles that provide lever action, with the fulcrum represented by the black dot. The result is a displacement D_f of the fluid, which moves with the oval window. The oval window has an area A_w. I_t is the intensity of the sound wave that is transmitted into the fluid, which I have indicated with a direction to the right even though the oval window is moving to the left at the instant shown in the figure.

Let us begin by discussing the effect of having different areas. The transmittance ultimately tells us what fraction of energy of the incident wave is transmitted – or alternatively what fraction of **power** P is transmitted. Thus,

$$T = \frac{P_\text{t}}{P_\text{i}}. \tag{9.12}$$

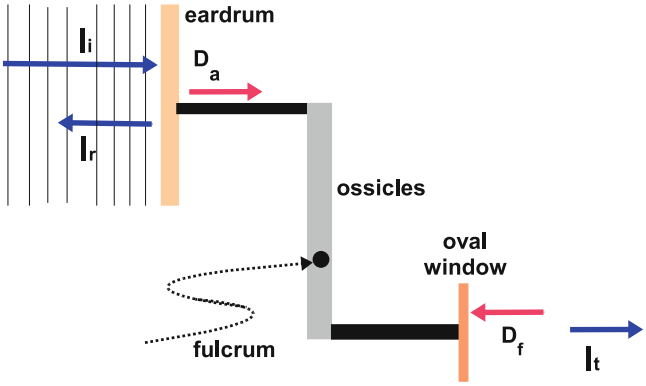


Fig. 9.8 Schematic of the operation of the ossicles

If the incident wave falls directly on the boundary with a second medium, the two media have a common area and

$$T = \frac{P_\mathrm{t}}{P_\mathrm{i}} = \frac{I_\mathrm{t} A}{I_\mathrm{i} A} = \frac{I_\mathrm{t}}{I_\mathrm{i}}. \tag{9.13}$$

Since the areas are different, we will have

$$T = \frac{P_\mathrm{t}}{P_\mathrm{i}} = \frac{I_\mathrm{t} A_\mathrm{w}}{I_\mathrm{i} A_\mathrm{d}}. \tag{9.14}$$

Instead of conservation of energy being represented by $I_\mathrm{t} + I_r = I_\mathrm{i}$, we have

$$P_\mathrm{t} + P_\mathrm{r} = P_\mathrm{i}. \tag{9.15}$$

When a wave in a medium is incident upon a boundary with another medium, the amplitude of the displacement of the waves in the media must be equal at the boundary. In Fig. 9.8, the displacements D_a and D_f are the respective amplitudes of the displacement of air and fluid, respectively. With direct contact of the eardrum with the oval window of the cochlea, we would need

$$D_\mathrm{a} = D_\mathrm{f}. \tag{9.16}$$

However, as a result of the lever action:

$$D_\mathrm{a} = \mathbf{r} D_\mathrm{f}. \tag{9.17}$$

Next, if there is a reflected wave, the wave in the first medium consists of two displacement amplitudes added together – one from the incident wave and one from the reflected wave. We have

$$D_{\mathrm{a}} = D_{\mathrm{i}} - D_{\mathrm{r}}. \tag{9.18}$$

Note the minus sign because the reflected wave is reversed in direction as when a wave reaches the end of a closed pipe. (See Sect. 3.1.)

The only remaining relation that is needed to derive the final result in (9.11) is the following relation between the intensity and the amplitude of displacement:

$$I = \frac{1}{2}(2\pi f)^2 Z D^2. \tag{9.19}$$

9.6 The Cochlea

The **cochlea** (or **inner ear**) is shaped like a snail. (See Fig. 9.1.) It is by far the most complicated part of the ear. No more than about one-half centimeter in width and almost entirely buried within a bony mass of the skull, it makes use of an intricate apparatus whereby the final conversion of the sound wave into nerve impulses is made. It also achieves a **partial frequency analysis**, which contributes to our ability to discriminate pitch.

The cochlear chamber is a spiral about 35 mm in length. Within is contained but a few drops of a liquid which has about the same density (1.03 g/cm^3) as water and about twice the 'thickness' (technically referred to as the **viscosity**) of water. In Fig. 9.9, we see a schematic of the interior of the full length of the cochlea if it were uncoiled. The center of the coil would be at the left. The chamber is divided into three sub-chambers – the **scala vestibuli**, the **scala tympani**, and the **scala media**. Only the first two are shown in this figure, for simplicity. Between the chambers is the **basilar membrane**. It has a width that varies from 0.08 mm near the oval window to 0.5 mm near the **helicotrema**. The basilar membrane serves as one of the two partitions between the three chambers.

In Fig. 9.10, we see a cross section across the length of the cochlea. This figure reveals greater details, including **Reissner's membrane**, which is the partition between the scala tympani and scala vestibuli. The basilar membrane is set into motion by the motion of the cochlear fluid. It contains nerve endings within **hair cells**, which, through the motion of the basilar membrane, are stimulated into producing nerve signals that travel to the brain through the auditory nerve.

Experiments have revealed that when sound is exciting the ear, the oval window and round window move nearly in opposite directions (are nearly half a cycle out of phase). This indicates that the cochlear fluid moves essentially en-masse – that is,

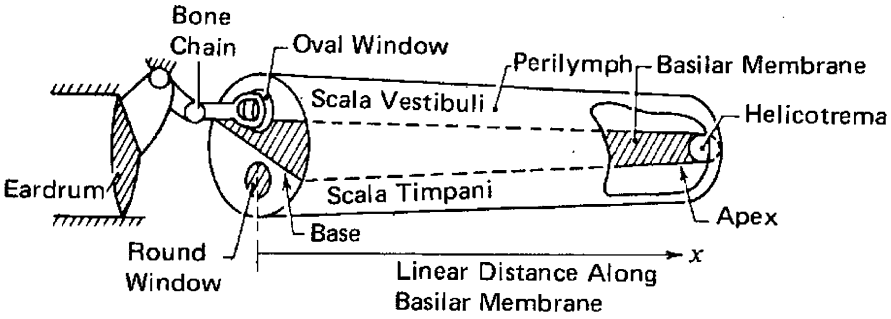


Fig. 9.9 Cochlea chamber, omitting the full complexity of the upper chamber (source: Roederer, op. cit.)

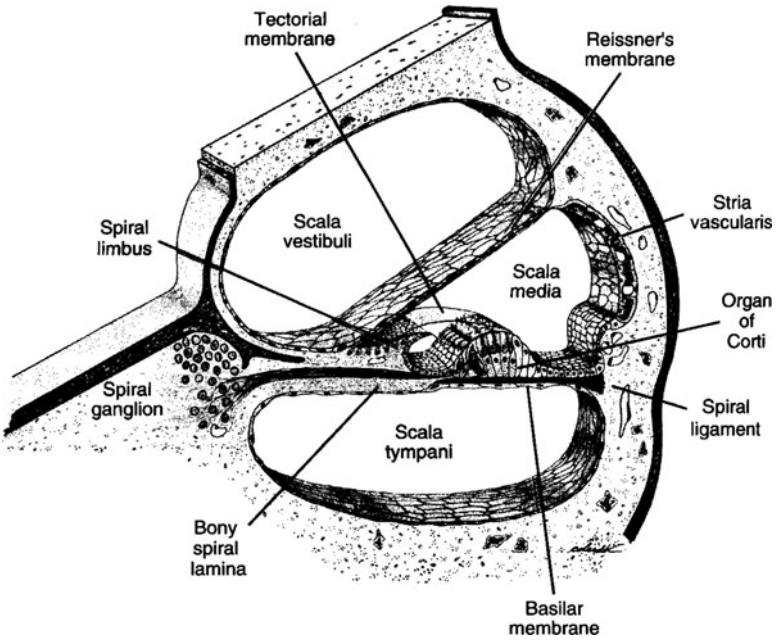


Fig. 9.10 Details of the cochlea (source: Roederer, op. cit.)

that there is negligible compression of the fluid. Thus, with negligible delay, while the oval window is moving to the right, the fluid in the scala vestibuli is moving to the right, fluid is flowing downward through the **helicotrema** and to the left in the

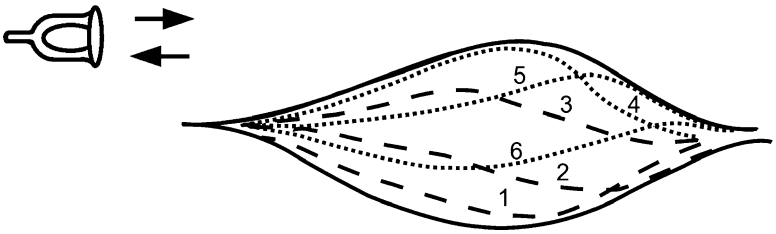


Fig. 9.11 Cochlea membrane wave with the envelope of the pulse

scala tympani, and the round window is moving to the left. Directions are, of course, reversed when the oval window is moving to the left.[3]

Extensive experiments have been performed so as to study the detailed motion of the fluid and the basilar membrane when the footplate is set into oscillation. The pioneer in this field of research was **Georg von Békésy**, who received the Nobel Prize in Medicine in 1961 for his work.[4] The picture that has evolved follows:

As was pointed out above, the fluid in the scala tympani moves in a direction opposite to that of the fluid in the scala vestibuli. The moving fluid exerts a frictional force on the membrane, in opposite directions on the two sides of a given point on the membrane. As a result, the membrane is displaced in a direction normal to the membrane. In effect, the fluid sets up a transverse vibrational motion of the membrane.

In Fig. 9.11, the dashed curves represent the shape of the membrane at a number of different stages of a complete sinusoidal cycle of motion of the stirrups.

Curves 1–3 correspond to moments when the stirrup is moving to the left, while curves 4–6 correspond to moments when the stirrup is moving to the right. The solid curve is the **envelope**, which is the smallest smooth curve within which are contained all the curves representing the shape of the membrane at all stages of a complete cycle.

The basilar membrane is lined with two rows of **hair cells**, which are connected to **nerve fibers**. (See Fig. 9.12.) *Physiology of the Ear*, A.F. Jahn and J. Santos-Sacchi Editors [Raven Press, N.Y. 1988] These fibers merge to form the **auditory nerve**, which leads to the brain. In the absence of sound, the nerve fibers emit impulses to the brain spontaneously and randomly.

If a sound is present, the relative motion of the basilar membrane is such as to produce a *bending* of the hair cells. This leads to an increase in the rate at which impulses are emitted by the nerve fibers and the perception of sound.

---

[3]The physical basis for the fact that the fluid moves essentially en-masse is that the wavelength of a sound wave in the cochlear fluid ranges from 75 mm to 75 m for audio frequencies, and is therefore at least twice the length of the cochlear chamber.

[4]See Georg von Békésy, *Experiments on Hearing* (McGraw-Hill Co., Inc., N.Y.,1960) which is an extensive treatise on the subject. His work has been of utmost importance in treating people with hearing difficulties.

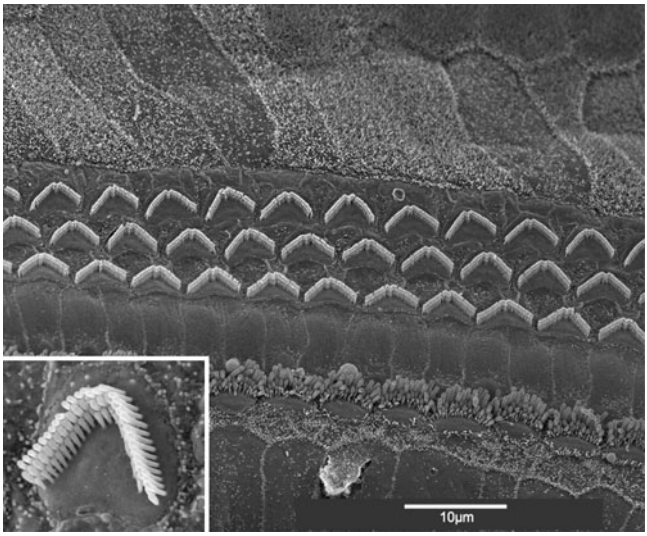


Fig. 9.12 Scanning electron micrograph of hair cells. Sensory epithelium of the **organ of Corti**. *D* Deiter's cells, *IHC* inner hair cells, *OHC* outer hair cells, *P* pillar cells (source: courtesy of Andrew Forge, University College of London Ear Institute)

As the sound intensity increases, two changes occur which lead to an increase in the rate at which impulses are emitted. First, the **amplitude** of the motion of the basilar membrane increases, resulting in an increase in the rate of impulse emission by *each* nerve fiber. Second, the envelope of the motion of the basilar membrane widens, resulting in an increase in the *number of hair cells* which are significantly bent and whose nerve impulse emission rate is increased. The increased rate of impulse emission results in an increase in the loudness of the sound.

9.6.1 Summary

We have seen how the eardrum and a set of three bones, which are far from beautiful in appearance nor simple in structure, serve to significantly increase the sound wave energy transmitted to the cochlea. Their beauty lies in their utter detail in the presence of puny size. As one grows older, these bones tend to become calcified (i.e., have excess calcium) and lose their flexibility at their joints. Consequently, the ossicles become an impediment to hearing and an operation is sometimes performed to allow the sound wave to bypass the ossicles and flow directly to the oval window. A person then suffers a severe hearing loss because the sound wave has to be transmitted to the air behind the eardrum and then on to the oval window; therefore one loses the benefits of the ossicles as discussed above. An alternative is to have

a microphone produce an amplified electronic signal that excites the nerves using a wire that is inserted into the cochlea. The device is called a **cochlear implant**. See Fig. 9.4. See comment on their operation later on in this chapter.

9.7 Pitch Discrimination

A sense of pitch requires an ability to discriminate frequencies – technically referred to as **pitch discrimination**. Specifically, tones of different frequency must affect the ear and/or brain differently. *All nerve impulses are alike.* As a result, there are only two ways in which auditory signals can be differentiated: *first*, by the specific nerve fibers which are transmitting the impulses and *second*, by the way the pattern of nerve impulses transmitted to the brain varies in time. There is strong evidence that the ear makes use of both approaches. The first approach is referred to as the **Place Theory of Pitch Perception**. We will refer to the second approach as the **Rhythm Theory of Pitch Perception**. Since both are operative, it is misleading, though unfortunately common, to refer to them as competing theories.

We begin with a discussion of the Place Theory of Pitch Perception, an idea that was first developed by **Hermann von Helmholtz** (Fig. 9.13).[5] Helmholtz pictured the basilar membrane as similar to a harp, consisting of a set of strings under tension, stretched transverse to the length of the membrane about a century ago.

The strings differed in length and in tension, and therefore differed in their fundamental frequencies. Then, a sound of definite frequency would excite only those strings having a fundamental frequency or an overtone frequency which is very close to the frequency of the sound wave. Very few strings would be excited, so that we would have a one-to-one correspondence between the sound frequency and the small set of strings excited. That one-to-one correspondence would be transmitted to the brain by the nerve fibers which were assumed connected to the individual strings, one-to-one, hence providing us with a sense of pitch.

Experiments by von Békésy and others showed that Helmholtz's idea was not far from the truth. In particular, von Békésy studied how the shape of the **envelope of the waves** (see Fig. 4.9) traveling along the basilar membrane varies with the frequency of pure tones. This variation in the shape of the envelope is depicted in Fig. 9.14.

We note that for frequencies above about 50 Hz, the envelope has a peak at some point along the membrane. The higher the frequency, the narrower the peak and

[5]Hermann Ludwig Ferdinand von Helmholtz (1821–1894), German physicist, anatomist, and physiologist. He worked on acoustics, hydrodynamics, electrodynamics, thermodynamics, meteorology, optics, non-Euclidian geometry and philosophy of natural sciences. He is known for his invention of the first ophthalmoscope, used by physicians to look into one's eye. In 1847 he formulated (independently of **Julius Robert Mayer** and Joule) the law of conservation of energy. Very often more than one scientist independently makes essentially the same discovery about the same time. Egos can lead to competition, arguments, public battles, and disappointment. Even the great **Newton** tried to blot out the name of **Gottfried Wilhelm Leibniz** for the latter's co-discovery (along with Newton) of the **differential and integral calculus** in mathematics . See the sad but

Fig. 9.13 Hermann Ludwig Ferdinand von Helmholtz (source: http://en.wikipedia.org/wiki/Hermann_von_Helmholtz)

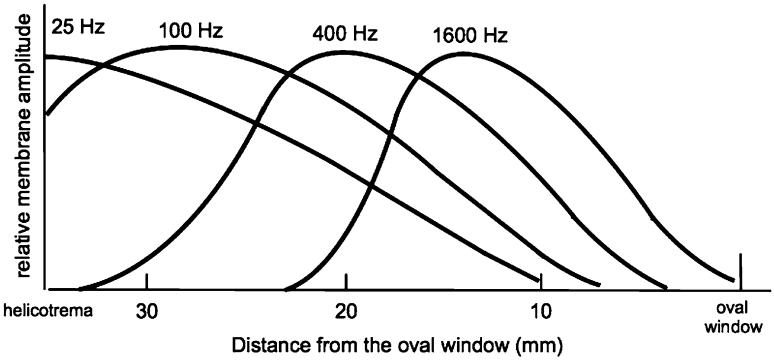


Fig. 9.14 Human basilar membrane response for various frequencies (figure based upon von Békésy, op. cit.)

the closer the peak is to the stirrup (stapes). In particular, it has been shown that the **distance** x_p of the peak from the far end of the membrane at helicotrema is approximately **proportional to the logarithm of the frequency**.

In Chap. 11, it will be shown that **pitch is essentially proportional to the logarithm of the frequency**. Therefore, the distance x_p is proportional to the pitch. Since those hair cells and attached nerve fibers which lie in the peak are most strongly stimulated, a sense of pitch is transmitted to the brain.

---

fascinating history of Mayer's work and his frustration from lack of recognition in the following website: http://www.uh.edu/engines/epi722.htm.

Fig. 9.15 Ernst Mach (source: http://en.wikipedia.org/wiki/Ernst_Mach)

There are deficiencies with the above theory. Our sense of pitch is very keen. We can discriminate frequencies which differ by only about 1%. On the other hand, the peak of the envelope is wide compared to the size of a hair cell, so that it is difficult to understand how the brain can discriminate between two envelopes which differ by only about 1%. The situation gets progressively worse as we get to lower frequencies. In fact, there is hardly any peak for frequencies below 50 Hz. Two solutions to these difficulties have been suggested[6]:

1. The first is an application of **Mach's Law of Simultaneous Contrast in Vision**, due to **Ernst Mach**[7] (Fig. 9.15) to hearing.
2. The second is the **Rhythm Theory of Pitch Perception**.

[6]There is a third factor that we mention here without details. Recently, Dennis Freeman has conducted research that reveals that the tectorial membrane (see Fig. 9.10) plays a considerable role in amplifying the effect of the frequency-dependent envelope in pitch discrimination. See MIT's *Technology Review*, Volume III, number 1, page M16.

[7]Ernst Mach (1838–1916), Austrian physicist and philosopher. The basis of Mach's natural philosophy was that all knowledge is a matter of sensations, so that what people call "laws of nature" are only summaries of experience provided by their own fallible senses. He discovered that if a body moves through the air at a speed faster than the speed of sound, it must produce a shock wave. The so-called Mach number is the speed of a body relative to the speed of sound. Thus, a speed of "Mach 3" is equal to three times the speed of sound in air.

9.7.1 *Some Mathematical Details on Pitch vs. the Peak of the Envelope***

What of Helmholtz's strings? Recall (see Chap. 2) that the fundamental frequency of a string is given by $f_1 = v/2R = (\mathcal{T}/\mu)^{1/2}/2R$, where, in particular, $\mathcal{T}$ is the tension. Tension, if present in the membrane, is too small to be measurable to date. It appears that the only (or primary) restoring force is the **stiffness** (i.e., resistance to stretching), so that $f_1 = v \times (\text{stiffness})^{1/2}$. In fact, Békésy showed that the logarithm of the stiffness was approximately proportional to the distance x_p; that is,

$$x_p \propto \log(\text{stiffness}). \tag{9.20}$$

Now suppose we ignore the coupling along the membrane and treat the membrane as a set of rods of equal length and cross-section but of varied stiffness to bending. The fundamental frequency of vibration f_1 of the rods can be shown to be proportional to the square root of the stiffness:

$$f_1 \propto \text{stiffness}^{1/2}, \tag{9.21}$$

so that

$$\text{stiffness} \propto {f_1}^2. \tag{9.22}$$

It follows that

$$x_p \propto \log({f_1}^2) = 2\log(f_1). \tag{9.23}$$

Thus the distance x_p is proportional to the logarithm of the fundamental frequency of the string and therefore to the sense of pitch. Finally, we can qualitatively account for the experimental results of Fig. 9.11 using this simple model of uncoupled strings, thus qualitatively confirming Helmholtz's original picture.

The variation in the width of the membrane (from about 0.1 mm at the stirrup to about 0.5 mm at the helicotrema) will further increase the variation of the frequency f_1 with x_p. This effect does not make the model invalid. Calculations using models of real membranes with the above varied *stiffness and width* confirm the basic validity of the above model.

9.7.2 *Mach's Law of Simultaneous Contrast in Vision*

Mach's Law of Simultaneous Contrast in Vision is based on the hypothesis (confirmed by experiment) that the nerve fibers emanating from different receptors on the retina of the eye (see Chap. 12) are not entirely independent. It is well known that a nerve impulse emitted by a receptor inhibits other receptors that

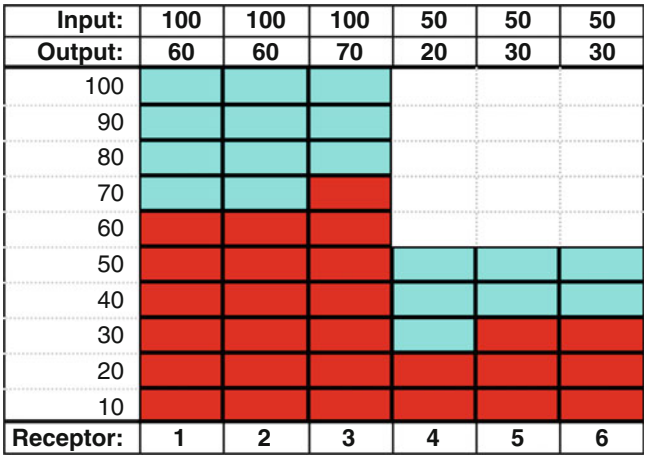


Fig. 9.16 Input and output of a string of six coupled receptors

are in its immediate vicinity from emitting nerve impulse. Consider then a simple situation wherein half of a region of mutually inhibiting receptors on the retina is uniformly stimulated by light, while the other half is stimulated at a lower intensity. In Fig. 9.16, we display the input and output of a string of six receptors. The inputs are 100 units for the first three and 50 units for the next three. The outputs are determined by the following inhibition: The output is equal to the input minus one-fifth the sum of the inputs of the two neighboring receptors. Thus, the output of the third receptor equals $100 - (100 + 50)/5 = 70$. The output of the fourth receptor equals $50 - (100 + 50)/5 = 20$.

We notice how the output exhibits a relative enhancement to the left of the boundary between the two regions (between #3 and #4) and a reduction to the right of that boundary. Without inhibition, the ratio of the outputs of the two neighboring receptors at the boundary is $100/50 = 2$. With inhibition, the ratio of the outputs of the two neighboring receptors at the boundary is $70/20 = 3.5$. Thus, the contrast between the two regions is increased at their boundary. Figure 9.17 illustrates the effect of the Law of Simultaneous Contrast in Vision quite dramatically in the so-called Mach bands. The darkness seems to increase dramatically close to the left side of a given rectangle. In fact, each given band is uniform.

If we assume that the auditory nerves on the basilar membrane interact in the above inhibitory manner, it can be shown that the peak in the wave envelope leads to a pattern along the membrane of the variation of the rate of nerve impulses which is more sharply peaked.

Mechanical explanations have also been proposed for a sharpening of the response curve. They are based on the idea that it is the *relative* displacement of the basilar membrane that produces a bending of the hair cell and a consequent

Fig. 9.17 Mach bands

stimulation of the nerve fibers. That is, a hair cell is stimulated to produce nerve impulses not on the basis of the degree of bending itself but rather on the basis of the *variation* of the degree of bending along the membrane.[8]

9.7.3 Rhythm Theory of Pitch Perception

We now turn to the Rhythm Theory of Pitch Perception. This theory is based on the experimentally proven fact that an increase in the nerve impulse rate occurs only during that part of a cycle of motion of the basilar membrane when the basilar membrane is moving toward the tectorial membrane. If a sinusoidal sound wave of frequency f is incident upon the ear, there will thus be an alternation at a frequency f between periods when the rate increases and periods when the rate does not increase. This situation is depicted in Fig. 9.18, wherein the vertical spikes of Fig. 9.18 represent the instants when a nerve impulse is emitted. We see from the figure that there is a resulting periodicity in the rate at which nerve impulses are emitted, which provides us with a sense of pitch.

What is the resulting basic understanding of pitch discrimination by the ear? The above **Place Theory** accounts for pitch discrimination of frequencies above about 4,000 Hz. Both the Place Theory and the Rhythm Theory are operative for frequencies between about 50 Hz and about 4,000 Hz. For frequencies below about 50 Hz, only the Rhythm Theory is operative. Over the years, there have been alternative theories; nevertheless, they are in essence enhancements of the above theories. Finally, we should note that a **cochlear implant** bypasses the outer and middle ears in having sound impinge upon a microphone which sends an electronic

[8] An analogy can be drawn between the **time** variation of the **displacement**, **velocity**, and **acceleration** of an automobile on the one hand, and the **spatial** variation along the length of the basilar membrane of its **displacement**, its **degree of bending**, and the above **variation** of the degree of bending, on the other hand.

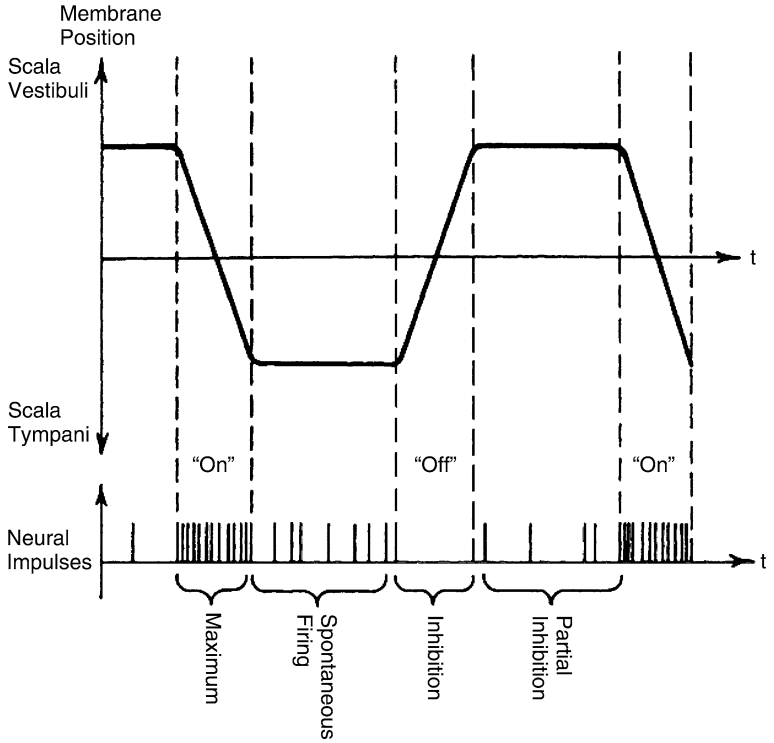


Fig. 9.18 Rhythm theory with nerve spikes (source: Roederer, op. cit.)

signal down a wire into the cochlea; this signal excites the nerves at the site of the hair cells. The individual wires, with corresponding sites in the cochlea, can number about five to ten. Therefore, there is quite limited pitch discrimination.

9.8 Terms

- Auditory canal
- Basilar membrane
- Cochlea
- Eardrum (tympanic membrane)
- Fulcrum
- Hair cells
- Law of Simultaneous Contrast
- Lever action
- Mach bands
- Mechanical advantage
- Ossicles
- Oval window
- Pitch discrimination
- Place theory of pitch perception
- Rhythm theory of
- Pitch perception

9.9 Problems for Chap. 9

1. Mechanical vibrations are transformed to nerve impulses in the

   a. Middle ear
   b. Eustachian tube
   c. Auditory canal
   d. Semicircular canals
   e. Cochlea

2. The correct order in which mechanical vibrations pass through the parts of the ear is

   a. Eardrum, cochlea, hammer, stirrup, and oval window
   b. Eardrum, hammer, anvil, stirrup, and cochlea
   c. Eardrum, Eustachian tube, oval window, and round window
   d. Eardrum, anvil, Eustachian tube, and organ of Corti
   e. Eardrum, cochlea, basilar membrane, and middle ear

3. What is the function of the ossicles of the ear?
4. What are the two theoretical mechanisms for pitch discrimination provided by the ear? Describe them in detail.
5. What are "Mach bands" and how do they relate to theories of pitch discrimination?
6. A mother of weight 120 lbs wants to balance herself against the weight of her child whose weight is 40 lbs. They are seated on opposite ends of a seesaw whose overall length is 12 ft. See Fig. 9.6. Where should she place the fulcrum? Start by determining the ratio of the distances from the fulcrum to each of the two of them.
7. Below is a table that is meant to illustrate the effect of Simultaneous Contrast. We have a set of inputs for a string of 14 receptors. The inputs are meant to represent the envelope of the wave that is travelling along the basilar membrane, with a broad peak at the ninth receptor.
   Calculate the respective outputs of the receptors using the recipe discussed in the text. That is, if we let I_n be the input to receptor n and O_n be the output from receptor n,
   $$O_n = I_n - (I_{n+1} + I_{n-1})/5. \tag{9.24}$$
   Produce two graphs – one for the input vs. the receptor number (1–14) and one for the output vs. the receptor number (1–14). Draw them, one below the other to emphasize the difference. Your result should indicate an enhancement of the peak of the response.

Receptor	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Input	20	22	24.5	26	29	33	37	40	42	39	36	33	31	30
Output														

Chapter 10
Psychoacoustics

In the last chapter, we learned about the last step in the trail from the source of sound to the brain via nerve signals from our ears. All these steps have been describable in what we refer to as *physical terms* and are *objective*. And yet we do not know in *physical terms* what it means when we say, "**I hear a sound**". This last step has eluded explanation and clarification. What is the *physical* nature of pain? Perhaps we are asking the wrong questions. Perhaps we are limited by language, which is after all a product of our conscious experience, and are therefore looking for the answer to a question that has no meaning and therefore is, shall we say, invalid as a question.[1]

The physicist knows how to characterize a sound *uniquely*. And yet, how people describe a given sound that is heard varies from one individual to another. Of course, a report must make use of the language of the person in relation to his/her own personal experience. Two people learn to associate the color red with a certain sensation they experience. What their individual experiences are, we do not know. Both may report that a given color is red, but what they actually experience may differ. We may both agree that two tones have the same or different pitch, but will never know the extent to which our perceptions of the sounds are similar.

In spite of the above difficult issues, we can ask people certain questions regarding their perceptive response to various sounds. For example: Which of two sounds is louder? Or, which of two sounds has a higher pitch? Or, sing a tone with a frequency of 440 Hz, recognizing that their ability to do so depends upon their previous exposure to this tone. Tests of individuals by psychologists have resulted, as you would expect, in a broad distribution of responses. Nevertheless, psychologists have summarized their results in terms of normal responses, and it is in these terms that we will discuss some of the characteristics of human

[1] Such a situation exists in the regime of phenomena for which the specific nature of Quantum Theory is manifest. We have found, for example, that it is impossible to describe an atom in terms of images that we have amassed for describing the world at the macroscopic level. For further ideas into the issue of the nature of perception and what it means to think or feel, see the fascinating book by **Daniel Dennett Consciousness Explained**, [Penguin Press, UK, 1992].

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_10,

psychoacoustics, which is the study of the relationship between the objectively characterizable sound incident upon a human ear and the corresponding perception of the sound.[2]

Note

Before we get into the subject I need to point out that the subjects covered in this chapter are extremely limited in relation to the incredibly wide range of areas of study of psychoacoustics. The chapter focuses on a few subjects: (1) measures of loudness vs. intensity and frequency; (2) "combination tones": a phenomenon wherein we hear frequencies that are not actually present in the sound incident upon our ears due to the nonlinear response of our ears; (3) duration of a note needed for pitch discrimination; and (4) "fusion of harmonics": The sound of a musical instrument generally has a large ensemble of harmonics. Nevertheless, we do not hear the individual harmonics; rather, the sound appears to have one source.

Both in this chapter on psychoacoustics and in Chap. 14 on color vision, which is a branch of psycho-optics, we will focus on the basic elements of sound and color, respectively: the musical note, with its multitude of timbres and the color patch. We represent them in Fig. 10.1 for us to focus our minds on these elements.[3]

It is interesting to compare these two images. The color patch displays a hue (red, green, ...), saturation (degree of paleness), and brightness – the details about which you can read about in Chap. 14. On the other hand, the note symbol on the staff of a musical score has by itself only the content of the frequency chosen for the "A," which is currently typically 440 Hz. This component of information is the analog of the hue of the color patch. However, the actual sound produced by a musician has a huge range of timbres, vibratos, and dynamics that can each be varied throughout the duration of the note played. It is reasonable to compare these components of the note with the relatively small range of degrees of saturation. On the other hand, the level of research being carried out with respect to a color patch is many orders of magnitude greater than that for the performance of individual notes. In parallel is the relative ease we have in remembering the appearance of a color patch in comparison with the characteristics of a performed note. Perhaps, we will learn how to pay more attention to these varied sound characteristics and find a clearer, more precise way to characterize their differences.

[2]The following Wikipedia website is a useful resource of links to many psychoacoustic phenomena: http://en.wikipedia.org/wiki/Psychoacoustics.

[3]The note is an A with a typical frequency of 440 Hz and a color patch with the color that the composer-pianist Alexandre Scriabin associated with the note A.

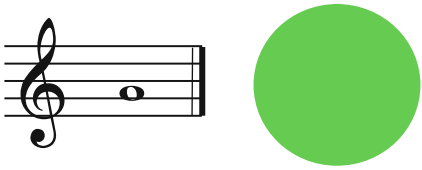


Fig. 10.1 A musical note – an A440 – representing an element of a piece of music; a color patch – representing an element of color

10.1 Equal Loudness Curves

For normal hearing, increasing the intensity always increases loudness. What is the quantitative relationship between intensity and loudness? To answer this question would require, for example, that we be able to clearly determine when one sound is twice as loud as another, a requirement that is impossible to meet. There are less demanding questions one could investigate. For example, if a person is exposed to two sounds of different frequency but equal intensity, they generally report that their sense of loudness of the two sounds differs. The person can then be asked to match the loudness of one tone by varying the intensity of the second tone. Back in the 1930s, extensive tests were carried out and resulted in the "so-called equal loudness curves." A more recent set of curves is shown in Fig. 10.2. We begin by discussing the significance and highlights of these curves.

We note that the vertical axis is marked "**intensity level**" (what we have referred to as the "sound level") in dB, while the frequency is marked along the horizontal axis. Points P and Q refer to two tones at frequencies of 100 and 400 Hz, respectively, and lie along the same equal loudness curve. According to the curve, 100 Hz at 68 dB will sound equal in loudness as 400 Hz at about 57 dB.

The **phon**, symbol ϕ, is a unit used to label the equal loudness curves. Suppose that a sound of given frequency has a certain intensity I (with a corresponding **sound level – SL**). To determine the number of phons for that sound, we look along the corresponding equal loudness curve for the intensity and SL for a frequency of 1,000 Hz. The number of phons for the given sound is equal to the SL for a frequency of 1,000 Hz.

For example, the two points **P** and **Q** both lie along the equal loudness curve labelled "60 phons" (on the 1,000 Hz vertical axis). They correspond to the frequencies 100 and 400 Hz, respectively. The number "60" refers to the sound level of a 1,000 Hz sound at the given loudness. Zero phons corresponds to the threshold of hearing for all frequencies while 120 phons corresponds to the threshold

For the reader who is interested in a more comprehensive discussion of psychoacoustics, I highly recommend the book by **Juan Roederer**, *Introduction to the Physics and Psychophysics of Music – 4th ed.* [Springer-Verlag, New York, 2008].

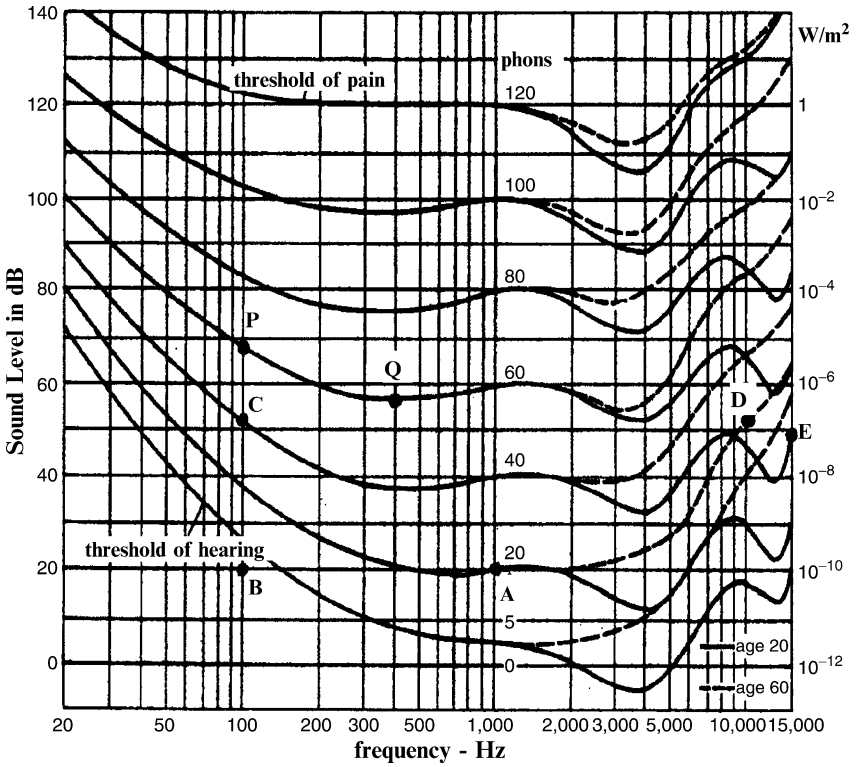


Fig. 10.2 Equal loudness curves (source: H.E. White and D. H. White, *Physics and Music*, (Holt, Rinehart, and Winston, Philadelphia, 1980))

of pain. The solid curves correspond to individuals at age 20, while the dashed curves correspond to individuals at age 60. We note that the hearing loss due to old age lies mostly in the high frequency range of 2,000 Hz and above.[4]

The most outstanding qualitative feature of the equal loudness curves is that the ear is most sensitive to frequencies lying between about 1,000 and 4,000 Hz. Typically, the maximum sensitivity lies at about 4,000 Hz. This fact accounts for the following interesting aspects of music: First, while the bass drum has a power that can far exceed the power output of any other instrument in an orchestra, it does not drown out the orchestra because of its low frequency range. Second, professional singers, who must project a loud sound in spite of the relatively low range of their voices (especially that of basses and baritones) and a lack of amplification

[4]There is a revised international standard (ISO 226 2003) to be found on the following website (1-22-2011): http://en.wikipedia.org/wiki/FletcherMunson_curves. I kept the older figure below because it includes the effect of aging. One reason for the difference are improved testing procedures.

in the performance hall, as can be the case in opera, can do so by singing with what is called a **squillo**. Effectively, their voices contain a high amplitude of higher harmonics lying in the above range – from 2,000 to 4,000 Hz if possible. In the **Estill method of voice development and theory**, the squillo is referred to as **twang**.

Finally, we should note that the region characterized by the equal loudness curves is bounded at the bottom and at the top. At the bottom, we have the curve referred to as the **threshold of hearing**. For each frequency, we see the lowest possible intensity that is audible. The upper boundary is called the **threshold of pain**, above which sound is regarded as being "painful" and masks a sense of pitch. These threshold of hearing is discussed in greater detail in Appendix H.[5]

10.2 The "Sone Scale" of Expressing Loudness**

It is well recognized that doubling the intensity corresponds to much less than a doubling of the sensation. Therefore, intensity is a very poor measure of loudness. The scale of **sound level**, which is logarithmic in the intensity, better represents our sense of loudness in which doubling the sound level reflects more accurately a doubling of the loudness than does a doubling of the intensity. However, this improvement **over**-compensates the inaccuracies of using the intensity as a measure of loudness. To obtain an even more accurate measure of loudness, the **phon** is sometimes mapped onto another parameter called the **sone**, which we will represent by the letter **s**.

The relation between the **sone** and the **phon** is shown in Fig. 10.3 as well as being represented in Table 10.1. Note that the vertical scale is not linear; it is logarithmic. All segments, running from 1 to 10 or from 10 to 100, and so on, are equal in length in the graph. Over the course of each segment, the loudness increases in sones by a factor of 10 (rather than adds to the loudness in sones by a fixed number).

The solid curve on the graph represents the sone value for all values of phons. The mathematical relation between sones and phons for $\phi > 40$ phons is given by

$$s = 2^{\frac{\phi - 40}{10}}. \tag{10.1}$$

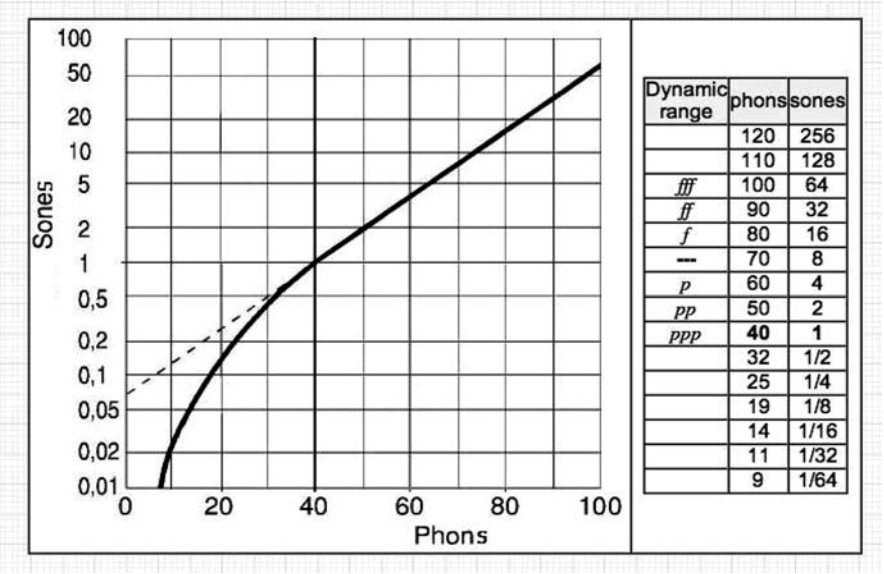
This relation produces a straight line in the figure, which is continued by a dashed curve for $\phi < 40$ for comparison with the actual behavior for this regime.

For $\phi < 40$,

$$s = (\phi/40)^{2.86} - 0.005. \tag{10.2}$$

We will restrict our discussion to the simple regime when $\phi > 40$ phons. The behavior is based on the observation of psychoacousticians that **any increase in**

[5] You can test your own hearing within the limitations of your level of training by using the applet on this website (1-22-2011): http://www.phys.unsw.edu.au/jw/hearing.html.



Dynamic range	phons	sones
	120	256
	110	128
fff	100	64
ff	90	32
f	80	16
---	70	8
p	60	4
pp	50	2
ppp	**40**	**1**
	32	1/2
	25	1/4
	19	1/8
	14	1/16
	11	1/32
	9	1/64

Fig. 10.3 Sones vs. phons; table relating sones to dynamics (loudness) symbols in musical scores (source: Sengpielaudio, (2-1-2011): http://www.sengpielaudio.com/calculatorSonephon.htm)

Table 10.1 Phons vs. sones

Phon level	40	50	60	70
Loudness in sones	1	2	4	8

the number of phons by ten doubles the loudness in sones. Thus, the loudness in sones will double if the number of phons increase from 40 phons to 50 phons or from 50 phons to 60 phons. The loudness in sones will quadruple of the number of phons increases from 30 phons to 50 phons. See a sample of values in Table 10.1.

It can be shown that the mathematical relation between the sone level and intensity is given by[6]

$$s = \left(\frac{I'}{I'_{40}}\right)^{0.3}. \tag{10.3}$$

Here I' is the intensity of a tone at 1,000 Hz that is equal in loudness to that of the given tone. I'_{40} is the intensity of a 1,000 Hz tone that has a sound level of 40 dB, that is, 10^{-8} W/m^2. The sone level increases more slowly than linearly but more rapidly than logarithmically with respect to the intensity I'.

[6]The number 0.3 in the equation is actually an approximation for $\log\ 2 = 0.3010\ldots$.

Note

Both the number of phons and the sound level in sones that correspond to a given sound wave depend not only upon the intensity, but also upon the frequency!

The relationships described above are quite complicated. Therefore, we will present as an example the following situation: Let us consider a sound of frequency 200 Hz and sound level 30 dB. According to Fig. 10.2, the sound corresponds to about 23 phons. From (10.1) we find that the loudness is

$$s = 2^{\frac{(23-40)}{10}} = 2^{-1.7} = 0.31 \text{ sones}. \tag{10.4}$$

Suppose that we want a sound that is doubly loud, that is $s = 2 \times 0.31 = 0.62$ sones. Let us determine the required number of phons.

$$s \equiv 2^{\frac{\phi-40}{10}} = 2 \times 0.31 = 2 \times 2^{\frac{(23-40)}{10}}. \tag{10.5}$$

Then we have

$$2 = 2^{\frac{\phi-23}{10}}, \tag{10.6}$$

so that $(\phi - 23)/10 = 1$ and $\phi = 33$ phons.

Note that we could obtain this result more directly by recalling that doubling the loudness in sones requires an addition of 10 phons to the phon level. Then we require $\phi = 23 + 10 = 33$ phons.

To repeat,

Generally, in order to double the loudness, measured in sones, one must add 10 *phons to the phon level.*

Sample Problem 10-1

What is the sound level corresponding to the above 200 Hz sound having 33 phons?

Solution

From Fig. 10.2, we find $SL = 37$ dB.

Note that the change in sound level is from 30 to 37 dB, corresponding to a change of 7 dB and an increase in intensity by factor of $10^{\Delta SL/10} = 10^{0.7} = 5$. In sum, in order to double the loudness of this particular sound, one must increase the intensity by a factor of 5.

The above results for a specific frequency of 200 Hz are summarized in Table 10.2. (The results generally depend upon the frequency.)

Table 10.2 Comparison: intensity, sound level (SL), phon level ϕ, and loudness in sones of a 200 Hz pure tone

Intensity I in W/m^2	Sound level SL	Phon level	Loudness in sones
10^{-9}	30	23	0.31
5×10^{-9}	37	33	0.62

10.3 Loudness from Many Sources[7]

Suppose that we have a number of sound sources. We address the question of how we would compute the resulting SL and the loudness in sones. To simplify the discussion, we will discuss only two sound sources; the generalization to more than two should be obvious.

There are a number of cases that distinguish the results:

1. The two sources have the same frequency and are coherent – that is, have a definite relative phase at the wave level. In this case, we determine the resulting amplitude as discussed in Chap. 7. Squaring the total amplitude gives us the resulting intensity, from which we can calculate the resulting phon level and then the resulting number of sones.

   As a simple example, we will consider two such sources that have the same frequency of 1,000 Hz, the same amplitude, and are in phase. The total amplitude is twice each. Therefore, the intensity is quadrupled, resulting in an increase of 6 dB and hence 6 phons because the frequency is 1,000 Hz. Therefore, the loudness in phons will change from ϕ to $\phi' = \phi + 6$. From (10.1), we see that the loudness in sones will increase by a factor of $2^{6/10} = 2^{0.6} \approx 1.52$.
2. The two sources are independent and have different frequencies. Let Δf be the magnitude of the difference in the two frequencies. In this case, there are two subcases that depend upon the two frequencies. We need to consider the **critical bandwidth**, which refers roughly to the range of frequency differences such that the two pitches cannot be distinguished.

   (a) Suppose that $\Delta f <$ critical bandwidth. Then we find the intensity of each of the individual sources. We add the intensities to obtain the total intensity. We then calculate the corresponding number of phons for the total intensity based on the average frequency. And finally, we calculate the number of sones based on this phon level. **Example**: As in the previous example, we assume that the frequencies are both close to 1,000 Hz and that both have one sone, which corresponds to 40 phons. If we double intensity, we add 3 phons, which results in a total of $40 + 3 = 43$ phons. Therefore, the resulting loudness in sones is $s = 2^{(43-40)/10} = 2^{0.3} = 1.23$ sones.

[7] See the website (1-9-2011): http://home.tm.tue.nl/dhermes/lectures/SoundPerception/05Loudness.html.

(b) Now suppose that $\Delta f >$ critical bandwidth. This case is simple, we simply add the number of sones. In the above example, we obtain 2 sones.

Sample Problem 10-2

Consider a swarm of 1,000 mosquitos, each producing a sound with SL=10 dB. Let us estimate the resulting sound level SL and the loudness in sones.

We can easily obtain the sound level since $\Delta SL = 10 \log 10^3 = 30$ dB. Therefore, the resulting sound level is $30 + 10 = 40$ dB.

To determine the loudness, sones is more complicated since we need to take into account the frequency spectrum of the sound. For simplicity, we will represent the spectrum by the most prominent frequency, which is about 300 Hz. From Fig. 10.2, we see that one mosquito's sound amounts to $\phi = 0$ phons. This corresponds to a loudness of $2^{-40/10} = 1/16$ sone. To compute the total loudness, we note from Fig. 10.2 that 40 dB at 300 Hz corresponds to about 42 phons. Thus, we obtain

$$s = 2^{(42-40)/10} = 2^{0.2} \approx 1.1 \text{ sones.} \tag{10.7}$$

Note

To close this section, we should note the following:
The intensity and the sound level are both **objective** measures of loudness.
On the contrary, the phon level and the loudness in sones are **subjective** measures of loudness in the sense that they are based on the results of testing the hearing of a number of individuals.

10.4 Combination Tones and the Nonlinear Response of the Cochlea

The phenomenon discussed in this section is quite unusual in which it is responsible for the perception of tones, the so-called combination tones. They are not at all present in the sound wave incident upon one's ears and may partially account for one's sense of musical consonance. In hi-fidelity terminology, it is responsible for **harmonic distortion**. In the early days of radio, it was a favorable characteristic of poor audio speakers: The radio could not respond well to low frequencies (say, about <400 Hz), so that the fundamental and possibly some overtones of a low-pitched tone of a musical instrument might be missing or very weak in the electrical

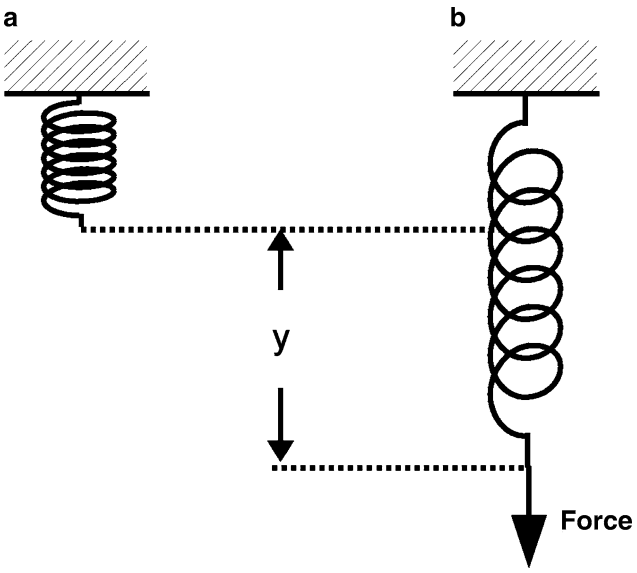


Fig. 10.4 Spring system

signal which excited the speaker. However, because of the **nonlinear response** of the speaker, the higher harmonics would cause the speaker to produce a tone with the missing fundamental and overtones included as components. The result might be likened to enriched, bleached flour.

To understand **nonlinear response**, it is helpful to first appreciate **linear response**, which we have thus far taken for granted. (Read Chap. 2 on the SHO, as a preparation for what follows.) Consider the spring illustrated in Fig. 10.4. A downward force is applied to the spring, which consequently increases in length by an amount y in Fig. 10.4.

If the spring were pushed upward by a force of the same magnitude, the spring would move upward by the same displacement, y. If the force is doubled, the displacement y will be doubled. Generally, the distance is proportional to the force.

$$\text{Displacement } y \propto F \quad \text{or} \quad y = \frac{F}{k}, \tag{10.8}$$

where k is the **spring constant**. If we graph (10.8), we obtain a straight line. **Positive** forces represent downward forces – hence a stretching of the spring and increasing y. **Negative** forces represent upward forces – hence a contraction of the spring and decreasing y. Because of the straight line graph, one says that the spring responds **linearly** to an external force. (The displacement is on the y-axis while the force is on the x-axis). The technical term describing this behavior is that there is **linear response**. (Alternatively, one says that "the response of the spring to the force is linear.")

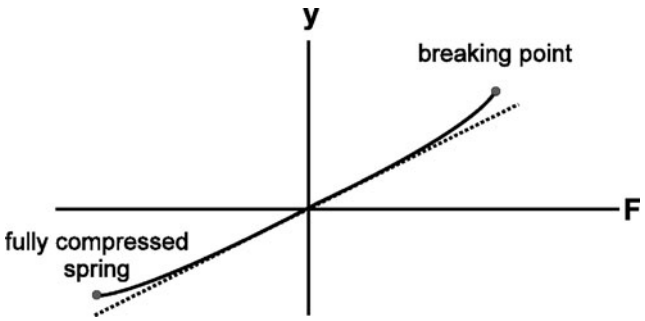


Fig. 10.5 Nonlinear displacement of a spring in response to a force

Suppose the force varies sinusoidally in time with a frequency f. Then the response, y, is just a sine wave multiplied by the constant $1/k$, and is therefore a sine wave of the same frequency f but with a different amplitude. Generally, whatever the pattern of variation of the force with time will be reproduced by the displacement.

Consider the process of a sound wave incident upon a microphone that is connected to an amplifier that is, in turn, connected to a loudspeaker. If there is linear response throughout, the pattern of the incident sound wave in time will be reproduced by the electrical signal from the microphone, by the electrical signal from the amplifier, and finally by the sound coming out of a loudspeaker. Thus,

linear response is associated with fidelity!

Now consider the response of a **real** spring as depicted in Fig. 10.5. The curve is not a straight line. The response is said to be **nonlinear**.[8]

In Table 10.3, we list a number of systems upon which a force or an analog of a force, a so-called generalized force is exerted. The quantity that measures the response of the system to the generalized force is listed in the right-hand column.

We now consider a special case known as **quadratic nonlinear response**. In this case, the displacement is proportional to the (force)^2, or

$$y = bF^2, \tag{10.9}$$

where F is the force and where b is a constant of proportionality.

[8]It is important to note that for small enough forces, the response of a real spring is essentially linear. That is, we can assume that the displacement of the spring is proportional to the force to a good approximation. Thus, the so-called ideal spring is an abstraction whose behavior is approached by a real spring for small forces. It is a remarkable fact that the bulk of physical theories and concepts are based on abstract models of the real world which assume linear response as an approximation, with deviations from linearity being second order effects which may or may not be essential to the phenomena of interest. Furthermore, the Principle of Superposition, which was discussed in Chap. 7, is dependent on a linear response of the system. Thus, to the extent that the response is nonlinear, this principle breaks down.

Table 10.3 Various systems – generalized force vs. the response

System	Generalized force	Response
Spring	Force on spring	Displacement of spring
Stereo receiver	Amplitude of radio wave	Electrical voltage output of stereo receiver
Speaker	Electrical voltage input to speaker	Displacement of the diaphragm of the speaker
Wishbone	Pull on a wishbone	Change in the angle between the "legs" of the wishbone

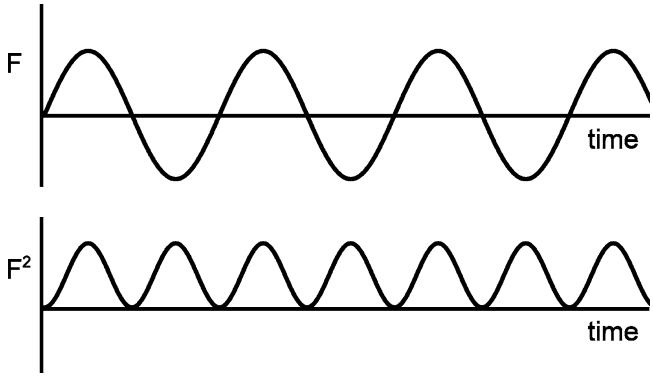


Fig. 10.6 A sine wave; the square of a sine wave

Suppose now that the force F is a pure sine wave of frequency f, as in Fig. 10.6. According to (10.9), the displacement will be the square of a sine wave and hence is always positive. The two functions, a sine wave and the square of a sine wave, are depicted in Fig. 10.6.

In Fig. 10.6, we see that the force F (the upper curve) is periodic and has twice the period of the square of the force F^2 (the lower curve). Alternatively, F^2 has double the frequency of F.[9] If the force has a frequency of 400 Hz, the output will be 800 Hz. If the response is a sum $y = aF + bF^2$, the output will be two frequencies, 400 and 800 Hz.

We get a more interesting result if the force is a mixture of two Fourier components, say, with frequencies f_1 and f_2. Let us represent the component forces by the symbols F_1 and F_2, respectively. Then the total force F is given by the sum of F_1 and F_2:

$$F = F_1 + F_2,$$

[9]This result is obtained from the trigonometric identity is $\sin^2\theta = 1/2 - (1/2)\cos 2\theta$.

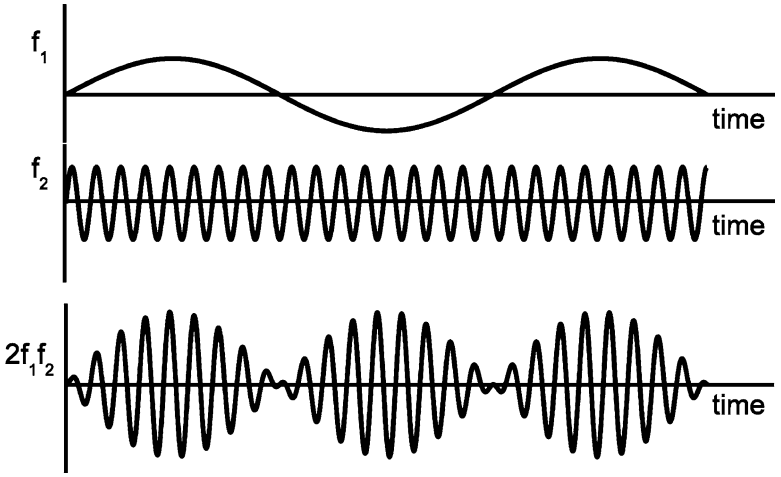


Fig. 10.7 Two sine waves, with $f_2 \gg f_1$, and the cross term of the square of their sum

while the square of F is given by

$$F^2 = (F_1 + F_2)^2 = F_1^2 + F_2^2 + 2F_1F_2.$$

According to our preceding result, F_1^2 has a frequency component of $2f_1$, while F_2^2 has a frequency component $2f_2$. Finally, we need to know the frequency composition associated with the product F_1F_2.

It can be shown[10] that the product $2F_1F_2$ has two components, with frequencies $(f_2 - f_1)$ and $(f_2 + f_1)$, respectively.

We can make this result somewhat plausible by considering the case when $f_2 \gg f_1$. Then, in the product F_1F_2, the factor F_1 produces an envelope of the sine curve F_2. The behavior of F_1, F_2, and the product $2F_1F_2$ are depicted in Fig. 10.7.

The pattern $2F_1F_2$ is that of the **beating** of two components, with a **beat frequency**

$$f_B = (f_2 + f_1) - (f_2 - f_1) = 2f_1.$$

In sum, if $y \propto bF^2$ and $F = F_1 + F_2$, y has Fourier components with frequencies

$$2f_1,\ 2f_2,\ (f_2 + f_1),\ \text{and}\ (f_2 - f_1).$$

Let us finally consider a second example of nonlinear response, wherein the displacement is proportional to cube of the applied force: $y = cF^3$, with c as a

[10]The result is based on the trigonometric identity: $2\sin\theta_1 \times \sin\theta_2 = cos(\theta_2 - \theta_1) - cos(\theta_2 + \theta_1)$.

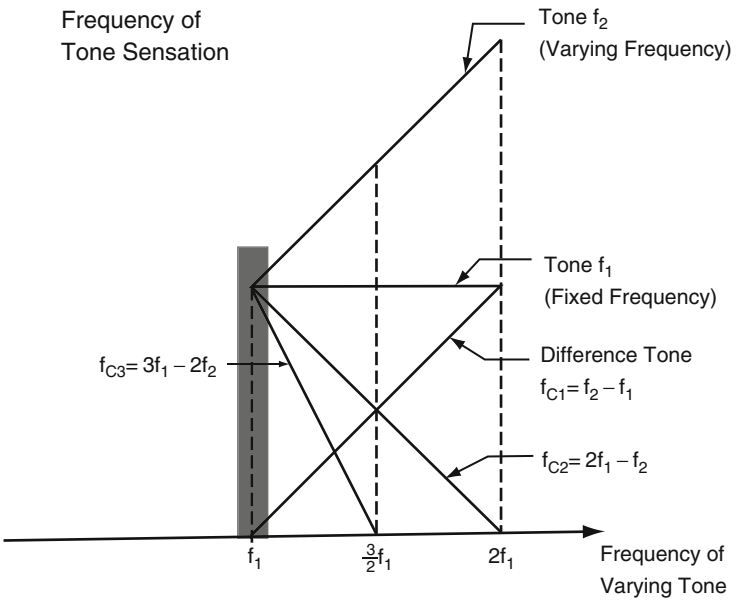


Fig. 10.8 Frequencies of combination tones (source: Roederer, op. cit.)

constant. Then, if $F = F_1 + F_2$, it can be shown that y has components with frequencies f_1, f_2, $3f_1$, $3f_2$, $2f_1 - f_2$, $2f_2 - f_1$, $2f_1 + f_2$, and $2f_2 + f_1$. (The reader can try to guess what result would obtain if $y \propto F^4$.)

Now that we have discussed how various nonlinearities in response are manifested by the outputs of two sine waves, we now summarize how nonlinearity manifested in the ear:

When a sound, having a mixture of two components, with frequencies f_1 and f_2, is incident upon the ear, tones having pitches corresponding to

$$\begin{aligned} f_{c1} &= f_2 - f_1 \\ f_{c2} &= 2f_1 - f_2 \\ f_{c3} &= 3f_1 - 2f_2 \end{aligned} \tag{10.10}$$

can be heard – in addition to the tones corresponding to the frequencies f_1 and f_2. These are the frequencies of the **combination tones**. Their existence indicates that the response of the ear is nonlinear. See Fig. 10.8, wherein these frequencies are graphed as a function of f_2, with f_1 kept fixed.

Experiments show that the basilar membrane within the cochlea responds nonlinearly to the fluid forces acting upon it. The fluid does respond linearly to the pressure exerted by the footplate at the oval window. Ultimately, the basilar membrane responds nonlinearly to the sound pressure p at the outer ear.

Now the sound pressure can be written as the sum of two components, corresponding to the two frequencies, f_1 and f_2:

$$p = p_1 + p_2. \tag{10.11}$$

The above perceived frequencies indicate that the response of the ear to the sound pressure is given by the sum

$$y = ap + bp^2 + cp^3 + ep^5, \tag{10.12}$$

where a, b, c, and e are constants. The first term accounts for the perception of the tones corresponding to f_1 and f_2, the second to $(f_2 - f_1)$, the third to $(2f_1 - f_2)$, and the fourth to $(3f_1 - 2f_2)$. A difficult question to answer is why the remaining frequencies, such as $f_2 + f_1$ or $2f_2 + f_1$, which are expected according to the above response relation between y and p, are not perceived.

10.5 The Blue Color of the Sea and Its Connection with Combination Tones

In Chap. 8, we discussed how the scattering of light by the air is responsible for our being able to see the blue sky. Certainly, when we look down into the sea, scattering of the light is responsible for our being able to see its blueness. We might therefore conclude that the preferential scattering of short wavelengths is the most important reason for the blue color of the sea. Evidence to the contrary is indicated by the fact that if you view sunlight from within the sea looking upward, the light will be blue – in contrast to the red appearance of the setting sun. The primary source of the blue color of the sea is the absorption by water of light in the red region of the visible spectrum. Figure 10.9 exhibits the blueness of water from above the surface as well as from below. The deeper you are in the water, the bluer the light will be because of the increased absorption of red.[11]

Interestingly, it was difficult to understand how water could exhibit strong absorption in this range of wavelengths since there are no corresponding energy level differences among the quantum states of water.[12] There are vibration modes in the infrared region; they are the key to the explanation. It turns out that the vibration of the atoms in a water molecule do not obey Hooke's law precisely. There is a *nonlinearity* in the interatomic forces. As a result, the spectrum of frequencies includes **combination modes** – analogous to the combination tones that we hear due to the nonlinear response of the ear! It has been shown that the strong absorption

[11] If you look down into the sea, the light you see is scattered light. However, the effect of preferential scattering *toward* the blue is more than compensated for by the preferential absorption *in* the red.

[12] See Chap. 6 for a review of the connection between absorption and energy level differences.

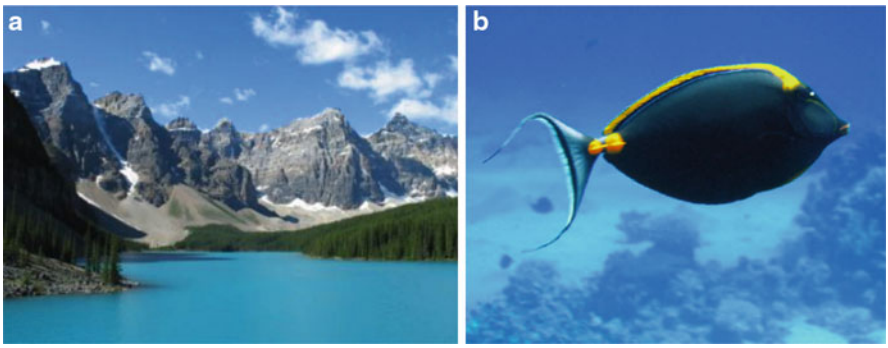


Fig. 10.9 (**a**) Moraine Lake, Bannf, Canada (**b**) Egypt Orange Spine Unicorn fish in the Red Sea (sources: (**a**) http://en.wikipedia.org/wiki/File:Moraine_Lake-Banff_NP.JPG (**b**) Photo by Sami Salmenkivi, http://seafishes.wordpress.com/category/family/surgeonfish/)

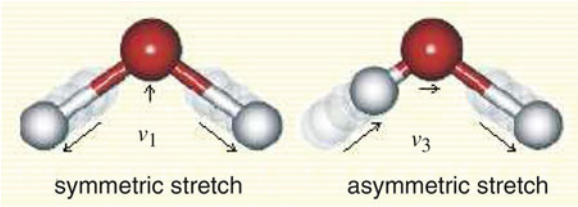


Fig. 10.10 Key vibration modes – $\nu 1$ and $\nu 3$ – responsible for the blueness of the sea (source: Martin Chaplin, http://www.lsbu.ac.uk/water/vibrat.html#2)

in the red is due to two modes of vibration – commonly labeled $\nu 1$ and $\nu 3$ – that produce a *fourth order combination mode* with frequency $f_c = f_1 + 3f_3$ in the red low wavelength region. These vibration modes are exhibited in Fig. 10.10.

10.6 Duration of a Note and Pitch Discrimination

How long must a single note be played for you to be able to have a clear sense of its pitch? Imagine a pianist playing a series of notes extremely rapidly – say at 32 notes per second. The duration of a single note is a mere 1/32 of a second. In case you have not done so, I suggest that you play a series notes on the piano by striking the keys sharply. See what happens as you move downward to ever lower notes. I am sure that you will find that it gets very difficult to sense the pitch when you play the very lowest of notes. Why is this so?

The answer to this question has to do with the frequency. Imagine if the frequency is but 20 Hz. Then with a duration of $\tau = 1/32$ s, there will be but $20/32 = 5/8$ oscillation. It is unreasonable that one can have any sense of pitch in this case. We would expect that the duration must be such that the number of oscillations is at least on the order of two or three.

In general, the number of oscillations N over a time interval τ is given by τ/period. Since $f = 1/\text{period}$, $N = f\tau$. For example, if the frequency is 440 Hz

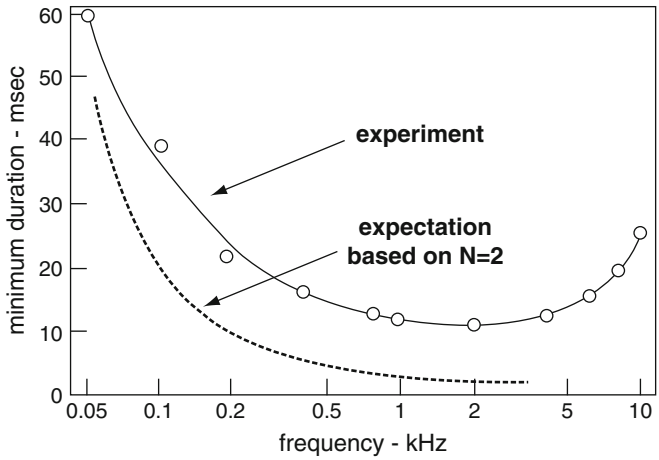


Fig. 10.11 Minimum Duration $|\tau$ in ms to discriminate a pitch vs. frequency on a log scale (source: Courtesy of Jouni Hiltunen, http://www.acoustics.hut.fi/teaching/S-89.3320/KA6b.pdf)

and the time interval is 1/100th of a second, the number of oscillations is $N = f\tau = 440 \times 0.01 = 4.4$.

For a given time interval, the lower the frequency, the poorer is our sense of pitch. As a consequence, it is often difficult to perceive clearly the pitch of the notes at the far lower end of the piano. We can express the minimum duration as

$$\tau = \frac{N}{f} = NT, \tag{10.13}$$

where T is the period.

We see how the necessary duration is inversely proportional to the frequency. We see a comparison of the Fig. 10.11 experimental results of Matti Karjalainen with the expectation – the dotted curve – according to this formula with $N = 2$. Note that the horizontal axis is actually the logarithm of the frequency.[13] The rise in the minimum duration for very high frequencies is not explainable on the basis of our simplified approach.

Note

A related question is how precisely a device can measure the frequency of a sine wave of finite duration τ. The answer is that the uncertainty in frequency Δf is given by

$$\Delta f \approx \frac{f}{N} = \frac{1}{\tau}. \tag{10.14}$$

[13]The expected curve is then $\tau \propto 10^{-b \log f}$, where b is a constant. We can see the exponential behavior of the dotted curve.

> Thus, if you are measuring a frequency of 400 Hz and you sample the sound for 10 s, you can typically rely on the result to within 1/10 Hz. Thus you might obtain a reading of 400.1 Hz – and therefore a result with four significant figures.

10.7 Fusion of Harmonics: A Marvel of Auditory Processing

We take for granted that when a musical instrument plays a note, we will hear the sound of but one source of sound. We have already learned in Chap. 2 that the frequency spectrum of the periodic wave of a musical instrument is a harmonic series, with the frequency of the wave. We have also learned that the timbre of an instrument is partly determined by the relative amplitudes and phases of the Fourier components associated with the instrument. On the other hand, more than three decades ago, I began to use a device made by the PASCO corporation for producing an electronic signal consisting of a periodic wave having up to nine harmonics, with a fundamental of 440 Hz. Such a device is called a **synthesizer**. The electronic signal was fed into an amplifier, which was connected to a loudspeaker. All listeners reported that they could hear the individual harmonics in the sound. Each harmonic was audibly separated, as if the harmonic had come from a separate source of sound. What is the difference between the wave produced by the Pasco synthesizer and the sound wave produced by a musical instrument? There is no visible characteristic on an oscilloscope trace that indicates a difference. Perhaps there is a significant difference that is too small to see on the trace with one's eyes.

We will discuss in this section the fact that the brain processes the sound so that we do not hear the individual harmonics. This process is referred to as **fusion of harmonics**.[14,15]

---

[14]I have been greatly helped in my attempt to weed out the known understandings of fusion by two audio-psychologists: Alan Bregman of McGill University, Brian Roberts of Aston University (Birmingham, England) and Oliver Knill of Harvard University. Alan Bregman is the author of a book entitled *Auditory Scene Analysis* [MIT Press, Cambridge, 1994], in which he discusses how the brain processes and ensemble of sound inputs and organizes them according to sources. In particular, he explains how the brain is able to focus on one source of sound and ignore or become almost oblivious to other concurrent sources. As a result, we are able to hear one person speak in the midst of a dense crowd at a party.

[15]For a resource of introductory material and references on this subject see: (1-2-2011): http://jjensen.org/VirtualPitch.html#use.

A more general concept than fusion of hamonics is the concept of **virtual pitch**. It takes into account the tendency of the brain to choose a pitch to be perceived even of frequency spectrum is not perfectly in a harmonic series and there is ambiguity. The concept of virtual pitch is attributed to **Ernst Terhardt**. See his publications: **Pitch, consonance, and harmony**, Journal of the Acoustical Society of America, **55** #5 1974. p.1061-1069, and **Calculating Virtual Pitch**,

Fusion in the Taste of Food

It might be unclear to some readers what I mean by fusion of a mixture of harmonics. We can get some idea of fusion by considering the taste of a homogeneous dish of food. I recall many years ago, finding Lobster Cantonese an extremely delicious dish. When I first attempted to prepare the dish myself, I was amazed to learn that the essential ingredients were lobster and garlic. How well I knew the taste of garlic and yet how surprised I was that garlic was an essential ingredient in the recipe. Somehow, the blend of garlic and lobster produced a taste all its own – that of Lobster Cantonese – with the flavor of neither ingredient standing out. And so it seems to be with most superb dishes – as long as they are prepared properly. As another example, we can consider curries. Most are such that the ingredients are individually recognizable; there are fortunately some that have a wonderful homogenized flavor all their own. And so it seems to be with the fusion of a mixture of harmonics from a musical instrument!

It is interesting to consider and to try to perceive what the world of music would be like in the absence of fusion: The sound of a musical instrument would be heard as an ensemble of harmonics that would be superimposed with those of other musical instruments. We would lose our ability to separate out the sounds of the ensemble of instruments. Vibratos might lose their sweetness of tone. And so on.

Fusion of Harmonics is responsible for the rich beauty of musical instruments.

Here are some important interesting questions to be investigated:

1. What accounts for the ubiquitous fusion of the sound produced by musical instruments?
2. What are the conditions under which a sound wave with a superposition of harmonics will be fused and will be perceived as having a single source? Factors that seem to be important include:

   (a) The frequency of the fundamental studies indicate that the lower the fundamental frequency, the greater is the degree of fusion.
   (b) The number of harmonics present and their relative amplitude.

Hearing Research **1** 1979. p.155-182. Below are references to fascinating illusory responses to sequences of complex sounds:

(1) **Shepard's Staircase**, wherein a sound seems to be ever decreasing in pitch but is actually cycling around like M.C. Escher's staircase. For an incredibly hilarious representation of this illusion see (1-21-2011): http://www.flixxy.com/escher.htm. To listen to Shepard's staircase see (1-21-2011):
http://www.cycleback.com/sonicbarber.html
Reference: **Shepard, R.N.** (1964). **Circularity in judgments of relative pitch**, J. Acoust. Soc. Am., **36**, 2346-2353.
(2) **Diana Deutsch** (1-21-2011):
http://www.philomel.com/musical_illusions/octave.php
Included are a number of sound files that allow you to listen to illusions.

(c) The presence of "proportional modulation," sometimes referred to as "parallel modulation." Consider frequency modulation, which is characterized by two parameters. There is a rate at which the frequency is modulated – label it $f_{\rm m}$. Next, there is an amplitude of variation of the frequency – label it Δf. Proportional frequency modulation would involve each harmonic being modulated with the same frequency of modulation $f_{\rm m}$ but with a variation $f_{\rm a}$ in proportion to the harmonic number n.

Here are some possibly relevant sources of frequency modulation:

Acoustic stringed instrument
A major source of frequency modulation is **vibrato**. A small periodic variation in the length of string that is free to vibrate will lead to proportional variation in each of the harmonics. Here is another possible source of frequency modulation, small as it may be: Normally we think of the vibrating string as having two fixed ends. However, the transmission of sound waves involves the string moving the bridge. Therefore, the string is not absolutely fixed at the bridge. Therefore, the string's length and its tension are modulated.

Wind instrument such as a flute
Here too – there is a modulation of the frequency due to the vibration of the mouth of the musician.

3. Are there strong variations in the auditory processing of people such that sounds that are fused for some people are not fused for others. It is reported that some people can sometimes distinctly hear the individual harmonics produced by a musical instrument.

10.7.1 Mathematica File

Below we provide a Mathematica file for producing a variety of waves that can be heard using the PLAY command of Mathematica. The reader can use this file to test his/her auditory processing a superposition of proportionally frequency modulated harmonics. When you run the commands, a window will appear that allows you to listen to the wave form.

SYMBOLS:

Text: fn = central frequency for a given harmonic, $n = 1, 2, 3, ...$ [f2 = 2 f1, ..., f6 = 6 f1]
$\equiv \Delta f$ = maximum change in frequency
hn = $\delta f n / f n$ = **modulation index**
An = amplitudes

We begin with all modulation indices $hn \equiv \Delta f/fn$ the same:
$h = fv/f1$.

INPUT LINES
f1 = f
f = 440
f2 = 2*f
f3 = 3*f
f4 = 4*f
f5 = 5 f
f6 = 6 f
fv = 10
phi = f*t + (fv/(2 Pi f))*Cos[2 Pi f * t]
phi2 = f2*t + (fv/(2 Pi f2))*Cos[2 Pi* f2 * t]
phi3 = f3*t + (fv/(2 Pi f3))*Cos[2 Pi* f3 * t]
phi4 = f4*t + (fv/(2 Pi f4))*Cos[2 Pi* f4 * t]
phi5 = f5*t + (fv/(2 Pi f5))*Cos[2 Pi* f5* t]
phi6 = f6*t + (fv/(2 Pi f6))*Cos[2 Pi* f6* t]
A1 = 1
A2 = 1
A3 = 1
A4 = 0.01
A5 = 0.01
A6 = 0.001

INPUT COMMAND
Play[A1*Sin[2 *Pi* phi] + A2*Sin[2* Pi * phi2] + A3*Sin[2* Pi * phi3] + A4* Sin[2* Pi * phi4] + A5* Sin[2* Pi * phi5] + A6*Sin[2* Pi * phi6], t, 0, 5, SampleRate → 40000]

Next, the modulation indices are proportional to the harmonic:
$hn = n * fv/f1$

INPUT LINES
phi = f*t + (fv/(2 Pi f))*Cos[2 Pi f * t]
phi2a = f2*t + (fv/(2 Pi f))*Cos[2 Pi* f2 * t]
phi3a = f3*t + (fv/(2 Pi f))*Cos[2 Pi* f3 * t]
phi4a = f4*t + (fv/(2 Pi f))*Cos[2 Pi* f4 * t]
phi5a = f5*t + (fv/(2 Pi f))*Cos[2 Pi* f5* t]
phi6a = f6*t + (fv/(2 Pi f))*Cos[2 Pi* f6* t]

INPUT COMMAND
Play[A1*Sin[2 *Pi* phi] + A2*Sin[2* Pi * phi2a] + A3*Sin[2* Pi * phi3a] + A4* Sin[2* Pi * phi4a] + A5* Sin[2* Pi * phi5a] + A6*Sin[2* Pi * phi6a], t, 0, 5, SampleRate → 40000]

10.8 Additional Psychoacoustic Phenomena

We summarize below additional psychoacoustic phenomena not covered in this text:

1. The variation of pitch with loudness: Our sense of pitch varies with loudness. This phenomenon reflects a dramatic difference between a **subjective perception** and an **objective input**. We see the results of experiments of this phenomenon in Fig. 10.12.[16] The central features of the graphs are the following: At about 2 kHz, pitch does not depend upon loudness. Below 2 kHz, the pitch decreases with increasing loudness, while above 2 kHz it decreases with increasing loudness. Note that when musicians tune their instruments for the purposes of having a shared tuning, they share their notes at a low sound level.
2. **Just noticeable difference in frequency** as a function of frequency: How different can two sounds be with respect to frequency and still be distinguishable?
3. **Just noticeable difference of loudness** as a function of frequency: How different in loudness can two sounds be and still be distinguishable?
4. **Second order beats**, also referred to as **mistuned consonances**, result in a sense of beating between the corresponding harmonics of the two tones.
5. **masking**: One sufficiently intense tone of a certain frequency will mask a second tone having a different frequency and much lower intensity.
6. A psychoacoustic basis for **consonance**.

The reader is encouraged to read other resources that describe the incredibly rich experiences connected with both psychoacoustics and auditory processing. A recent outstanding book at a layman's level is *This is Your Brain on Music*, by Daniel

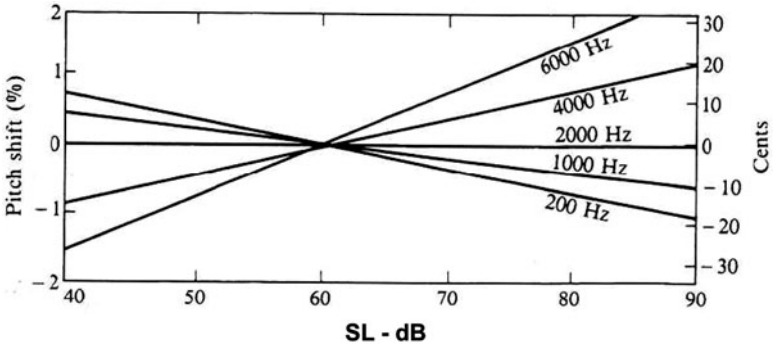


Fig. 10.12 Change in pitch with respect to sound level (source: Based upon *Hearing Research*, vol. 1, p. 162, (1979), "Calculating Virtual Pitch", Ernst Terhardt, with permission from Elsevier)

[16]This website has sound bites that allow you to hear the change of pitch with increasing intensity. (1-21-2011):
http://www.santafevisions.com/csf/demos/audio/412_dependence_pitch_intensity.htm.

Levitin [Penguin Group, New York, 2006]. Another book at a higher level is *Tuning, Timbre, Spectrum, Scale*, by William A. Sethares [Springer-Verlag, London, 2nd edition]. To appreciate the latter book, the reader should first study Chap. 11 on musical scales. And finally we mention *Musicophilia: Tales of Music and the Brain*, by Oliver Sachs [Alfred A. Knopf, New York, 2007].

10.9 Terms

- Aural harmonics
- Combination tone
- Difference tone
- Equal loudness curve
- General force
- Harmonic distortion
- Linear response
- Masking
- Nonlinear response
- Phon

10.10 Important Equations

Mathematical relation between the sone and the phon:

$$s = 2^{\frac{\phi - 40}{10}}. \tag{10.15}$$

Combination tones:

$$(f_2 - f_1), \quad (2f_1 - f_2), \quad (3f_1 - 2f_2).$$

10.11 Problems for Chap. 10

Five **pure tones** (which we will call A, B, C, D, and E) are sounded with the physical characteristics shown in table above. Use the equal loudness curve in the text to answer the following three questions.

Tone	Frequency (Hz)	Intensity (W/m^2)
A	1,000	1.0×10^{-10}
B	100	1.0×10^{-10}
C	100	1.5×10^{-7}
D	10,000	1.5×10^{-8}
E	15,000	9×10^{-8}

1. For a normal young ear, which tone is probably the loudest?

   (a) A
   (b) B
   (c) C
   (d) D
   (e) E

2. For a 60-year-old person, which tone or tones are probably inaudible?

   (a) B
   (b) E
   (c) B and E
   (d) A, B, and E
   (e) A, B, D, and E

3. For a normal young ear, which tone or tones sound equal in loudness to tone C?

   (a) B
   (b) D
   (c) E
   (d) D and E

4. (a) What is the **phon level** of a 60 dB sound at 100 Hz?
   (b) What is the **sound level** in dB of a 10,000 Hz sound of 20 phons?
   (c) What is the loudness in sones corresponding to 80 phons?
   (d) How many phons corresponds to a loudness of 1/4 sone?
5. (a) What are the frequencies of the significant **combination tones** produced by two tones having frequencies 500 and 750 Hz?
   (b) What is the hypothesized physical basis for the perception of combination tones?
6. Consider a sound of frequency of 400 Hz and sound level of 40 dB.

   (a) What is the number of phons?
   (b) Find the loudness in sones.
   (c) Suppose that we want to double the loudness in sones of the sound. Find the required number of phons and the sound level.

7. Suppose that we have one million bees, each producing a sound of frequency 200 Hz and a sound level of 15 dB. The determination of the loudness of the total sound produced is the object of this problem.
   See Sample Problem 10.3.

   (a) Find the corresponding phon level of the sound from one bee.
   (b) Find the corresponding loudness in sones of the sound from one bee.
   (c) Find the corresponding sound level in **decibels** and the loudness in **phons** of the sound of one-million bees.
   (d) Find the corresponding loudness in sones of the sound of one million bees.

8. Again referring to Sample Problem 10.3, note that when the two sources with about the same frequency have one sone each, the resulting sone level was about 1.1 sones, which is less than the two sones resulting had the frequencies not been close. Yet, we should note that this value of 1.1 depended upon the frequency being 1,000 Hz, for which a change in SL equals a change in phons.

   (a) Explain why, for the change in sones to exceed two, the frequency must be such that a change in SL of 3 dB amounts to a change of at least 10 phons.
   (b) Study the equal loudness curves in Fig. 10.2 and see whether there is any such frequency.

9. Here is a problem that professionals who produce musical recordings must contend with. It is also an issue that affects all music we hear, whether live or from recordings.

   For simplicity, suppose that we mix two pure tones, one at 200 Hz and the other at 1,000 Hz. We want them to have the same loudness. We mix at a level that the 1,000 Hz tone has a level of 20 phons.

   (a) Determine the sound level of the 200 Hz tone?
   (b) Now suppose that the **intensity** of both tones is increased by a factor of 100. Determine the resulting sound level and the phon level of each tone, noting that the latter are now unequal.

10. Most people would regard the moth as being quite primitive and helpless against attacks by a predator. However, it has been known for over 50 years that moths have auditory perception using eardrums and a few sensory cells that can detect an impending attack by predator bats. Recently,[17] it was reported that a moth's eardrum is not static but adjusts itself to the changing sound from bats. Bats use ultrasound to locate their prey and, of course, to avoid obstacles in their path. The range of frequencies of a bat during general flight is usually in the range of 20–40 kHz. This is a range of frequencies for which the moth's eardrum is most responsive in its resting state.

    However, there are two interesting changes in bat and moth behavior in the course of a predatory attack.

    Explain the physical principles that account for these changes.

    (a) When homing in on prey, a bat's frequency is increased to a much higher range, up to about 80 kHz.
    (b) When a bat is approaching a moth, the increased sound intensity leads the moth's eardrum to become stiffer.

[17] http://www.sciencedaily.com/releases/2006/12/061218122629.htm.

Chapter 11
Tuning, Intonation, and Temperament: Choosing Frequencies for Musical Notes

Imagine yourself seated in a concert hall, anticipating the beginning of a symphony orchestra performance. The musicians are all seated. The concertmaster rises and calls to the oboist to sound the "A," which will be the standard pitch that all others will use to tune their instruments. In advance the oboist has tuned the oboe to the standard frequency chosen by the orchestra, which is usually a frequency of 440 Hz. The winds and brass and tympani tune their instruments accordingly.

Next, the strings tune their A-strings to the oboe's "A." Following that, the strings tune their other three strings accordingly. For example, the violins tune the "E," which is a musical "fifth" above the "A," played simultaneously with the "A." The violinist strives for the maximal beauty of a resonance between the two strings. Tests reveal that the ratio of the two frequencies is very close (within a fraction of a percent) to $f_E/f_A = 3:2$. The violinist continues in a similar manner with the "D" a fifth below the "A" and then the "G" a fifth below the "D." The frequency ratios – high to low string – again will be close to 3:2. The harpist has already tuned the harp's 47 strings to the standard frequency for the A, but in advance and while on stage so as to avoid changes that might result from moving the harp. Upon hearing the oboe, the harpist might have to quickly make fine adjustments.[1]

If a piano is to be played, its 230 strings (!) would have been previously tuned to the standard "A" before the time of the performance. The pianist will not have an opportunity to make any changes during the concert.

[1] A harpist acquaintance of mine, Judith Ross, told me the following: "Harpists are always nervous about the tuning. I tune a little higher than 440, even though the oboe blows 440, because the violins always seem to tune or go sharp. And I'm used to adjusting as I go. When I play in the pit for shows, I'm constantly re-tuning during the dialogues. I've even replaced broken strings during dialogues. In orchestral concerts if I hear something wacky, I try to tune the offending string(s) as inconspicuously as possible. Stravinsky is believed to have said that 'Harpists spend 90% of their time tuning their harps and 10% playing out of tune.'" Here is a source of this quote (1-23-2011): http://en.wikiquote.org/wiki/Igor_Stravinsky.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_11,

Once tuned, the musicians will perform the concert by producing strings of notes with frequencies that are limited by their respective instruments. Of course, musicians will frequently have to make adjustments of the tuning of their instruments during a concert due to changes that are wont to take place – such as temperature changes, changes in the tension of a string that might not have been stabilized when first tuned, and so on. The string players will have liberty to choose notes that involve shortening the length of string that is free to vibrate by varying the placement of their fingers on the strings. Winds and brass can choose notes by covering holes on the instrument or changing the length of the pipe in the case of a trombone. The frequency can be moderately varied by the winds and brass by changing the way the mouth blows into the instrument or covering the end of the instrument in the case of the brass. The harpist and pianist cannot make such adjustments.

Those notes that cannot be changed significantly are referred to as **fix tuned**. A piano is an entirely fix tuned instrument. The strings, winds, and brass are only partly fix tuned.

The whole process of tuning an orchestra is a grand display of majesty, reflecting the goal and ability of a group of people to get together to produce an extremely well-organized act of cooperation leading to the heavenly sound of the music of a symphony orchestra.

The focus of this chapter is to study the complex aspects of the choice of frequencies that are played once given a standard "A." This choice has two components: First is the **intended frequency**, which we refer to as **temperament** or **tuning**. Second is the **actual frequency** in a performance. The latter is best referred to as **intonation** and can reflect either intended choice, where choice is possible, or a mistaken, unintended outcome. "Bad intonation" refers to a disagreeable resulting pitch. Above all, in this chapter we will learn about the central role of **numerology** in the process of choosing frequencies and of the fact that a compromise cannot be avoided in an attempt to maximize consonance. Total consonance in its usual meaning can be shown to be impossible mathematically!

Since music is sound and sound consists of waves, what are the waves that music is made of? The focus in this text is on music that has a definite pitch. And, as we have seen, pitch is, for the most part (but not entirely), determined by the frequency of a periodic wave.[2] Given that the audible range of frequencies spans $\sim$20 Hz to $\sim$ 20,000 Hz, a musical composition could call for musical notes whose fundamental frequencies span this entire range, *continuously*. Instead, cultures have produced musical compositions which make use of certain *discrete* sets of frequencies. These sets are called **musical scales**.

To some extent, the choice of frequencies is analogous to the set of colors of paints that a painter places on his/her palette, putting aside the fact that a painter uses the set of colors to produce a continuum of other colors by mixing the base set

[2]Note that I refer here to a "**periodic wave**" , not a sinusoidal wave that is associated with a "**pure tone**".

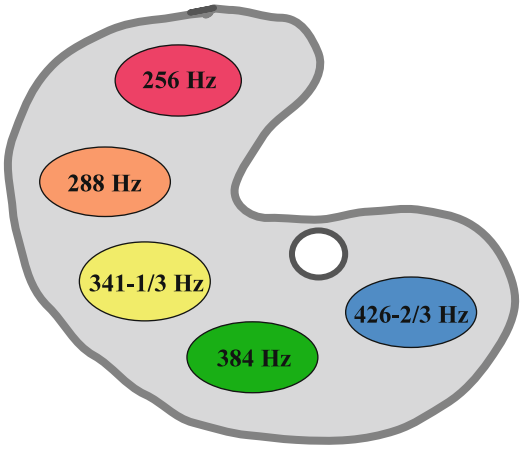


Fig. 11.1 Frequency palette

in various proportions. See Fig. 11.1, in which we exhibit this analogy with a set of frequencies corresponding to a **pentatonic scale** (to be discussed below) laid out on a palette.

The **pentatonic scale** is found throughout the world. The two other most common musical scales are the **diatonic scale** and the **chromatic scale**, which are used in the Western world. We will discuss only these three. Also, we will use the symbols and terminology of Western music. The key question that any serious musician must deal with is:

What should be the frequencies of the musical notes?

The goal of this chapter is to study the bases for the choices in tuning or the choice of the frequencies. We will begin with a discussion of musical scales, which are the backbone of what is referred to as tonal Western music. We then briefly discuss Pythagorean tuning, which is one of the oldest mathematically defined tunings in Western music. We then move on to Just tuning and discuss its drawbacks. Finally, we discuss the most widely used tempered tuning – equal temperament. In Problem 14 of this chapter, we discuss Werkmeister I(III) temperament, which was one of a number of temperaments that were popular in the Baroque era (ca. first half of the eighteenth century).

11.1 Musical Scales

Every scale has a **key note**, which is the first note of the scale: A musical composition tends to be drawn to that note so strongly that almost invariably, the last note of a Western composition is the key note. The composition tends to feel

incomplete and produces a certain tension unless this is so.[3] To be specific, we will assume that we are dealing with a scale that has "**C**" as its key note. We say that the "scale is in the **key of C**."

Most fundamental in determining the discrete set of frequencies chosen for a scale is the **octave**: Two frequencies which are in a 2:1 ratio are said to be an **octave** apart. They sound much alike. Some people cannot tell them apart. Once we choose a certain frequency for so-called middle-C on the piano, we can generate an infinite set of other frequencies that are octaves apart. Consequently, the entire spectrum of notes consists of a series of identical octaves of notes. We will use the symbol C_4 to denote middle-C. (See Fig. 11.2 of a **piano keyboard**.) The frequencies indicated are those associated with the A above middle-C set at 440 Hz and the other notes tuned with **equal temperament** , which will be discussed in detail in this chapter.

Suppose, for simplicity, we choose $f_{C_4} = 250\,\text{Hz}$. Then an octave above C_4 is C_5, with $f_{C_5} = 2 \times 250 = 500\,\text{Hz}$. Continuing on, we obtain

$$f_{C_6} = 2 \times f_{C_5} = 4 \times f_{C_4} = 1{,}000\,\text{Hz}$$
$$f_{C_7} = 2{,}000\,\text{Hz}$$
$$f_{C_8} = 4{,}000\,\text{Hz}$$
$$\ldots$$

Moving downward, we obtain $f_{C_3} = (1/2)\, f_{C_4} = 125\,\text{Hz}$, ...etc. The question is: How should we fill in the notes with discrete frequencies between a C and its octave above, C_5?

> **Note**
>
> *Once this range is filled in, all other notes are determined by octave relationships, as above*. In this connection, we should note that the term **octave** has an additional meaning: It is also used to refer to the *set of notes* spanning from a given note to its octave note above. Thus, we may also refer to the **octave of notes** ranging from C_4 to C_5 or from C_5 to C_6.

Moving from one note to another is referred to as taking **steps** – e.g., making a *semitone step*.

Considering all the fuss we will be making about tuning and intonation in theory, how do instrumentalists *actually* tend to choose their frequencies? There has been much study of this last question; unfortunately, the question is beyond the scope of the text.[4]

---

[3] In fact, some composers *intentionally* end on a note other than the key note so as to leave the listener in a state of tension!

[4] For further details, see the classic text on the psychology of music: Seashore, Carl Emil, 1866–1949. Psychology of music, by **Carl E. Seashore**. New York, Dover Publications [1967] ML3830.S32 P8 1967.

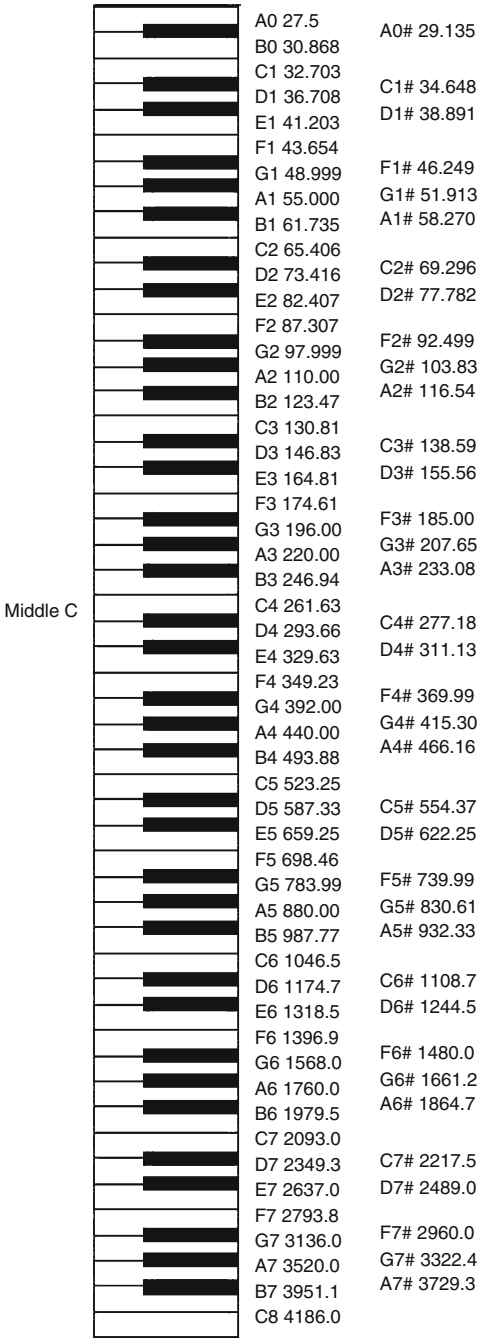


Fig. 11.2 Piano keyboard with a corresponding **musical staff**. Shown to the *right* of each key is the frequency according to equal tempered tuning, which will be discussed later on in this chapter (source: http://www.vibrationdata.com/piano.htm, courtesy of Tom Irvine)

11.2 The Major Diatonic Scale

The notes of the diatonic scale in the key of C are **C, D, E, F, G, A, B, C'**. They are represented by the white keys on a piano keyboard (Fig. 11.3).

*The **musical interval** expresses the relationship between the pitches and hence frequencies between a pair of notes.* Musical intervals have the following names:

- C to C': **Octave**
  so called because there are eight notes in the diatonic scale.
- Intervals between neighboring notes in the diatonic scale are either **semitones** or **whole tones**.
- A semitone is also called a **minor second**.
  Examples: E-F and B-C
- A whole tone is also called a **major second**.
  Examples: C-D, D-E, G-A, and A-B
- **Minor third**=1-1/2 whole tones=3 semitones
  Examples: D-F and A-C
- **Major third**=2 whole tones=4 semitones
  Examples: C-E and G-B
- **Fourth**=2-1/2 whole tones=5 semitones
  Examples: C-F, D-G, and A-D
- **Fifth**=3-1/2 whole tones=7 semitones
  Examples: C-G, G-D, and D-A
- **Minor sixth**=4 whole tones=8 semitones
  Examples: B-G and E-C
- **Major sixth**=4-1/2 whole tones=9 semitones
  Examples: C-A and G-E

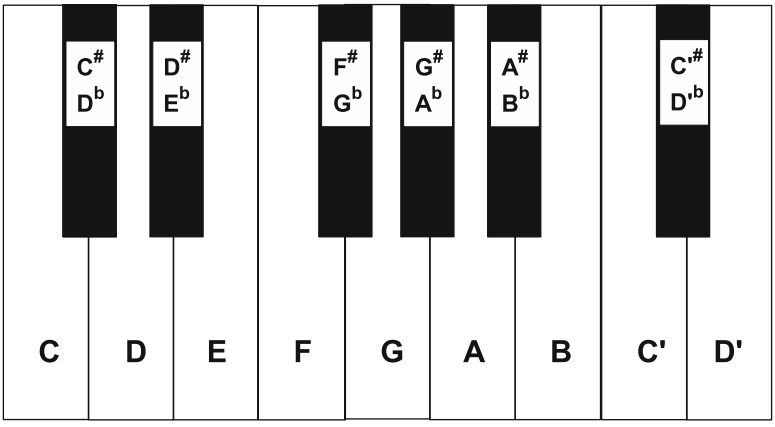


Fig. 11.3 Piano keyboard

– **Seventh**=5-1/2 whole tones=11 semitones
  Examples: C-B

Thus, in relation to the base note of C, the notes of the diatonic scale in C major are

– D: 2 semitones or one whole tone ... **major second**
– E: 4 semitones or 2 whole tones ... **major third**
– F: 5 semitones or 2-1/2 whole tones ... **fourth**
– G: 7 semitones or 3-1/2 whole tones ... **fifth**
– A: 9 semitones or 4-1/2 whole tones ... **major sixth**
– B: 11 semitones or 5-1/2 whole tones ... **seventh**
– C: 12 semitones or 6 whole tones ... **octave**

The *essential question* now, as before, is the following: *How should the frequencies associated with the notes in the **diatonic scale** be chosen?* This choice is referred to as **tuning**. We will discuss in detail the three most important tunings, **Pythagorean Tuning**, **Just Tuning**, and **Equal Tempered Tuning**.

The fact that a choice is open to us and must be made might seem foreign to someone who regards music as having a certain natural state of existence. Let it be known that even primitive peoples were aware of the necessity of making a choice. Reread the tale of the Huang Chung in Chap. 1, as a reminder. Most people are mere "consumers" of the music performed by musicians and are not aware of the details that musicians have to dabble with. Musicians, in turn, have to rely upon the still more diligent studies of others.

The most important fact that must be recognized is that

musical intervals correspond to frequency ratios.

This fact can be understood from the following exemplary situation: Suppose that we choose a certain frequency for f_G to be the fifth above the chosen frequency $f_C = 250\,\text{Hz}$, say 375 Hz, as in the case of so-called Just tuning. Then the pair of octaves above these two notes, C' and G', must also be a fifth apart.

Now, since $f_{G'} = 2f_G = 750\,\text{Hz}$ and $f_{C'} = 2f_C = 500\,\text{Hz}$, $f_{G'}/f_{C'} = 2f_G/2f_C$, or

$$\frac{f_{G'}}{f_{C'}} = \frac{f_G}{f_C} = \frac{3}{2}.$$

*The ratio 3/2 defines the **Just** interval for a fifth.* Alternatively, consider the sequence of notes, C-E-G. C-E is a **major 3rd**, while E-G is a minor 3rd. Together, they add up to the musical interval of a fifth, corresponding to C-G. We write

major third + minor third = fifth

This fact is exhibited in Fig. 11.4.

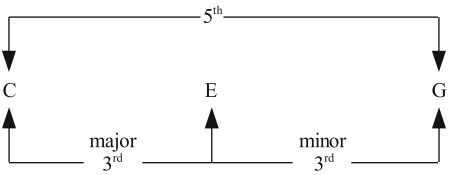


Fig. 11.4 Adding of the major third and the minor third to a fifth

This equation is reflected by the following mathematical equation:

$$\frac{f_E}{f_C} \cdot \frac{f_G}{f_E} = \frac{f_G}{f_C}. \tag{11.1}$$

We thus see that it is

1. The *ratio* of frequencies that defines the musical interval
2. *Adding intervals* amounts to *multiplying by frequency ratios*

We have arrived at the following stage in our study of tuning: We first choose the frequency of one note – say middle C. All octaves are determined by the unchangeable ratio 2:1. Then, since musical intervals correspond to frequency ratios, our task is to decide what the frequency ratios should be corresponding to the musical intervals introduced previously.

11.3 Comments Regarding Western Music

Before we discuss the three most important tunings, we would like to note some important aspects of Western music. We have indicated that compositions of Western classical music are drawn strongly toward the **key note**: When we are away from the key note, we feel a pull toward it. Interestingly, the pull is strongest when we are close to a key note. It is true that other notes are pulling too: Most strongly is the fifth above the key note. (The G in the key of C.) Next in strength are the major third (the E) and the minor third (the E-flat). (Those knowledgeable about musical theory will note that which is the stronger of these last two depends upon whether the key is a major one or a minor one.) Next in strength is the fourth-above the C (in this case, the F). (This author must grant that some people feel that the fourth above is stronger than the major and minor thirds.) And finally, we have the major and minor sixths.

The pulling effect is associated only partly with our sense of **consonance** between the key note and each of the above notes – its opposite being **dissonance**. I suggest that this is so in spite of the fact that consonance is associated with a satisfying sound when two notes are played together: The key note is always stored in the memory bank of our brain and other notes are compared with it, whether or not the key note is sounded.

This desire for consonance plays a role in determining the tuning. The relevant question then is: What properties of a musical interval tend to produce a sense of

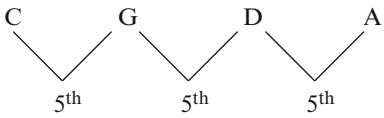


Fig. 11.5 Steps of fifths

consonance? Often it is said that a sense of consonance guides the tuner. But what characteristics are actually associated with consonance?[5] One definite characteristic traditionally is a sense of the richness of tone that is associated with resonance. To see how resonance is relevant, we will consider the tuning of a viola.

First, the A string (which is the A above middle C) is tuned to some standard or reference frequency. These days, that frequency is taken to be a value between 440 and 444 Hz. In Bach's time, it may have been as low as 415 Hz! This frequency seems to be steadily increasing, reflecting an ever-increasing preference for bright sounds.

Next, the remaining strings are tuned so as to be fifths apart, as indicated in Fig. 11.5:

Now, suppose that the D is set at a frequency that is exactly $2/3$ that of A, as in the case of Pythagorean or Just tuning. For concreteness, let $f_A = 441\,\text{Hz}$. Then $f_D = 294\,\text{Hz}$.

When the D-string is excited, the harmonics of 294 Hz are generally present. These are: 294 Hz, $2 \times 294 = 588\,\text{Hz}$, $3 \times 294 = 882\,\text{Hz}$, ...

When the A-string is excited, the harmonics of 441 Hz are generally present. These are: 441 Hz, $2 \times 441\,\text{Hz} = 882\,\text{Hz}$, ...

Note the match between the third harmonic of the D and the second harmonic of the A. A consequence is resonance between the vibrating strings. If the D and A are played together, this resonance enriches the tone quality. (A fascinating question is the extent to which there is a resonance taking place in the ear and possibly in the brain.[6]) It is interesting to note that it is much more difficult to adjust by ear two **pure tones** to a 3:2 ratio of frequencies. There is no sharp maximum of consonance at this ratio. Highly accurate tuning can, however, be accomplished by listening to the *combination tones* produced in the hearing process. (See Chap. 10.)

Generally, it is found that pairs of notes sound most consonant when their frequencies are in a ratio of small integers, an observation that leads us to **Just tuning**, which will be treated in Sect. 11.5. However, we will first study the simpler Pythagorean intonation.

[5] Of course, if the tuner has perfect pitch (i.e., has an internal sense of what the pitch of an isolated note is without the need of a reference sound such as tuning fork), then consonance may not be a factor. We should want to understand better the nature of perfect pitch in this context.

[6] See The Physics of Musical Sounds by C.A. Taylor (The English Universities Press, London, 1965), Introduction to the Physics and Psychophysics of Music by Juan G. Roederer, (The English Universities Press, London, 1973), and On the Sensations of Tone, by **Hermann L.F. von Helmholtz** (originally published in German in 1877, reissued in English by Dover, N.Y., 1954).

11.4 Pythagorean Tuning and the Pentatonic Scale

Pythagorean tuning is based on assigning a frequency ratio of $3:2$ to the fifth and treating all the remaining intervals as subservient to the fifth, as follows. We start by choosing the frequency f_{C_4}. Then, we determine the frequencies for the notes that arise as we ascend and descend in a series of fifths:

NOTE

$F_3 \rightarrow$	$C_4 \rightarrow$	$G_4 \rightarrow$	$D_5 \rightarrow$	$A_5 \rightarrow$	$E_6 \rightarrow$	B_6
$2/3$	1	$3/2$	$(3/2)^2$	$(3/2)^3$	$(3/2)^4$	$(3/2)^5$

Note how in descending down by a fifth from C_4 to F_3, we multiply by $2/3$. The reason is that in doing so we must divide by $3/2$, which is equivalent to multiplying by $2/3$.

We will discuss only the *five* notes C, D , F, G, and A. These notes form a **pentatonic scale**, which is the essential scale in many parts of the non-Western world.[7] The process of determining the remaining frequencies that are needed to form the diatonic and chromatic scales is a mere continuation of the process we will demonstrate below.

What we want are the frequencies in the octave starting with C_4. They are determined by a process called **reducing to the octave** the above frequencies, as follows: $f_{F_4} = 2f_{F_3}$, $f_{D_4} = (1/2)f_{D_5}$, $f_{A_4} = (1/2)f_{A_5}$.

Thus, using $2(2/3) = 4/3$, $(1/2)(3/2)^2 = 9/8$, and $(1/2)(3/2)^3 = 27/16$, we obtain the pentatonic scale in **Pythagorean tuning** (see Fig. 11.6).

On the last line, we have indicated the intervals between neighboring notes. For example, for the interval between D and F, we have:

$$\frac{f_F}{f_D} = \frac{\frac{f_F}{f_C}}{\frac{f_D}{f_C}} = \frac{\frac{4}{3}}{\frac{9}{8}} = \frac{4}{3} \times \frac{8}{9} = \frac{32}{27}. \tag{11.2}$$

We note the results for the following **Pythagorean intervals** in Pythagorean tuning:

Whole tone (C-D) : $9/8$
Minor third (D-F) : $32/27$

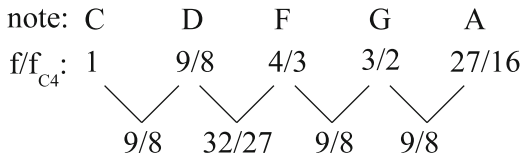


Fig. 11.6 Construction of the pythagorean scale

[7] An easy way to hear this scale is to play the five black keys of the piano in order, starting with whatever first note you wish.

Major third (F-A) : 81/64
Fourth (C-F) : 4/3
Major sixth (C-A) : 27/16

We note that the two intervals of a third (the minor third and the major third) each involve a ratio of rather large integers. As such, these intervals do not sound very consonant in the traditional sense.

11.5 Just Tuning and the Just Scale

Just tuning sets a *priority* on ratios of small integers for the frequency ratios of two musical intervals, the Fifth – 3/2, and Major third – 5/4.

As we pointed out above, these two musical intervals are the central ones in Western classical music. We will now see how this choice for the two intervals is used to *generate* the Just notes.

As in the case of generating the Pythagorean scale, we start by choosing the frequency f_C of the key note C. From thereon, we are interested only in the ratio of the frequency of interest to this frequency. Let us lay out the notes of the diatonic scale with the following notation. We obtain the E and G straightforwardly, since C-E is a major third and C-G is a fifth. See Fig. 11.7.

We can obtain A by descending from E to A by a fifth and then reducing to the octave by ascending by an octave. This involves the product

$$2\left(\frac{2}{3}\right)\left(\frac{5}{4}\right) = 2\left(\frac{5}{6}\right) = \frac{5}{3}.$$

See Fig. 11.8.

We can obtain the F by descending from A by a major third, thus obtaining the ratio $(4/5)(5/3) = 4/3$. See Fig. 11.9.

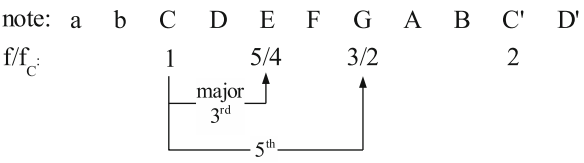


Fig. 11.7 Obtaining G from C in Just tuning

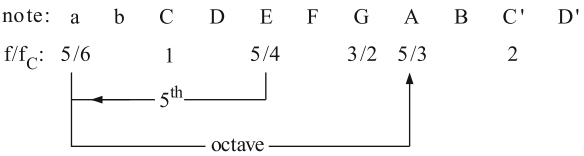


Fig. 11.8 Obtaining A from C in Just tuning

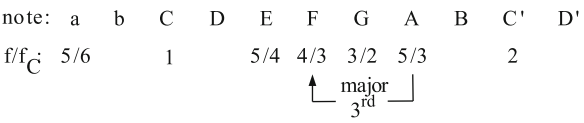


Fig. 11.9 Obtaining F from C in Just tuning

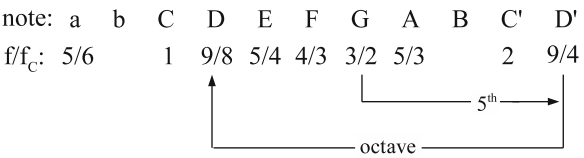


Fig. 11.10 Obtaining D from C in Just tuning

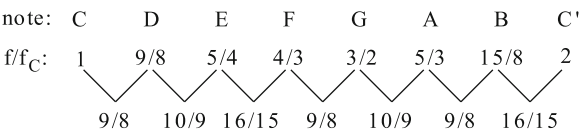


Fig. 11.11 Full Just diatonic scale

We can obtain the D by ascending from G by a fifth and then descending by an octave. The product is $(1/2)(3/2)(3/2) = (1/2)(9/4) = 9/8$. See Fig. 11.10.

Finally, we can obtain the B by ascending from E by a fifth. The product involved is $(3/2)(5/4) = 15/8$. The full Just diatonic scale is shown in Fig. 11.11.

Intervals between neighboring notes are indicated on the last line. They are obtained by taking ratios of the corresponding pair of ratios. For example,

$$\frac{f_F}{f_E} = \frac{\frac{f_F}{f_C}}{\frac{f_E}{f_C}} = \frac{\frac{4}{3}}{\frac{5}{4}} = \frac{4}{3} \times \frac{4}{5} = \frac{16}{15}. \tag{11.3}$$

The basic music intervals in Just tuning are listed below:

Semitone	16/15	Fifth	3/2
Whole tone	10/9 and 9/8	Minor sixth	8/5
Minor third	6/5	Major sixth	5/3
Major third	5/4	Seventh	15/8
Fourth	4/3		

We obtained the minor third, $6/5$, from the interval E-G as follows:

$$\frac{f_G}{f_E} = \frac{\frac{f_G}{f_C}}{\frac{f_E}{f_C}} = \frac{\frac{3}{2}}{\frac{5}{4}} = \frac{3}{2} \times \frac{4}{5} = \frac{6}{5}. \tag{11.4}$$

We obtained the minor sixth, $8/5$, from the interval A-F', in a similar way.

In Fig. 11.12 is a chart that lays out the important intervals in Just tuning in a clear way.

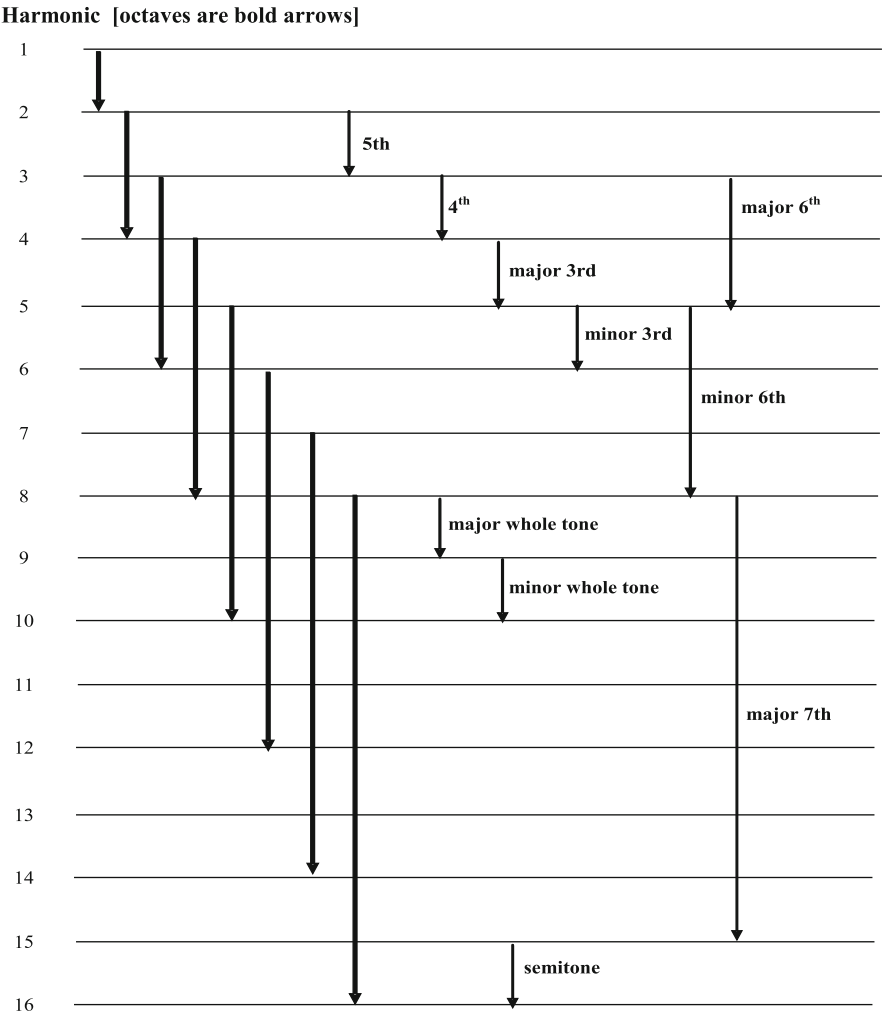


Fig. 11.12 Chart of important Just intervals

11.6 The Just Chromatic Scale

The **Chromatic Scale** includes all the so-called sharps and flats of the notes of the diatonic scale in the key of C. These are, respectively, semitones above and below the corresponding notes. Thus, $C^{\#}$ is a semitone above C, and B^{b} is a semitone below B.

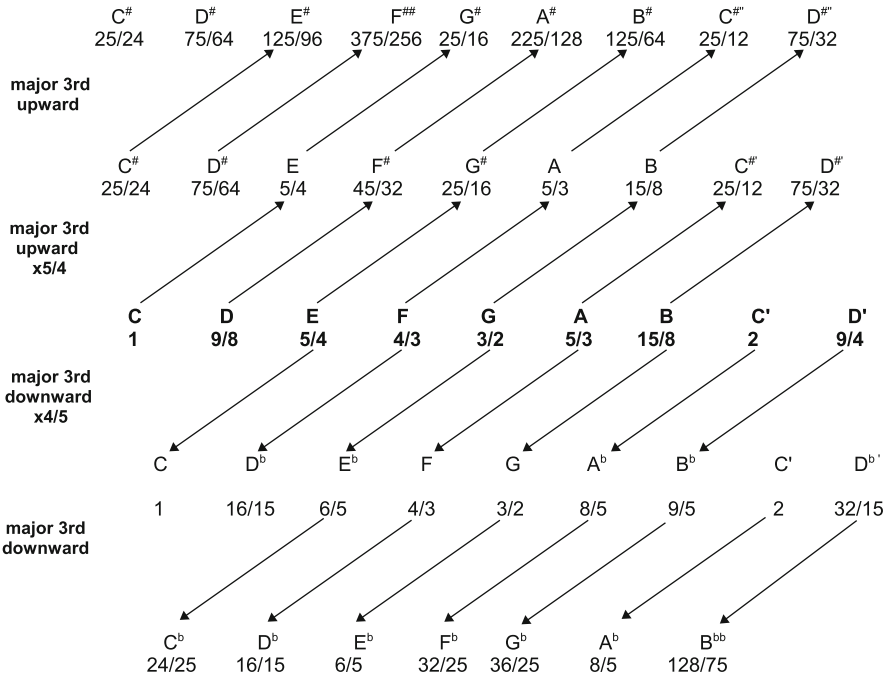


Fig. 11.13 Building the Just chromatic scale

To obtain the notes with **sharps** and **flats** involves a procedure of ascending and descending by major thirds, as shown on the next page.

Note that $f_{G^\#} \neq f_{A^\flat}$, $f_{A^\#} \neq f_{B^\flat}$. Such pairs of notes are called **enharmonic equivalents**. On a piano, they are not distinguishable since they are produced by the same key. The interval between the above pairs is $125/128 = 0.9755$:

$$\frac{f_{G^\#}}{f_{A^\flat}} = \frac{25/16}{8/5} = \frac{125}{128}$$

$$\frac{f_{A^\#}}{f_{B^\flat}} = \frac{225/128}{9/5} = \frac{125}{128}.$$

We obtain the same ratio for the enharmonic equivalents, $f_E/f_{F^\flat}$ and $f_{B^{\flat\flat}}/f_A$.

The detailed construction of the Just chromatic scale is laid out in Fig. 11.13. It is interesting to note the inclusion of double sharps and double flats, which are necessary for some scales.

Many musicians, who are not constrained in playing with a fixed tuned instrument such as the piano, will distinguish between enharmonic equivalents. As a violinist, this author happens to have a strong tendency to do so.

11.7 Intrinsic Problems with Just Tuning

Just intervals have the beauty of rich consonance. Unfortunately, for fundamental, mathematical reasons, they are impossible to realize fully in a composition, as we will observe in the following three examples.

1. In the key of C, D-F is a minor third and yet has a frequency ratio of $(4/3)/(9/8) = 33/27$ instead of the standard Just minor third of $6/5$. Thus, there are two different minor thirds in the scale. The ratio $32/27$ is rather harsh and defeats the goal of Just tuning. We would hope that this interval could be avoided; unfortunately, mathematical analysis shows this to be impossible.
2. Suppose a violinist tunes the strings with fifths having $3/2$ frequency ratios. Recall that the A-string is A_4. Then $f_{E_5} = (3/2)\, f_{A_4}$ and $f_{D_4} = (2/3)\, f_{A_4}$.

   Now, suppose that the violinist plays a B_4 so as to be a Just interval with respect to E_5, which is a fourth above. Then $f_{E_5}/f_{B_4} = 4/3$ or

$$f_{B_4} = \frac{3}{4} f_{E_5} = \left(\frac{3}{4}\right)\left(\frac{3}{2}\right) f_{A_4} = \left(\frac{9}{8}\right) f_{A_4}.$$

   We now compute the interval between this B_4 and the D_4 below:

$$\frac{f_{B_4}}{f_{D_4}} = \frac{\frac{9}{8} f_{A_4}}{\frac{2}{3} f_{A_4}} = \frac{27}{16}.$$

   This interval is the major sixth in Pythagorean tuning. It differs from the Just major sixth of $5/3$ by a small amount, the interval being equal to $(5/3)/(27/16) = 80/81$.

   The interval $81/80$ is well known, being referred to as the **syntonic comma**. While it is equivalent to only about one-fifth of a semitone, it is large enough to make the interval $27/16$ noticeably dissonant. We will see the syntonic comma a number of times in this chapter since it is ubiquitous in the mathematics of tunings.[8,9]

   Our result is summarized in Fig. 11.14.

   Alternatively, one could choose to play the major sixth between the D and the B as a Just major sixth, that is, $5/3$. But then, the interval of a major fourth

[8] It is interesting to calculate the ratio $80/81$ using a pocket calculator. The ratio consists of a repeated set of integers – called a "repeated decimal." Amazingly, each digit from 1 to 9 appears once and only once. The reader should also calculate other ratios, such as $31/81$ or $13/81$.

[9] The syntonic comma is often mistaken for the **Pythagorean comma**. The latter is the interval that we obtain if we start with a base frequency and increase it by a product of twelve fifths, each with a ratio of $3/2$. Thus, if we start with the note F, we would obtain the series of notes: $F \Rightarrow C \Rightarrow G \Rightarrow D \Rightarrow A \Rightarrow E \Rightarrow B \Rightarrow F\# \Rightarrow C\# \Rightarrow G\# \Rightarrow D\# \Rightarrow A\# \Rightarrow E\#$. $E\#$ is enharmonic with F and is very close to being seven octaves above (not below as we might expect) the original note F by a factor of $(3/2)^{12}/2^7 \sim 1.0136...$. This number is the **Pythagorean comma**.

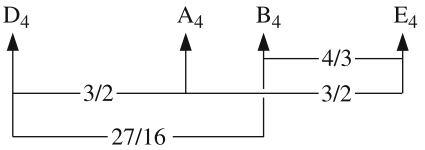


Fig. 11.14 Syntonic comma reflected by the ratio 27/16

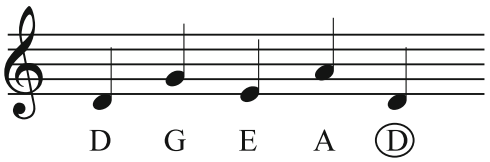


Fig. 11.15 Problematic sequence of notes

between the B and the E would be $[(3/2)(3/2)]/(5/3) = 27/20$, rather than the Just ratio of $4/3$. The ratio between these two numbers is $(27/20)/(4/3) = 81/80$, the syntonic comma.

We see then that it is absolutely impossible to perform pieces fully with just intervals between all pairs of significant notes. Furthermore, the errors are significant.

Now note that we have pieces wherein the violinist has to play these three notes, D, B, & E in rapid sequence, with open D and E strings desired, or worse yet, with the pair DB followed immediately by the pair BE. There would be too little time for the violinist to move the finger for the B to produce two Just intervals. We see that the violinist would have to compromise a bit and to make the two intervals sound acceptable. On the other hand, there is a natural help here: if the violinist plays the B with a **vibrato**,[10] the desired consonances are not badly affected by the lack of Just intervals.

3. As we progress through a piece of music, we might choose to have the interval between every neighboring note be a Just interval. If we do so, we may follow such a path that there is no guarantee that we return to the frequencies of the initial notes. An example of such a problematical sequence of notes is seen in Fig. 11.15.

   Consider the intervals between neighboring notes:

   – D to G above: perfect fourth
   – G to E below: minor third
   – E to A above: perfect fourth
   – A to D below: fifth

   With Just intervals between neighboring notes, we obtain

$$f_D = f_D \times \frac{4}{3} \times \frac{5}{6} \times \frac{4}{3} \times \frac{2}{3} = \frac{80}{81} f_D \neq f_D.$$

[10] See Sect. 2.15.

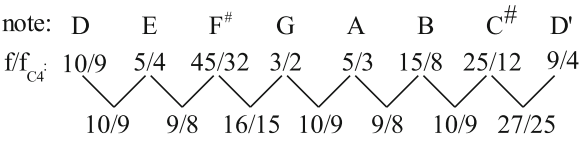


Fig. 11.16 Major D scale

If Pythagorean intervals are used, the above problem is avoided; however, their major and minor thirds are less consonant than corresponding Just intervals.

Note that the above sequence, DGEA, is the theme of the first movement of Ralph Vaughan-Williams' Symphony #8. It is an interesting fact that the resulting A, as determined above, has a ratio of 40/27 with respect to the opening note of D instead of 3/2, and results in a psychoacoustic tension that is musically exciting.

Homework: Check this last statement by calculating the same product using Pythagorean intervals. Recall that the Pythagorean minor third is the ratio 32/27.

> ### Note
>
> I have been told that a Mozart opera was analyzed in the above manner to see whether the final keynote had any shift relative to the opening note. One would expect shifting by 80/81 or 81/80 to occur many times in a long piece. Upward shifts by $81/80$ and downward shifts by $80/81$ would not be expected to cancel if they were random, much as when a drunkard carries on a random walk: Starting from the center of the top of a mesa, we would certainly expect the drunkard to eventually fall off. Amazingly, after the entire length of the Mozart opera, there was no shift at all! This result probably reflects well of the quality of Mozart's music.

4. Let us now examine the Diatonic Major scale in the key of D, using the frequencies that were determined earlier **based on the key of C**. We need two sharps from the chromatic scale, $F^{\#}$ and $C^{\#}$. The result we obtain is depicted in Fig. 11.16.

   Note that the last semitone is equal to $27/25$, rather than $16/15$. The ratio of the two numbers is $(27/25)/(16/15) = 81/80$, the syntonic comma.

   In order to handle such a change, harpsichords used to be repeatedly tuned, according to the key of the piece being performed and in the manner by which we constructed the scale for the key of C, above. In order to avoid this labor, equal-tempered tuning was introduced. We discuss this tuning in the next section.

11.8 Equal Tempered Tuning

In **Equal Tempered Tuning**, the *octave is divided into 12 equal semitone intervals*. The gain is that the problems of Just tuning are removed. The loss is that resonances

Table 11.1 A comparison of some intervals in Just vs. equal tempered tuning

Interval	Just (J)	Equal tempered (ET)
Semitone	$16/15 = 1.066\ldots$	$2^{1/12} = 1.059\ldots$
Whole tone	$9/8 = 1.125$, $10/9 = 1.11\ldots$	$2^{2/12} = 1.122\ldots$
Minor third	$6/5 = 1.2$	$2^{3/12} = 2^{1/4} = 1.189\ldots$
Major third	$5/4 = 1.25$	$2^{4/12} = 2^{1/3} = 1.26\ldots$
Fourth	$4/3 = 1.333\ldots$	$2^{5/12} = 1.335\ldots$
Fifth	$3/2 = 1.5$	$2^{7/12} = 1.498\ldots$
Minor sixth	$8/5 = 1.6$	$2^{8/12} = 2^{2/3} = 1.587\ldots$
Major sixth	$5/3 = 1.666\ldots$	$2^{9/12} = 2^{3/4} = 1.68\ldots$
Octave	2	2

and consonances are not as strong as they are in Just tuning. Consonant intervals such as the fifth or the major third are harsher. The greatest gain, perhaps, is that fixed tuned instruments such as the piano do not have to be retuned for each key change. Furthermore, single compositions that have changes of key can be played in a consistent manner on fixed tuned instruments, that is, without biasing one key.[11]

For equal tempered tuning, all we need is to determine the frequency ratio corresponding to a semitone. Label it with the symbol r. Since there are 12 semitones to the octave, we must have

$$r^{12} = 2$$

$$r = 2^{1/12} = 1.05946\ldots.$$

For comparison sake, the Just semitone is $16/15 = 1.066\ldots$ and is therefore slightly larger. To compensate for this increased value within the octave, Just tuning has a whole tone interval, $10/9 = 1.1111\ldots$, which is less than the equal tempered whole tone (= two semitones) interval of

$$r^2 = 2^{2/12} = 2^{1/6} = 1.12246\ldots.$$

On the other hand, recall that Just tuning uses two different whole tones; the other Just whole tone interval of $9/8 = 1.125$ is *greater* than r^2.

Table 11.1 compares the important intervals in Just (J) and Equal Tempered (ET) tuning.

[11] Bach's **Well-Tempered Clavier**, a collection of 48 short pieces, two for each of the possible keys (including only those with at most one sharp or flat), was famous in promoting what is referred to as one of the numerous **well tempered tunings**. According to Anton Kellner (http://plaza.ufl.edu/wnb/baroque_temperament.htm#German) Bach's temperament was not at all equal tempered. One possibility is that it was similar to one of the tunings of **Andreas Werkmeister** (1645–1706) (http://en.wikipedia.org/wiki/Werckmeister_temperament). See the problem on Werkmeister tuning at the end of the chapter.

11.9 The Cents System of Expressing Musical Intervals

To express intervals that are smaller than a semitone in a quantitative way, so as to reflect very small changes of pitch, the equal tempered semitone, $r = 2^{1/12}$, is divided into one hundred (100) units, called **Cents**, as follows:

$$100 \text{ cents} = 1 \text{ ET semitone}$$
$$1{,}200 \text{ cents} = 1 \text{ octave}$$

Generally, in the cents system the interval between two frequencies, f_2 and f_1, is given by:

$$\mathcal{C} \equiv \frac{1{,}200}{\log 2} \log \frac{f_2}{f_1}. \tag{11.5}$$

where the unit of the parameter C is the number of cents in the interval.[12]

We have here an explicit, *mathematical expression for the musical interval as a function of the corresponding frequency ratio*.[13]

[12]This equation can be written in a simpler form using the logarithm to the base 2 instead of the current base 10:

$$\mathcal{C} = \log_2 \frac{f_2}{f_1}. \tag{11.6}$$

[13]"Musical instrument digital interface" (**MIDI**) is a protocol for allowing various devices, such as musical instruments and computers, to communicate musical scores. It gives a numerical value to all notes based on the frequency. The value is clearly closely related to our sense of pitch and is given by

$$\mathbf{p} = 69 + 12 \log_2 \frac{f}{440}. \tag{11.7}$$

Note that $\mathbf{p} = 69$ is associated with A440. The number 69 was chosen so that the C below A440 will have the value $\mathbf{p} = 60$. In addition, for each semi-tone in ET $\mathbf{p}$ changes by unity:

$$\Delta\mathbf{p} = 12 \log_2 \frac{f_2}{f_1} = 12 \log_2 2^{1/12} = 1. \tag{11.8}$$

A related objective measure of pitch is the **mel scale** defined by the equation

$$m = \frac{1{,}000}{\log\ 2} \log\left(\frac{f}{1{,}000} + 1\right). \tag{11.9}$$

Note that $m = 1{,}000$ mels at a frequency of 1,000 Hz and zero mels at 0 Hz.

Discussion and applications of this equation follow in the examples below.

Sample Problem 11-1

Calculate the number of cents in an ET semitone.

Solution

We have

$$\frac{f_2}{f_1} = 2^{12}. \tag{11.10}$$

$$\mathcal{C} = \frac{1{,}200}{\log 2} \log\, 2^{12} = \frac{1{,}200}{\log 2} \times \frac{1}{12} \times \log 2 = 100\,\text{cents}.$$

Sample Problem 11-2

What is the frequency ratio corresponding to 1 cent?

Solution

Since a ratio of $2^{1/12}$ corresponds to 100 cents, we have

$$\frac{f_2}{f_1} = \left(2^{1/12}\right)^{1/100} = 2^{1/1{,}200} \simeq 1.00058.$$

As a check, we note that

$$\begin{aligned}\mathcal{C} &= \frac{1{,}200}{\log 2} \log 2^{1/1{,}200} \\ &= \left(\frac{1{,}200}{\log 2}\right)\left(\frac{1}{1{,}200}\right) \log 2 = 1\,\text{cent}.\end{aligned}$$

Sample Problem 11-3

By how many cents do the Just and equal tempered fifth differ?

Solution

For the J-fifth, we have

$$\mathcal{C} = \left(\frac{1{,}200}{\log 2}\right) \log(3/2) \cong 702\,\text{cents}.$$

while for the ET-fifth (7 semitones) we have precisely $C = 700$ cents.

Therefore, the ET-fifth is 2 cents ($\sim 2/100 = 1/50$ or 2% of a semitone) smaller than the J-fifth.

We now demonstrate how **adding intervals in cents corresponds to multiplying ratios**.

Suppose that we have three frequencies, $f_3 > f_2 > f_1$. We know that the interval $f_1 \to f_3$ must equal the *sum* of the two intervals $f_1 \to f_2$ and $f_2 \to f_3$. This fact is reflected mathematically as follows. Let:

$$\mathcal{C}_{21} = \text{Interval for the frequency ratio } \frac{f_2}{f_1}$$

$$\mathcal{C}_{32} = \text{Interval for the frequency ratio } \frac{f_3}{f_2}$$

$$\mathcal{C}_{31} = \text{Interval for the frequency ratio } \frac{f_3}{f_1},$$

so that we can then add them to become

$$\mathcal{C}_{32} + \mathcal{C}_{21} = \frac{1{,}200}{\log 2} \log \frac{f_3}{f_2} + \frac{1{,}200}{\log 2} \log \frac{f_2}{f_1}. \tag{11.11}$$

or

$$\mathcal{C}_{32} + \mathcal{C}_{21} = \frac{1{,}200}{\log 2} \left[\log \frac{f_3}{f_2} + \log \frac{f_2}{f1} \right]. \tag{11.12}$$

or

$$\mathcal{C}_{32} + \mathcal{C}_{21} = \frac{1{,}200}{\log 2} \log \frac{f_3}{f_1} = C_{31}. \tag{11.13}$$

That is, $\mathcal{C}_{32} + \mathcal{C}_{21} = \mathcal{C}_{31}$.

The cents system is convenient for expressing in a precise quantitative way musical intervals. Below, we exhibit a table that compares the three tunings discussed in this chapter with respect to the musical interval from C to various notes in the chromatic scale. Frequency ratios are given as well as the number of cents.

11.10 Debussy's Six-Tone Scale

Given the power of impressionistic music, such as the music of **Claude Debussy** or **Maurice Ravel**, it is worth mentioning Debussy's **six-tone Scale**. These scales give the listener a powerful sense of the mystical. Why this is so is not clear. There are two scales:

F	G	A	B	$C^{\#}$	$D^{\#}$
$F^{\#}$	$G^{\#}$	$A^{\#}$	C	D	E

Each scale consists of ET whole-tone intervals. They are often played juxtaposed against each other.

11.11 Terms

- Building a musical scale
- Cent
- Cents system of expressing musical intervals
- Chromatic scale
- Consonance
- Diatonic scale
- Dissonance
- Enharmonic equivalents
- Equal tempered tuning
- Fifth
- Fix tuned instrument
- Flat
- Fourth
- Intonation
- Just temperament
- Just scale
- Key note
- Major second
- Major sixth
- Major third
- Minor second
- Minor third
- Minor sixth
- Musical interval
- Musical scale
- Musical staff
- Octave
- Pentatonic scale
- Pythagorean comma
- Pythagorean temperament
- Pythagorean scale
- Semitone
- Seventh
- Sharp
- Six-tone scale (of Debussy)
- Syntonic comma
- Temperament
- Tuning
- Vibrato
- Well tempered tuning
- Werkmeister temperament
- Whole tone

11.12 Important Equations

- Objective numerical expression of the musical interval between two frequencies, f_1 and f_2

$$\frac{f_2}{f_1}.$$

- Frequency ratio for musical intervals in equal tempered intonation:

$$\text{frequency ratio} = 2^{n/12}, \tag{11.14}$$

  where n is the number of semitones.
- Equation defining **cents** in relation to frequency ratio

$$\mathcal{C} \equiv \frac{1{,}200}{\log 2} \log \frac{f_2}{f_1}. \tag{11.15}$$

11.13 Problems for Chap. 11

1. One tone is produced with a frequency of 150 Hz. The tone three octaves higher is produced with a frequency of (choose one).

   (a) 300 Hz
   (b) 450 Hz
   (c) 600 Hz
   (d) 900 Hz
   (e) 1,200 Hz

2. (a) How many *semitones* are there in each of the following intervals: major second, minor third, major third, fourth, fifth, minor sixth, major sixth, octave?
   (b) What are the frequency ratios of the above intervals in equal tempered tuning?
   (c) Show that the sum of a major sixth and a minor third is an octave by adding up the number of semitones for each interval. Check whether the product of the corresponding frequency ratios is two for Pythagorean tuning, as shown in the table of musical intervals Table 11.2.
3. (a) What advantages does equal tempered tuning have over Just tuning?
   (b) What advantages does Just tuning have over equal tempered tuning?
4. Suppose the note A above middle-C is tuned at 440 Hz.

   (a) To what frequency should the E above be tuned to produce a *Just* fifth in the key of C? Note that the third harmonic of the A will equal the second harmonic of the E, so that there will be *resonance*.
   (b) To what frequency should the E above be tuned so as to produce an *equal-tempered* fifth?
   (c) In the latter case, what will be the **beat frequency** between the third harmonic of the A and the second harmonic of the E?

5. We learned that in Just tuning, there are two different frequency ratios for a whole tone. Show that the ratio of these two ratios is a syntonic comma, 81/80.
6. The **tritone** is a sum of a perfect fourth and a minor second and can be regarded as half of a full octave. The interval is famous for its association with horror and gloom. Calculate the corresponding frequency ratio in both Just tuning and equal temperament.
7. The six strings of a guitar are tuned to the following notes, in order of increasing pitch:

   $$\text{E} - \text{A} - \text{D} - \text{G} - \text{B} - \text{E''}$$

   so that the two E's – E and E" – are two octaves apart.

   (a) What are the intervals (by name) D-G and G-B?
   (b) Suppose all the above intervals are set as JUST intervals. What will be the frequency ratio – $f_{E''}/f_E$? (It will **not** be 4 : 1!)

Table 11.2 Table of musical intervals

NOTE		Pythagorean frequency	Cents	Just frequency	Cents	Equal temperament frequency	Cents
C		1	0	1	0	1	0
	$C^{\#}$	2,187/2,048	114	25/24	71	$2^{1/12}$	100
	D^{b}	256/243	89	16/15	112	$2^{1/12}$	100
D		9/8	204	9/8	204	$2^{1/6}$	200
	$D^{\#}$	19,683/16,384	317	75/64	275	$2^{1/4}$	300
	E^{b}	32/27	294	6/5	316	$2^{1/4}$	300
E		81/64	409	5/4	386	$2^{1/3}$	400
	$E^{\#}$	177,147/131,072	522	125/96	457		
	F^{b}	8,192/6,561	385	32/25	427		
F		4/3	498	4/3	498	$2^{5/12}$	500
	$F^{\#}$	729/512	612	45/32	590	$2^{1/2}$	600
	G^{b}	1,024/729	588	36/25	631	$2^{1/2}$	600
G		3/2	702	3/2	702	$2^{7/12}$	700
	$G^{\#}$	6,561/4,096	816	25/16	773	$2^{2/3}$	800
	A^{b}	128/81	792	8/5	814	$2^{2/3}$	800
A		27/16	906	5/3	884	$2^{3/4}$	900
	$A^{\#}$	59,049/32,768	1,019	225/128	977	$2^{5/6}$	1,000
	B^{b}	16/9	996	9/5	1,018	$2^{5/6}$	1,000
B		243/128	1,109	15/8	1,088	$2^{11/12}$	1,100
	$B^{\#}$	531,441/262,144	1,228	125/64	1,159		
	C^{b}	4,096/2,187	1,086	48/25	1,129		
C		2	1,200	2	1,200	2	1,200

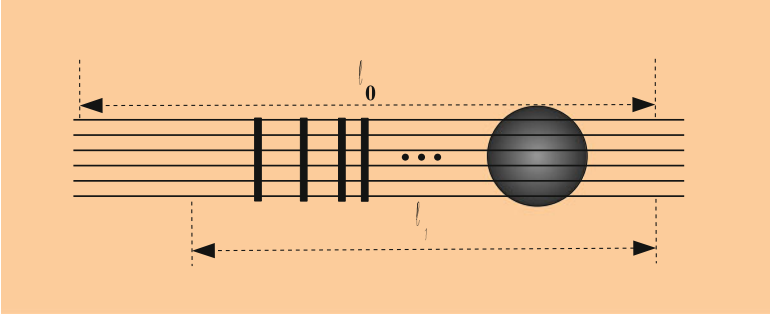


Fig. 11.17 Guitar frets

(c) By how many cents does $f_{E''}/f_E$ differ from 2,400 cents (corresponding to two octaves)?
(d) Show how the use of equal temperament eliminates this discrepancy.

8. The frets on a guitar are set so as to produce equal tempered frequencies as shown in Fig. 11.17.

Let ℓ_0 be the length of the "open" string (without any fingers on the frets). Let ℓ_n be the length of string that is free to vibrate corresponding to the nth note above that of the open string. The length ℓ_1 is shown in Fig. 11.17. We see in the figure how the spacing between neighboring frets decreases as the frequency is increased.

Show that the length ℓ_n is given by

$$\frac{\ell_n}{\ell_0} = 2^{-n/12}. \tag{11.16}$$

This function is *mathematically* similar to the exponential decay in the attenuation of a wave. Note that

$$\log\ \ell_n = \log\ \ell_0 - n\ \frac{\log\ 2}{12}. \tag{11.17}$$

As a result, a plot of the *logarithm* of the length ℓ vs. n will yield a straight line with a slope given by $-(\log\ 2)/12$.

9. One summer, at a day camp of two of my grandchildren, we were sitting outdoors on a lawn that was close to a small airport. Suddenly a propeller-driven airplane passed directly overhead. We could hear a distinct tonality to the whirring sound that varied from one note to another that was a major third below.

Use this information to estimate the speed of the airplane.
HINT: **Refer to Problem 32 in Chap. 8**.

Table 11.3 Table of intervals in Just tuning in the key of D – J(D)

NOTE	f/f_C in J(C)	Frequency in J(C)	f/f_D in J(D)	Frequency in J(D)
C	1	264		
D	9/8	297	1	297
E	5/4		10/9	
F	4/3			
$F^{\#}$	45/25			
G	3/2			
A	5/3			
B	15/8			
$C'^{\#}$	25/12			
D'	9/4		2	

10. Suppose that the following musical passage is played with **Just intervals between neighboring notes**:

Major second	9/8	Minor third	6/5	Major third	5/4
Fourth	4/3	Fifth	3/2		

$$C \to E \to A \to B \to E' \to A \to F \to D \to C$$

(a) Find the frequency ratios for all the neighboring notes –

$$\frac{f_E}{f_C}, \quad \frac{f_A}{f_E}, \quad \frac{f_B}{f_A}, \quad \frac{f_E}{f_B}, \quad \ldots$$

(b) Show that the ratio of the frequency of the last C to the first C in the passage is equal to $80/81$ = inverse of the **syntonic comma**.

(c) Evaluate $80/81$ using a pocket calculator, exhibiting all decimal places in your answer.

(d) To how many cents does a syntonic comma correspond?

11. Let J(C) represent Just tuning in the key of C, as displayed in table of musical intervals and in Fig. 11.13. Suppose that we choose the frequency of the C, as $f_C = 264\,\text{Hz}$. This choice corresponds to a frequency of 440 Hz for A-440 in J(C). Suppose further that we have tuned a harpsichord in J(C) in order to play a piece in the key of C and then want to play another piece in the key of D.

A central question is: How will the frequencies of the strings compare to what we would need to have them tuned to J(D), that is, Just tuning in the key of D? You will use the table below to exhibit the results of your study.

(a) In order to answer this question, first calculate the actual frequencies in J(C) for the notes in the diatonic scale in the key of D: These notes are D E $F^{\#}$ G A B $C'^{\#}$ D'. Note that in the table, the ratios of the frequencies are exhibited, as taken from Table 11.2. You should fill in the third column with these frequencies of Table 11.3. The frequencies can be determined by multiplying $f_C = 264\,\text{Hz}$ by the corresponding ratios in the second column.

(b) Fill in the fourth column with the ratio of each these frequencies to the frequency $f_D = 297\,\text{Hz}$: 1, 10/9, ...

(c) You will next obtain the frequencies in J(D) using $f_D = 264\,\text{Hz} \times (9/8) = 297\,\text{Hz}$. All you have to do is to multiply 297 by the correspond ratios in the second column – that is, 1, 9/8, ... 15/8, 2. Enter the resulting frequencies in the fifth column.

The differences between the corresponding frequencies in J(D) and J(C) help us to appreciate the problem in using Just tuning in situations where one wants to play a piece of music within which the key changes.

12. Suppose a car is moving toward you and produces a **Doppler effect** on its tooted horn corresponding to a Just semitone ($f'/f = 16/15$). Determine how fast the car is moving.

13. **Stretch Tuning of Pianos**

In Chap. 2, we learned that the frequency spectrum of a string is not a harmonic series on account of stiffness. In particular, the frequency of the nth mode is given by[14]

$$f_n = n\frac{\sqrt{\frac{\mathcal{T}}{\mu}}}{2\ell}\sqrt{1+\mathcal{B}n^2}. \tag{11.18}$$

An important consequence is that the harmonics of a given piano string will not be consonant with some of the harmonics of other strings, as discussed in Chap. 2 and we will lose resonance. Of course, the use of equal temperament tuning already destroys this resonance *except for* the set of octaves of a given note.

Consider for simplicity the A-440 string and the string an octave above, which we will refer to as A-880. All tunings assign a fundamental frequency of $f_1^{440} = 440$-Hz for the A-440. Moreover, all of the standard tunings discussed in this chapter would assign a frequency of 880 Hz to the A-880. The combination of these two frequencies will produce a consonant sound. On the other hand, the first overtone of the A-440 will be slightly higher than 880 Hz on account of stiffness.

[14] In detail, the constant $\mathcal{B}$ is given by

$$\mathcal{B} = \frac{\pi a^2 Y}{\mathcal{T}}\left(\frac{\pi a}{2\ell}\right)^2, \tag{11.19}$$

where a is the radius of the string (assumed to be a solid cylinder), Y is Young's modulus of the string's material, $\mathcal{T}$ is the tension, and ℓ is the length of the string. Readers who have a background in the basic physics of elasticity will recognize that the ratio $(\mathcal{T}/(\pi a^2 Y))$ is the fractional increase $\Delta\ell/\ell$ in the length of the string due to the tension.

(a) Show that the frequency of the first overtone of A-440 in the presence of stiffness is given by

$$f_2^{440} = 880\frac{\sqrt{1+4\mathcal{B}}}{\sqrt{1+\mathcal{B}}}. \tag{11.20}$$

Note that if A-880 is tuned to this frequency, we would sacrifice the resonance in our hearing a 2:1 ratio of frequencies for resonance between the two strings, A-440 and A-880. Such is the case in **stretch tuning of pianos**.

(b) Calculate this frequency if $\mathcal{B} = 0.008$.

(c) What would be the beat frequency between 880-Hz and this frequency?

(d) Calculate the interval $\mathcal{C}$ in cents for the two frequencies f_2^{440} above and 880-Hz.

Now let us examine the next octave above A-440, namely, A-1760. We have two choices to make if we want to achieve resonance between the three strings under consideration thus far: We could tune A-1760 to so as to have a frequency equal to the second overtone of A-440 or we can choose a frequency equal to the first overtone of A-880.

(e) Explain why these two choices lead to different frequencies. You can simplify your response with formulas if you assume that the A-440 and A-880 string have the same value of the constant $\mathcal{B}$.

(f) Suppose that we apply one of the above stretch tunings for each of the 11 notes in the octave lying between A-440 and A-880. Will such a stretch tuning lead to resonances between corresponding harmonics of two different notes in the octave?

14. Recently, because of the development of electronic keyboards and pianos, one can change the frequencies of the notes in a small fraction of a second. As a result, access to temperaments used centuries ago that have a different sound (some might say more consonant sound) is available and is attractive to some musicians. Many expensive keyboards and pianos now provide these temperaments. In this problem, you will learn about one of these temperaments **Werkmeister I(III) Temperament**. It is based on the Pythagorean temperament.

(a) Consider the sequence of notes starting with C and increasing by fifths: C, G, D, A, E, B, $F^{\#}$, $C^{\#}$, $G^{\#}$. Thus, $G^{\#}$ is eight fifths above C.

Find the frequency $f_{G^{\#}}$ in proportion to C in Pythagorean tuning by multiplying by factors of (3/2) and reducing to the octave.

(b) Note that the interval C-E (a major third) is a series of four fifths followed by a reduction down by two octaves. The Pythagorean ratio is $(3/2)^4/2^2 = 81/64$ and differs significantly from the more consonant (5/4) of Just tuning.

Show that the ratio of these two ratios is the **syntonic comma** (81/80).

(c) **Werkmeister I(III)** temperament was introduced by Andreas Werkmeister. It "**well-tempers**" Pythagorean temperament most of all to produce a Just major third. To do so, Werkmeister makes a compromise. Instead of

producing the Pythagorean scale by multiplying by a sequence of ratios (3/2), he multiplies an additional ratio of $(80/81)^{1/4}$ for each fifth. The ratio $(81/80)^{1/4}$ is referred to as a **quarter comma**; thus, each multiplication by (3/2) is accompanied by a **reduction** by a quarter comma.

Show that the result is that the interval between C and E is a Just major third.

Note that the intervals between C and notes other than E are no longer Just. In particular, the interval of fifth between C and G is not equal to the "sacred value" of 3/2 of Just Temperament. Rather, it is reduced by a quarter comma, or $(80/81)^{1/4}$.

(d) Evaluate the correction factor $(80/81)^{1/4}$ and express it in cents.

(e) Consider the sequence of notes starting with C and decreasing by four intervals of a fifth each: C, F, B^b, E^b, A^b.

Find the frequency f_{A^b} in proportion to C in Pythagorean tuning by dividing by factors of (3/2) and reducing to the octave. Note first that the resulting f_{A^b} is a Just interval above C. Note, too, that $f_{A^b} \neq f_{G^\#}$. (You can show that this inequality persists with Werkmeister I(III) temperament.)

15. Flutists often have a tendency to sway when they play. We might wonder whether this motion can have a serious effect due to the Doppler effect. To answer this question, we need to know how much of a shift in frequency is tolerable. Tests reveal that a good listener can detect a frequency shift of about 3.5 Hz for an A440. This is the **just noticeable difference of pitch** vs. frequency. (Experienced ears can do better.)

Determine the velocity that will lead to a Doppler shift of 3.5 Hz.

Chapter 12
The Eye

The human eye carries out its function of converting information contained in light to nerve impulses which are sent to the brain in a manner which, to a great extent, parallels the operation of a camera. This fact is illustrated in Fig. 12.1 on the following page. Scientific American article by **George Wald**, in the August, 1950 issue.

The principle structures of the eye are the:

- **Cornea** – it acts like a primary lens
- **Iris** – it provides a variable aperture
- **Lens** – it provides a variable focal length
- **Retina** – it acts like camera film in responding to light by producing nerve impulses which are sent to the brain down the **optic nerve**

The eye serves to provide:

1. A two-dimensional representation of a scene, including both **light intensity** and a **sense of color** as a function of position in space
2. A sense of distance of light sources from our eyes – that is, **depth perception**

12.1 The Cornea and Lens

To appreciate fully this chapter, you should have a good understanding of the material on lenses in Chap. 8. The lens and cornea act as a **compound lens** system with variable effective focal length f. The effective image distance of the eye-lens system, d_{ie}, is *fixed* because the positions of the lens and retina are fixed. This effective distance is about 24 mm. According to the lens equation (see (12.1) below), the farther an object is, that is, the larger the object distance d_o is, the larger must be the focal length. At infinite object distance, the effective focal length f will be a maximum and equal to the image distance d_{ie}, that is about 24 mm. Conversely, the closer an object is, the smaller must be the focal length. Surprisingly, we will see

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_12,

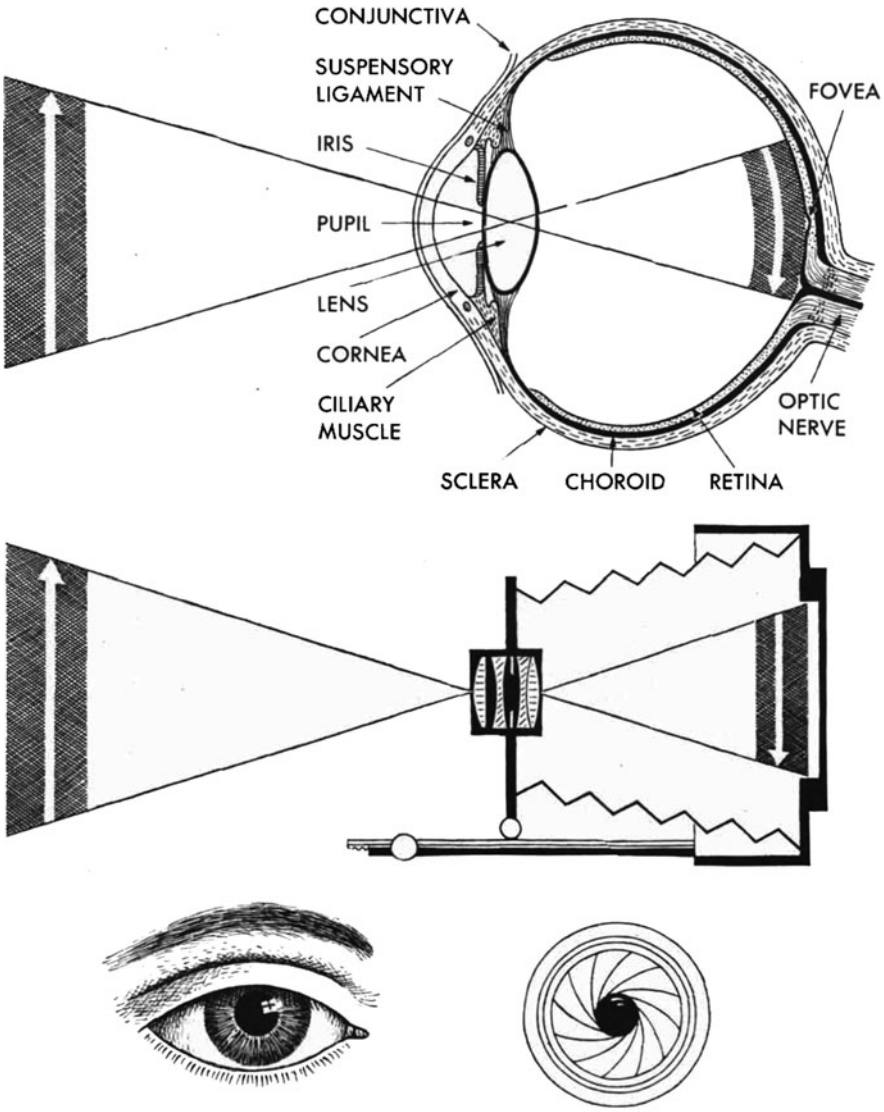


Fig. 12.1 The camera and eye compared (source: **George Wald**. Reproduced with permission. © (1950) *Scientific American*, Inc. All rights reserved.)

that the normal requirement of being able to see objects, whose distance ranges from as close as a foot or so all the way to infinity, do not require a very large variation in focal length. As usual, the focal length, object distance d_o, and image distance d_{ie} must satisfy the lens equation, which we write as

$$\frac{1}{f} = \frac{1}{d_o} + \frac{1}{d_{ie}}. \tag{12.1}$$

In order to increase the **focal length**, **ciliary muscles**, which act on the lens, must relax (not tighten) so as to flatten the lens. The process of varying the focal length is known as **accommodation**. Accommodation thus provides for a range of focal lengths, from some minimum, min **f** to some maximum, max **f**.

A **myopic** (or **near-sighted**) eye has a difficulty accommodating to great distances. It has a maximum focal length, max f, which is less than d_{ie}. As a result, the maximum object distance, max d_o that can be brought into focus is *not* at infinity, as one would wish. Using the lens formula we obtain:

$$\frac{1}{\max\ d_o} = \frac{1}{\max\ f} - \frac{1}{d_{ie}}. \tag{12.2}$$

To see near objects, the focal length of the eye must be decreased so that the lens must have a larger "bulging." The ciliary muscles are tightened to reduce f. A **hyperopic** (**far-sighted**) eye has a minimum focal length, min f, which is so large that the closest an object can be to the eye, and still be brought into focus on the retina, is too great for the person's needs. That closest distance, min $d_0 \equiv d_{np}$, is called the **near point**. We have:

$$\frac{1}{d_{np}} = \frac{1}{\min\ f} - \frac{1}{d_{ie}}. \tag{12.3}$$

Sample Problem 12-1

An eye has a retina that is 24.0 mm from the eye lens. It focuses an object that is 2.00 m away and has a height of 40 cm.

Find the focal length of the eye lens and the height of the image on the retina.

Solution

We have

$$\frac{1}{d_o} = \frac{1}{f} - \frac{1}{d_{ie}} \tag{12.4}$$

with $d_o = 200\,\text{cm}$, $d_{ie} = 200\,\text{cm}$, and $h_o = 40\,\text{cm}$.
Then

$$\frac{1}{d_o} = \frac{1}{200\,\text{cm}} - \frac{1}{2.4\,\text{cm}} = 0.422\,\text{cm}^{-1} \tag{12.5}$$

so that $f = 2.37\,\text{cm}$.
Note that f is extremely close to the image distance because of the large object distance. Also, relatively large changes in object distances do not require significant changes in the focal length.

Now we turn to the image height.
We do not need to know the focal length to determine the magnification.
It is determined by the object distance and the image distance.
Thus

$$h_{\mathrm{i}} = M h_{\mathrm{o}} = \left| \frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} \right| h_{\mathrm{o}} = \frac{2.4\,\mathrm{cm}}{200\,\mathrm{cm}} 40\,\mathrm{cm} = 0.48\,\mathrm{cm} = 4.8\,\mathrm{mm}. \quad (12.6)$$

> **Note**
>
> It is useful to appreciate the slight ambiguity of the near point: As a **home exercise**, estimate the near point of each of your eyes separately, as well as when used together. I suggest that you hold text in front of you having various font sizes. How different are the results?

As one ages, min f tends to increase because of increasing stiffness in the lens. This stiffening involves a process of crystallization of the lens material.

> **Note**
>
> For the *myopic* eye, min f < max f < d_{ie}, while for the *hyperopic* eye, min f **is too large for practical purposes**.

For an excellent applet that explains myopia and hyperopia, go to the website (accessible on 6/13/06) **http://webphysics.davidson.edu/physlet_resources/dav_optics/Examples/eye_demo.html**.

The lens is an excellent filter in the ultraviolet (UV) region of the spectrum. However, filtering extends into the violet end of the visible region. Thus, the fraction of incident light that is transmitted drops from close to 100% for $700 > \lambda > 500\,\mathrm{nm}$, to 15% for $\lambda = 400\,\mathrm{nm}$, to 0.1% for $\lambda = 365\,\mathrm{nm}$. (See Chap. 13 for further discussion on the **transmittance**, also called the **transmission coefficient**.) Thus, people whose lens has been removed because of cataracts experience a new world of vision in the UV.

12.2 The Iris

Experts claim that the human eye can see a steady light intensity, without suffering pain or eye damage, that spans about 12 orders of magnitude. (This range compares well with the span of 12 orders of magnitude of the ear.) To help the eye deal with such a huge range, the iris varies the aperture area from 0.02 to $0.3\,\mathrm{cm}^2$. This function of the iris is analogous to the role of the ossicles of the ear in protecting the ear from excessively loud sounds by restraining their mobility.

12.3 The "Humorous" Liquids of the Eye

The volume of the eye between the cornea and the lens is filled with a water-like liquid called **aqueous humor**. One of its purposes is to nourish the eye in place of blood, which would be opaque to light. Fluid is constantly flowing into and out of the chamber containing aqueous humor. Unfortunately, the pressure in the fluid can become excessive, resulting in the dangerous disease known as **glaucoma**.

The volume within the eye between the lens and the retina is filled with a gelatinous material called **vitreous humor**. Its consistency is close to that of egg white. Both humors serve to keep the eye's size and shape firm. Their index of refraction is about 1.3, in comparison with the index of refraction 1.4 of the lens. There is therefore refraction at the two **lens-humor interfaces**.

Clearly, both humors should remain as clear as possible so as not to block light rays heading for the retina. Fortunately, because many light rays emanating from a point source of light enter the eye on their way to being focused on a corresponding position on the retina, strands of material floating within the vitreous humor do not necessarily severely disturb the image produced on the retina.[1]

12.4 The Retina

A photograph of the retina is shown in Fig. 12.2, taken using an **ophthalmoscope**. This figure is a copy taken from the beautiful book by **R. L. Gregory** entitled *Eye and Brain*, (McGraw-Hill, New York, 1978). The retina contains a layer of two types of light-sensitive cells together referred to as **light receptors**. These are the **rods** and the **cones**, so called because of their respective shapes. In Fig. 12.3 we see a beautiful electron microscope image of rods and cones. They are capable of responding to the absorption of a single photon of light by producing a nerve impulse that is sent down a nerve fiber. Typically, though, a few photons are necessary to produce such excitation.

Near the middle of Fig. 12.2, we see a yellow region about 2 mm × 1.5 mm across, within which the optic nerve fibers pass through the retina, so that there are no rods or cones. This region produces a blind spot in our vision, which we usually don't notice. Towards to left of the yellow region is a dark orange region called the **macula**, with its even darker **fovea** at the center. The fovea is about one-millimeter in diameter. Fig. 12.4 is a schematic drawing of a section of the retina.

[1] In order to appreciate this fact: If you have an opportunity to be present when slides are being projected onto a screen by a slide projector, place some fingers over the lens and notice that your fingers do not cast a shadow on the screen; instead, the image on the screen is simply dimmed nonuniformly.

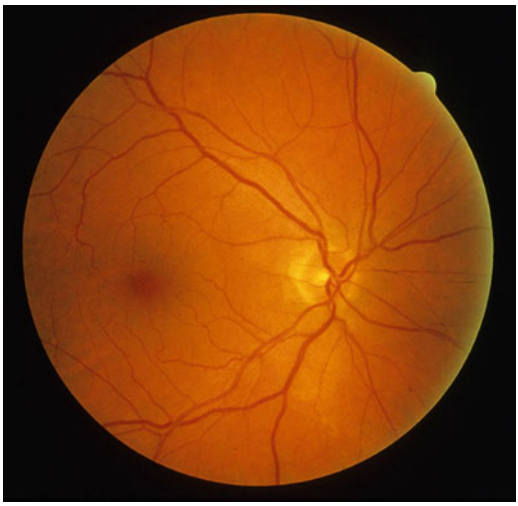


Fig. 12.2 The human retina (source: "Phenx Toolkit", funded by the National Human Genome Research Institute, https://www.phenxtoolkit.org/)

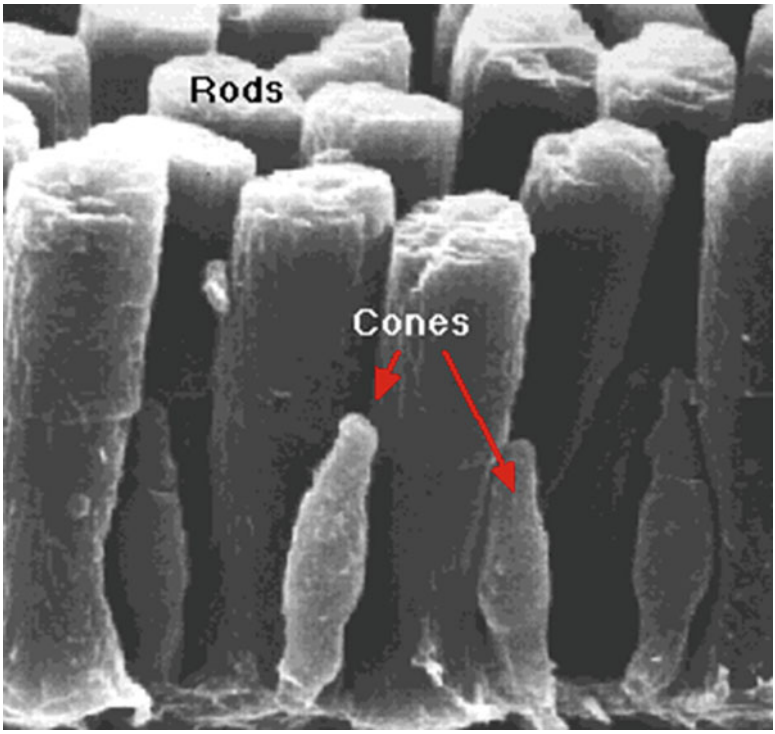


Fig. 12.3 Electron microscope image of rods and cones (source: http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/V/Vision.html$, courtesy of David Copenhagen; produced by David Copenhagen, Scott Mittman, and Maria Maglio)

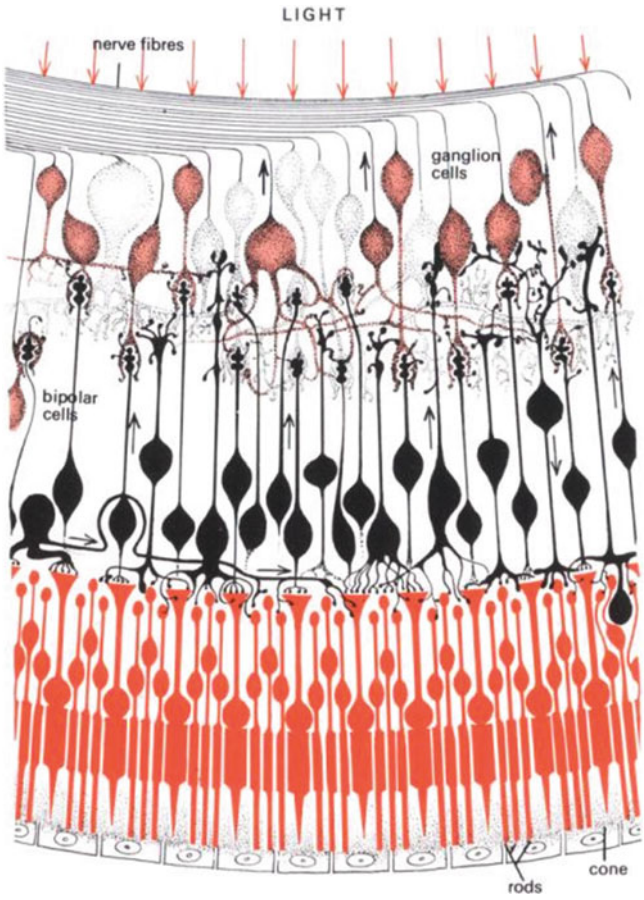


Fig. 12.4 Rods and cones in a section of the retina (source: GREGORY, Richard; *EYE AND BRAIN*. © 1990 Princeton University Press, Reprinted by permission of Princeton University Press)

Note the amazing fact that the nerve fibers leading to the optic nerve leave the retina within the vitreous humor, and thus lie in the path of incoming light. In contrast to invertebrates, vertebrates have this so-called **inverted retina**.

The following are some of the important characteristics of rods and cones:

Rods:

1. There exists one type of rod, with a peak frequency sensitivity at around 500 nm (greenish). As we will see in Chap. 14 on color vision, there being one type is connected with their not being used for color vision.
2. The rods are about 1,000 times more sensitive than the cones.

   This property is connected with their providing for **night vision**, referred to technically as **scotopic vision**.

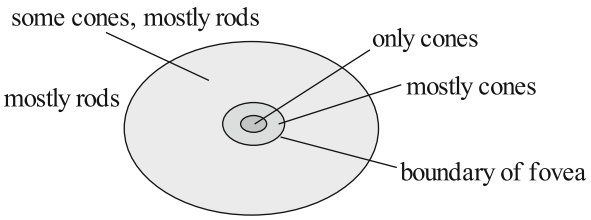


Fig. 12.5 Rough sketch of the distribution of rods and cones in the retina

3. Rods have a relatively long recovery time – about 25 min. That is, once they have been excited and have transmitted a nerve impulse, they require about 25 min to return to a receptive state. This property is connected with the observation that light temporarily **bleaches** rods from a purple color to a transparent state indicative of their inability to absorb light. The light-sensitive material in rods is called **rhodopsin** or **visual purple** and was extensively studied by **George Wald**, who received the Nobel Prize for his work.
4. About 120 million in number, rods are distributed mostly peripherally to the centrally located **fovea**, which has a diameter of about 1 mm. (See Fig. 12.5.)

Cones:

1. The cones provide for color vision by the existence of three types of cones. They have peaks in their sensitivity as a function of frequency at the following respective wavelengths: $\lambda \sim 440$ nm (blue), $\lambda \sim 520$ nm (green), and $\lambda \sim 570$ nm (orange). The three types are often called **blue cones**, **green cones**, and **red cones**, respectively. **Genes** for each type of cone have recently been isolated at Johns Hopkins University. How the three provide for color vision will be described later, in Chap. 14.
2. About six million in number, cones are concentrated in the region of the fovea. It is estimated that about 64% are red cones, 32% are green cones, and 2% are blue cones.
3. Cones provide for **day vision** or **photopic vision**.
4. As with all nerves, when a cone is excited and emits a nerve impulse, it takes time for it to be able to respond again. This time is referred to as the **recovery time**. Cones have a relatively short recovery time. Recovery time is responsible for **after images**. In Fig. 12.6 is a reproduction of a famous painting of the flag of the USA. The reader should stare at the center of the flag for about 30 s, with minimal eye motion. Then the reader should look to the side at a blank part of the page. The original red, white, and blue colors will appear. They are the **complementary colors** discussed in Chap. 14 to the colors in Fig. 12.6. This phenomenon results from fatigue of the cones corresponding to the color of a region.

In Fig. 12.7, we see the distribution along a line running from the top to the bottom of the eye that runs through the fovea and the blind spot at the exit to the optic nerve. From Figs. 12.5 and 12.7, it is clear that the rods provide for excellent peripheral

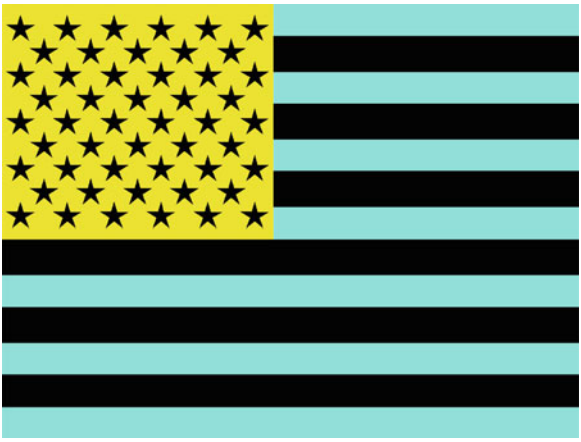


Fig. 12.6 Complement of the US flag

but poor head-on **scotopic** (*night*) vision. On the other hand, the cones provide for excellent head-on but poor peripheral **photopic** (*day*) vision. See Fig. 12.7.

12.5 Dark Adaptation

Suppose that you are in a well lit room and then suddenly turning off the lights. Most of us are familiar with the experience that we do not immediately see well. Our eyes need some time to adjust and increase their sensitivity to low light intensities. This adjustment is called **dark adaptation**. A minor reason is the need for our pupils to dilate – but that response is relatively fast – about 10–20 s.[2] Full adaptation, that is maximum sensitivity, is achieved by the cones in about 7 min, that by the rods in about 1 h. The reason is the fatigue of the rods and cones, which will be discussed in Chap. 14.

12.6 Depth Perception

Depth perception is apparently achieved by two primary means, both depending upon the use of two eyes:

1. **Convergence**: In order to produce the same field of vision on the two retinas, the two eyes must be turned through different angles. See Fig. 12.8. That difference decreases with increasing distance of the object.

[2] See Wikipedia (1-28-2011): http://en.wikipedia.org/wiki/Pupillary_light_reflex and M.H. Pirenne, *Vision and the Eye*, (Associated Book Publshers, London, 1987).

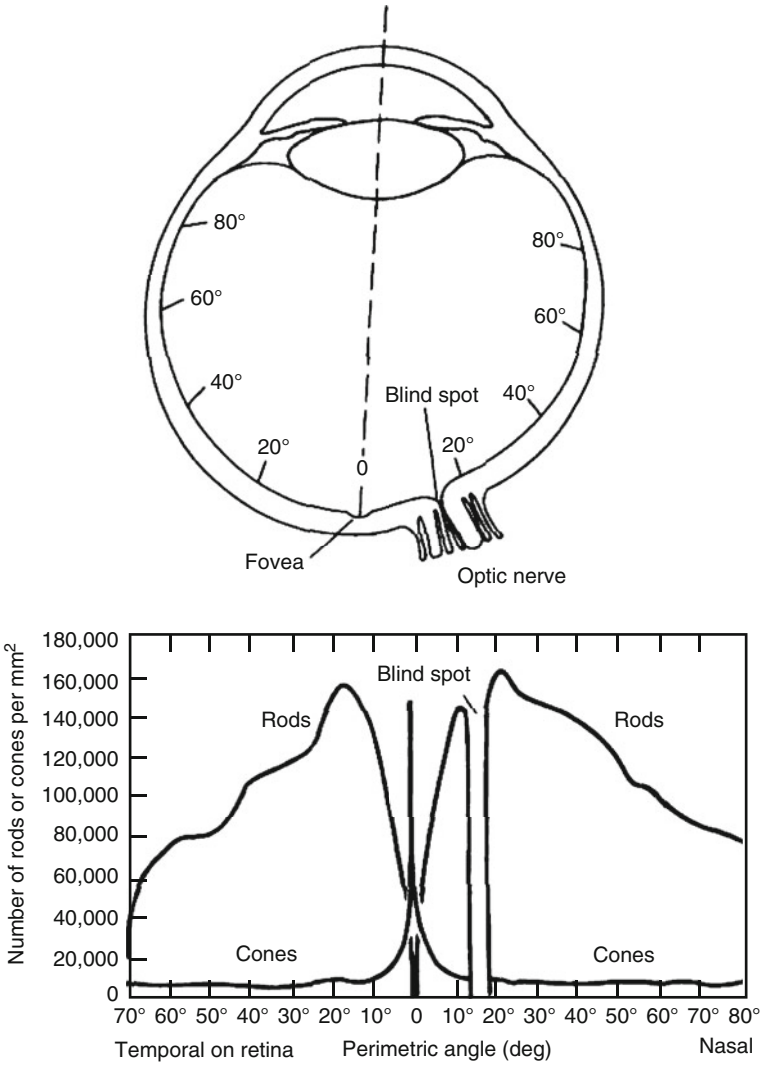


Fig. 12.7 Distribution of rods and cones (source: M.H. Pirenne, *Vision and the Eye*, (Associated Book Publshers, London, 1987))

2. **Disparity of eye position**: Because of the physical separation of the two eyes (by about 6.3 cm according to R. L. Gregory's *Eye and Brain*, the two retinas necessarily end up receiving images which are slightly shifted with respect to one another. This shift is interpreted by the brain in terms of the distance to the light source. **Stereoscopes** use this phenomenon to produce a sense of depth: The light coming from a given two-dimensional image – as on a photo – is split into two identical beams of light, with one beam incident upon one eye and the other

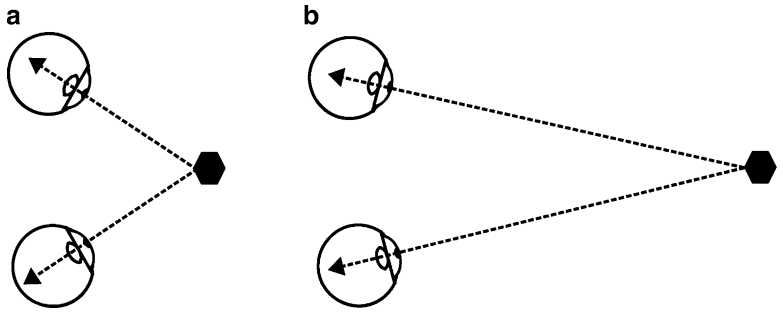


Fig. 12.8 Depth perception via convergence of eyes

upon the second eye. The beams are such that they are incident with different positions with respect to the two eyes, thereby reproducing the disparity of eye position of a real 3D scene.

12.7 Terms

- Accommodation
- Aqueous humor
- Ciliary muscles
- Compound lens
- Cones – blue, green, and red
- Convergence and disparity
- Cornea
- Day vision
- Depth perception
- Fovea
- Hyperopia (far-sightedness)
- Iris
- Lens
- Light receptors
- Myopia (near-sightedness)
- Near point
- Night vision
- Optic nerve
- Retina
- Rods
- Transmittance
- Vitreous humor

12.8 Problems for Chap. 12

1. (a) Roughly what is the value of the effective focal length of the eye?
   (b) What is the light-sensitive layer in the eye called?
   (c) What is myopia and how is it corrected?
   (d) What is hyperopia and how is it corrected?
   (e) What is the area of the retina called that has the greatest visual acuity?
   (f) What function does the crystalline lens of the eye have?

(g) For what object distances is the eye focused when relaxed – *near* or *far*?
(h) What is the **near point**?
(i) Why does the eye have a blind spot?
(j) What are rods and cones? Which are numerous?
(k) Which type of receptor cell does the fovea contain?
(l) Which is more sensitive to light, photopic or scotopic vision? By how much more?
(m) List two methods of binocular depth perception.

2. The greatest amount of refraction in the lens system of the eye occurs when the light

   (a) Passes through the middle of the lens
   (b) Passes from the vitreous humor to the aqueous humor
   (c) Passes from air into the vitreous humor
   (d) Enters the cornea

3. The old saying "at night, all cats appear gray" has the following scientific basis:

   (a) At low light levels, all cones are stimulated, so every object appears white or off-white.
   (b) At low light levels, the cones are not stimulated, and rods cannot distinguish colors.
   (c) At night, only the fovea reacts to light.
   (d) In low light levels, the lens transmits only blue light, which we interpret as gray.
   (e) At night, the optic nerve only reacts to light and dark.

4. Can an eye be both near-sighted and far-sighted? Explain.
5. Explain why in *dim light* objects appear more distinct if they lie off to the side of the field of vision.
6. Why do you suppose the eye's lens must have a *bulging much greater* than it would need if it were suspended in air, to provide its focal length?
7. Suppose that the image distance from the effective lens to the retina is 2.25 cm.

   (a) Find the focal length of the eye lens when viewing an extremely distant object.
   (b) Suppose that the eye now views an object a distance 30 cm away from the eye. Do you expect the focal length to increase or decrease? Calculate the focal length.

8. Suppose that an eye has a range of object distances which can be brought into focus from 40 to 200 cm. Assuming that $d_{ie} = 24$ mm, **calculate** this eye's range of focal lengths.
9. Consider an object with a height of 2 m. Estimate the minimum distance it can have from the eye so that its image can be cast entirely on the fovea.

10. The author had his eyes checked on December 8, 2005. The prescription for his left eye read -0.25 for far vision and $-0.25 + 2.25 = 2.00$ for near vision. These numbers refer to what is known as the **diopter** value. The diopter value **D** is a measure of the strength of a lens. Mathematically

$$D \equiv \frac{1}{f}, \tag{12.7}$$

where f is the focal length *expressed in meters*. Thus, for far vision I need a diverging lens with a focal length

$$f_{\text{lens}} = \frac{1}{D} = \frac{1}{-0.25} = -4.00\,\text{m}. \tag{12.8}$$

Let us now again assume that the fixed image distance in my eye is 24 mm.

(a) Calculate the focal length f_{far} needed by my eye to bring into focus objects at infinity **without** any corrective help.

(b) The focal length with my corrective lens is determined by the maximum focal length of my eye max f_{eye} and f_{far} as follows. If we neglect the fact that the corrective lens is held outside the eyeball and assume as an approximation that it is coincident with the center of the eye lens, it can be shown that the effective focal length of the combination of the two lenses is obtained by adding the inverses of the respective focal lengths:

$$\frac{1}{f_{\text{far}}} = \frac{1}{f_{\text{lens}}} + \frac{1}{\max\ f_{\text{eye}}}. \tag{12.9}$$

Expressed in terms of diopters, we have direct addition:

$$D_{\text{far}} = D_{\text{lens}} + \min\ D_{\text{eye}}. \tag{12.10}$$

Calculate D_{far} and min D_{eye}.

(c) Calculate min f_{eye}.

(d) The maximum focal length of my eye used for near vision can be determined by the desired near point, which we choose here to be $d_{\text{np}} = 25\,\text{cm}$, and the focal length of the corrective lens. First calculate the needed focal length without correction:

$$\frac{1}{f_{\text{near}}} = \frac{1}{d_{\text{np}}} + \frac{1}{24\,\text{mm}}. \tag{12.11}$$

(e) Next calculate

$$D_{\text{near}} = D_{\text{lens}} + \max\ D_{\text{eye}}. \tag{12.12}$$

(f) Finally calculate min f_{eye}.

It is useful to mark off on a line the positions of the eye-lens, f_{far}, f_{near}, max f_{eye}, and min f_{eye}.

11. A tree of height 5 m is a distance 10 m away from you. How high is the image of the tree on your retina if your lens-to-retina distance is 24 mm?
12. Suppose we treat the human eye as having a uniform index of refraction of 1.37. What fraction of light intensity, incident directly at the eye, is *transmitted* into the eye from air?
13. The **effective focal length** of a particular person's eye is 22 mm when the lens is relaxed. The effective lens-to-retina distance is 24 mm.

    (a) Is the person near-sighted or far-sighted?
    (b) To bring distant objects into focus, a diverging lens is used.

    Assuming that the **corrective lens** is *adjacent* to the eye: Determine the focal length of the lens so that it will produce a **virtual image** of an infinitely distant object, such that the virtual image will serve as an object which the eye will focus on the retina in the relaxed-lens state.
14. It has been reported that the density of rods on the retina can be as high as 160,000 per mm^2. On this basis, assuming that the rods are touching each other, **estimate the diameter** of a single rod. To do so, imagine the rods distributed as squares on a checkerboard, each square having a side d, which will represent the diameter. Then write down an expression relating d to the density. How many squares, each of area d^2 would there be in an area of 1 mm^2?

Chapter 13
Characterizing Light Sources Color Filters and Pigments

13.1 Characterization of a Light Beam

If we are to be able to understand the way the eye transmits nerve impulses, we need to be able to characterize accurately the light that is incident upon the eye. Obviously, the light generally casts an image with great variation in detail, with respect to both color and intensity. In this chapter, we will restrict ourselves to an image that consists of a field of **uniform color and intensity**. Such an image can be produced by having a light beam be cast upon and then reflected by a white screen. Therefore, we will focus our attention upon the characterization of a light beam.[1]

A complete characterization of a light beam and hence its source would include a specification of the wave pattern, that is, the magnitude of the electric field, as it varies in space and time, as well as the state of polarization. In this chapter, we will ignore the state of polarization of the beam. The reason is that our goal is to relate the physical characteristics of a beam with color perception. And, polarization plays little role in this regard.

For the purposes of characterizing the color and brightness alone, a spectral analysis is sufficient. That is, all we have to know is the intensity of all the Fourier (monochromatic) components in the visible region of the electromagnetic (EM) spectrum. A specification of the intensity with respect to frequency is called the **spectral intensity**, and will be symbolized by $I(\lambda)$ – A Fourier analysis of the wave pattern determines the spectral intensity. In practice, the spectral intensity of a light beam can be determined by a prism or by a diffraction grating. In the figure below, a light beam is shown passing through a diffraction grating. Equation (13.2) provides us with a relationship between the angle θ and the wavelength. Therefore, the spectral intensity of the light is measured by a light meter set at various angles θ.

[1] See S. J. Williamson and H. Z. Cummins, *Light and Color*, (John Wiley and Sons, New York, 1983), for a more detailed treatment of this subject.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_13,

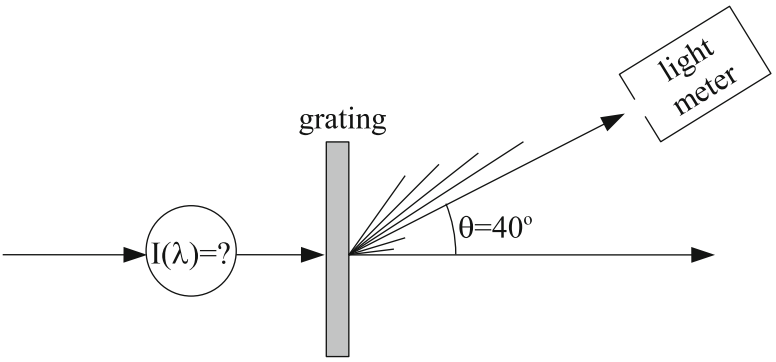


Fig. 13.1 Diffraction by a grating

In order to simplify our discussion, we will be concrete by specifying the line spacing of the grating as **10,000 lines per centimeter** (see Chap. 8). The line spacing d is equal to $(10{,}000)^{-1}$ cm $= 1{,}000$ nm. The first order diffraction pattern has the angle θ related to the wavelength via the equation:

$$\sin\theta = \frac{\lambda}{d} = \frac{\lambda\ (\text{nm})}{1{,}000}. \tag{13.1}$$

Alternatively, we could write

$$\lambda\ (\text{nm}) = 1{,}000\ \sin\theta. \tag{13.2}$$

The symbol λ (nm) means that the use of (13.1) or (13.2) requires that the wavelength be expressed in nanometers (nm).

We will restrict the wavelength λ to the range 400–700 nm (the approximate visible range). Then $\sin\theta$ ranges from 0.40 to 0.70 and, correspondingly, the angle θ ranges from 23.6° to 44.4°.

In Fig. 13.1, the light meter is set at 40°. It therefore measures the intensity of the spectral component having a wavelength of $\lambda = (\sin 40°)(1{,}000) = 643$ nm.

Examples of some spectral intensities are exhibited below:

1. **Monochromatic light**, with a wavelength of 580 nm is represented by a spike. See Fig. 13.2. Absolutely monochromatic light is a nonexistent ideal. If a quantum system were to have a transition from one state to another that leads to absolutely monochromatic light and if it were initially in the lower of the two states, the quantum system would not absorb any incident light that is not absolutely monochromatic light of precisely the same wavelength.
2. Neon laser light; see Fig. 13.3
   The spectral intensity of the laser light has a peak at 632.8 nm. The **bandwidth** $\Delta\lambda$, shown in the figure, gives us a measure of how monochromatic the light is. It is usually defined as the difference between the two wavelengths for which

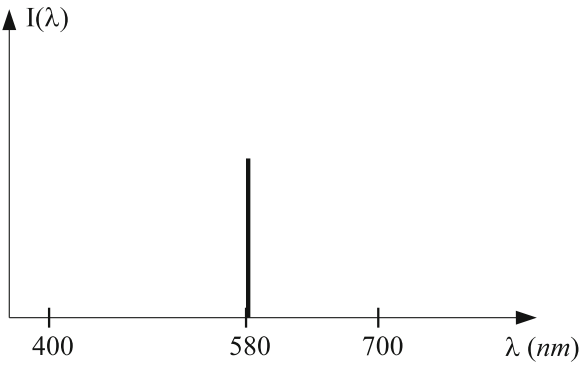


Fig. 13.2 Absolute monochromatic *yellow*

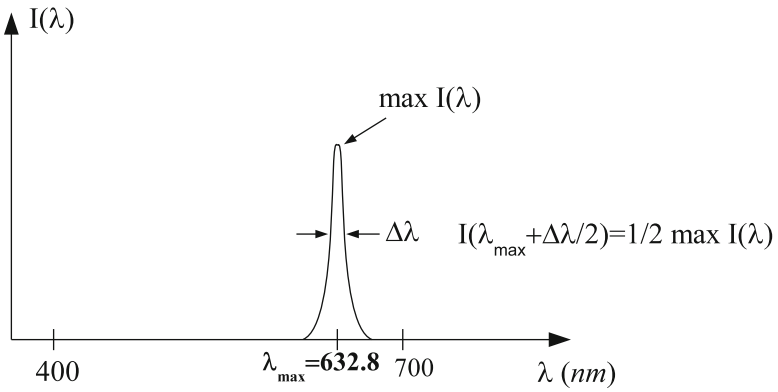


Fig. 13.3 Quasi-monochromatic *red*

$I(\lambda)$ is equal to one-half the maximum value of $I(\lambda)$. Laser light is very highly monochromatic in which $\Delta\lambda$ is a minuscule fraction of the wavelength $\lambda_{\max}$ at which $I(\lambda)$ is a maximum. For He–Ne laser light, $\Delta\lambda = 0.002\,\text{nm}$. Clearly $\Delta\lambda$ is highly exaggerated in the above figure.

The color sensations produced by monochromatic light spans the huge range of colors of the rainbow. These color sensations are called **spectral colors**. For simplicity, the spectral colors are grouped according to certain ranges of wavelengths. For example, the term **spectral red** refers to a monochromatic light that is red in color. The groups are listed in the table below:

λ (nm)	Spectral color
400–420	Violet
420–455	Indigo
455–490	Blue
490–575	Green
575–585	Yellow
585–650	Orange
650–720	Red

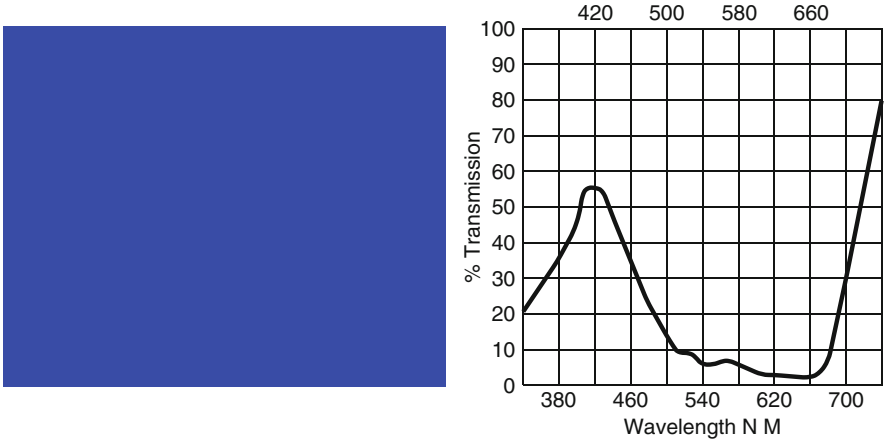


Fig. 13.4 Double Blue slide and corresponding spectral intensity (source: Courtesy of Rosco, Inc, Stamford, CT)

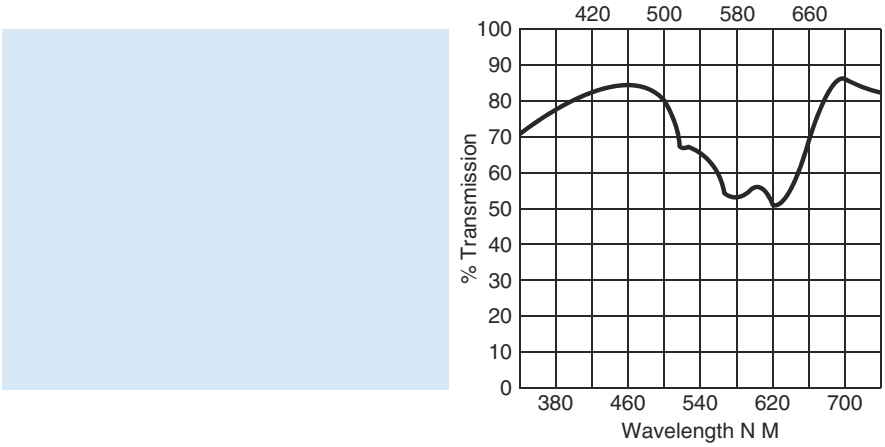


Fig. 13.5 Mist Blue slide and corresponding spectral intensity (source: Courtesy of Rosco, Inc, Stamford, CT)

According to this table, neon laser light would be labeled **orange**. Boundaries between color-labeled regions are not unanimously agreed upon!

It is important to note that certain color sensations cannot be produced by monochromatic light. They include: *white*, *magenta*, *cyan (turquoise)*, *brown*, and *gray*. These will be further discussed in Chap. 14.

3. In Fig. 13.4 is an example of a *broad* spectral intensity that produces a distinct bluish color sensation. Note that quite a bit of monochromatic light in the red region is present.

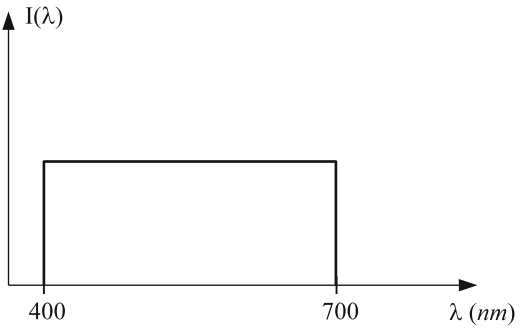


Fig. 13.6 *White* light spectrum

4. The pale blue color in Fig. 13.5 requires a much broader spectral intensity than the previous one in Fig. 13.4.
5. A spectral intensity that is a constant, that is, the same, for all visible wavelengths is referred to as **pure white** and is shown in Fig. 13.6. Another term for this spectrum is **equal energy spectrum**, for obvious reasons. Such a spectrum will appear white.

   As we move from a spectral blue to a broad spectrum blue to a pale blue, and on to pure white, the color sensation becomes less distinct. We say that the color becomes less **saturated**. A spectral color has maximum saturation. A pale blue is a low saturated blue, with "blue" referred to as the **hue**. It can be simply produced by adding white light to a saturated blue light. Similarly, **pink** is low saturated red.

   So far, we have characterized our psychological perception of a light source according to *hue* and *saturation*. Hue has so far been associated with a peak frequency in the spectrum. The **third** characteristic of a light source is its **brightness**, which is a **subjective** characteristic related to the overall intensity of the light source. The range of light intensities that can be seen without pain is from $\sim 10^{-10}\,\mathrm{W/m^2}$ to $\sim 100\,\mathrm{W/m^2}$ and therefore spans *12 orders of magnitude*, as in the case of sound.

   The following table summarizes the *psychological* perception characteristics with the *physical* characteristics of light and sound:

	Psychological	Physical
Light	Hue	Central frequency
	Saturation	Band width or white admixture
	Brightness	Intensity
Sound	Pitch	Fundamental frequency
	Timbre	Admixture of overtones
	Loudness	Intensity

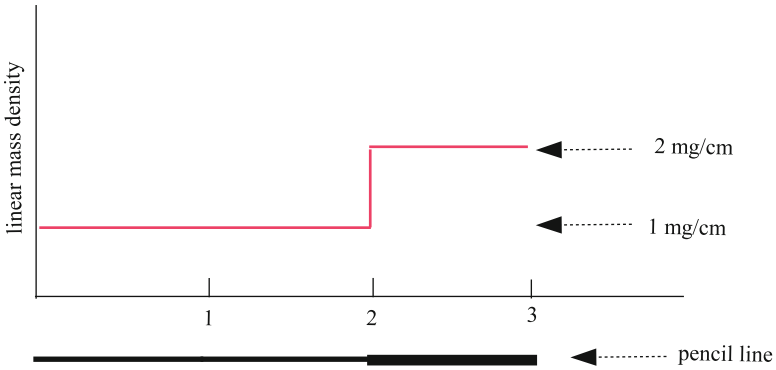


Fig. 13.7 Linear mass density of a *pencil line*

13.1.1 Spectral Intensity vs. Intensity

Question: What is the significance of the value of the **spectral intensity** at a specific wavelength? Is it the value of the **intensity** at that wavelength?

Let us consider a line on a paper, 3 cm long, made by a pencil with variable thickness. See Fig. 13.7.

There is no mass at a *given point* along the line. There is only mass for a line segment. We see a graph that plots the **linear mass density** of the material that constitutes the line. The total mass is 4 mg. There are 2 mg between the origin and the point 2 cm. This value is obtained by multiplying the linear mass density, 1 mg/cm by the length of 2 cm. Similarly, we obtain a mass of 2 mg for the segment that is 1 cm long with a linear mass density of 2 mg/cm. The reader will note that generally, the mass can be calculated by determining the area under the graph. Each unit area is a milligram. If we have a continuously changing thickness, the total mass will be the area under the curve of mass density vs. position.

Similarly, there is no intensity associated with the value of the spectral intensity at a given point. The spectral intensity is an **intensity density** being equal to the intensity per interval of wavelength. To be a bit more explicit – a range of wavelengths from λ_1 to λ_2 has a total intensity that is equal to the area under the curve of the spectral intensity vs. wavelength between these two wavelengths.

Sample Problem 13-1

In Fig. 13.8, we see a plot of a particular spectral intensity. Note that the unit of spectral intensity in the plot is 1 W/m^2 per nm. The wavelength difference between ticks on the plot is 100 nm. What is the total intensity?

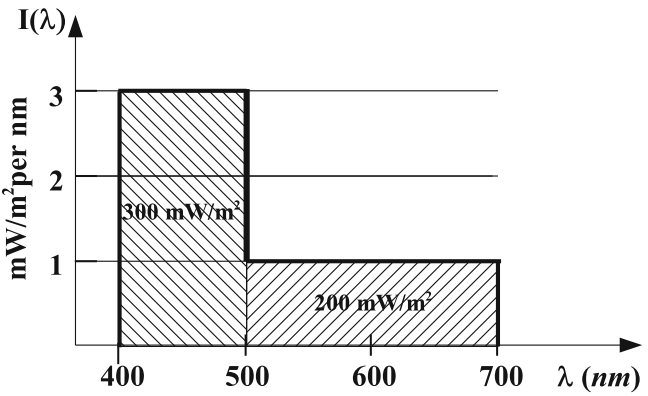


Fig. 13.8 Spectral intensity

Solution

We see two cross-hatched rectangles. From 400 to 500 nm, we have a spectral intensity of 1 mW/m^2 per nm, so that the area of this rectangle is 300 units, corresponding to 300 mW/m^2 = 0.3 W/m^2. With the same analysis, we obtain an intensity contribution of 0.2 W/m^2 from the 500 to 700 nm. The total intensity is therefore 0.5 W/m^2.

13.2 Color Filters

A color filter is transparent to light, with a fraction of the intensity transmitted that is dependent upon the wavelength. Color filters function by a process of **selective absorption** – that is, the more a spectral component is absorbed, the less is transmitted.

If a light beam of spectral intensity $I(\lambda)$ is passed through a color filter, the outgoing beam has a spectral intensity $I(\lambda)'$ that is given by the product of the **transmittance** $T(\lambda)$ and the incoming spectral intensity:

$$I'(\lambda) = T(\lambda)I(\lambda). \tag{13.3}$$

The transmittance can be expressed as a fraction (less than one) or a percentage. The process is exhibited in Fig. 13.9.

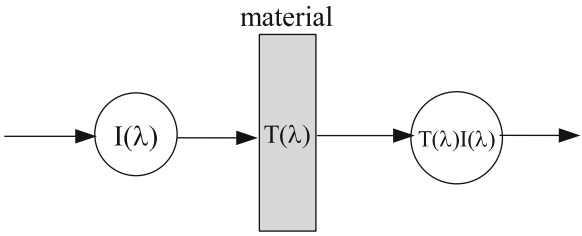


Fig. 13.9 Transmittance of a material

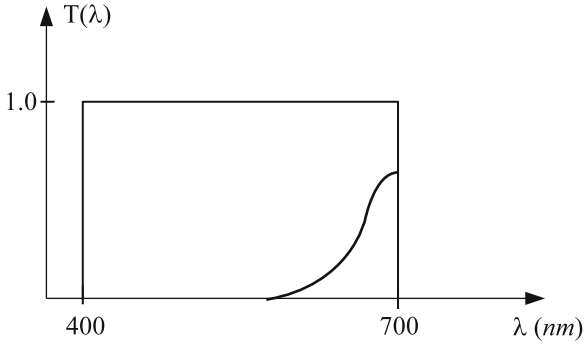


Fig. 13.10 Transmittance of a *red* filter

Note

If the incoming intensity is pure white, the outgoing beam will have a spectral intensity that is proportional to the transmittance. Thus, to produce the blue and pale blue colors in the previous figures requires filters with transmittances that are proportional to the respective spectral intensities.

An example of the transmittance of a red filter is shown in Fig. 13.10.

Homework: Sketch $T(\lambda)$ for a green filter.

Note

We do not have filters that transmit highly monochromatic light. To obtain highly monochromatic light from a white source (or non-chromatic source), a prism or a diffraction grating is used. In this capacity, these devices are referred to as **monochromators**.

Note

One might ask, "What happens to the incident light that is not transmitted through a filter?" The answer is that the remaining part of the light is represented by reflected light as well as light that is absorbed by the material out of which the filter is made.

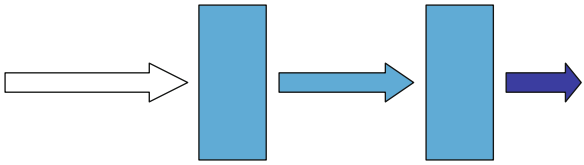


Fig. 13.11 Two identical filters

Note

Question: Does the value of the transmittance at a specific value of the wavelength have a direct significance experimentally? The answer is affirmative: The transmittance at a given wavelength is the fraction of incident **monochromatic light at that wavelength** that is transmitted.

13.2.1 Stacking Filters (Filters in Series)

When two filters are laid one on top of the other, in series, we say that they are **stacked**. The transmittance of a **series of two filters** is obtained by *multiplying* the two respective transmittances.

Thus,

$$T(\lambda) = T_1(\lambda)T_2(\lambda). \tag{13.4}$$

Note that the order in which the filters are stacked is not relevant!

To illustrate an application of this formula, we will examine a process whereby an incoming beam of white light has its spectral intensity increasingly narrowed (hence made increasingly saturated) by a series of *two identical* filters, so that $T_1(\lambda) = T_2(\lambda)$. See Fig. 13.11.

The length and color of the arrows in the figure are meant to show that on passing through a filter, the outgoing light has a reduced intensity and an increased saturation. This effect can be achieved by using one filter as follows: We simply have the light that passes through the single filter reflect off a mirror or a diffuse reflecting white surface and then pass through the filter a second time. (The term **white** here means that all frequencies are reflected to an equal extent.) See Fig. 13.12.

Let us consider a simple **numerical** example to illustrate the increase in saturation by a series of two identical filters. Suppose that $T_1(\lambda)$ and $T_2(\lambda)$ are given by the following:

$$T_1(\lambda) = T_2(\lambda) = T(\lambda) = \begin{cases} 0.8 \text{ for } 500\,\text{nm} < \lambda < 600\,\text{nm} \\ 0.3 \text{ otherwise} \end{cases}.$$

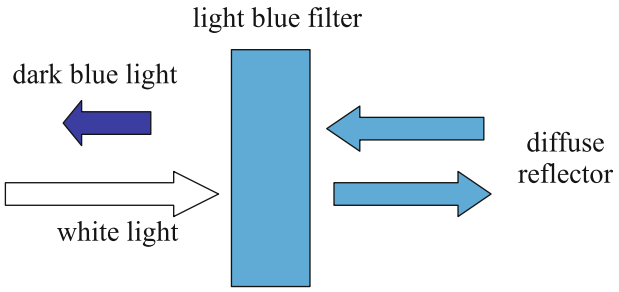


Fig. 13.12 Light through a filter, reflected back through the filter

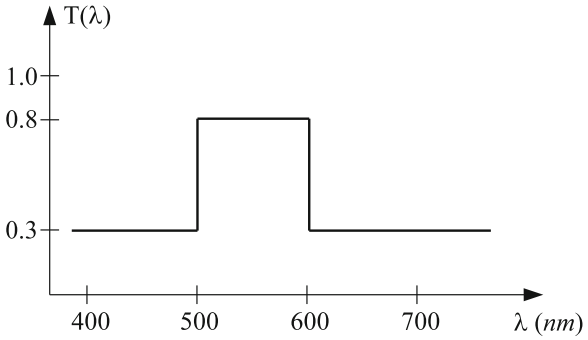


Fig. 13.13 Graphical representation of the transmittance in the above example

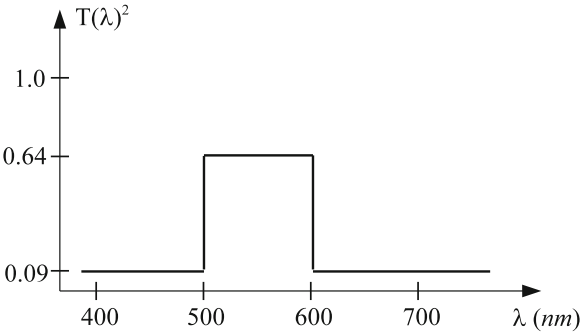


Fig. 13.14 Net transmittance for the sequence of two identical filters

This transmittance is shown in Fig. 13.13.

$$T(\lambda)^2 = \begin{cases} 0.64 \text{ for } 500\,\text{nm} < \lambda < 600\,\text{nm} \\ 0.09 \text{ otherwise} \end{cases}.$$

The resulting net transmittance is shown in Fig. 13.14.

We see that the resulting transmittance is more sharply peaked. A series of identical filters or a single thick filter can be used to make a white beam more monochromatic. However, note that *the price paid is reduced intensity*.

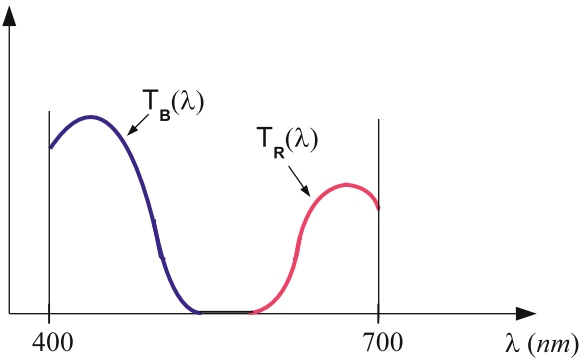


Fig. 13.15 Two filters together filtering out all light

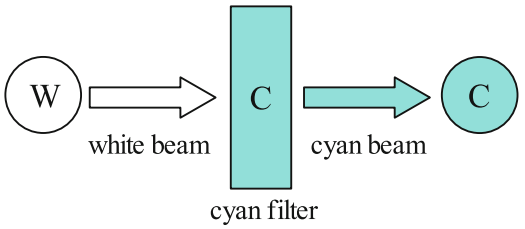


Fig. 13.16 Cyan filter

Example Two: We consider a series of two filters, one a red filter and the other a blue one. We see that wherever $T_{\lambda_R}(\lambda)$ is nonzero, $T_{\lambda_B}(\lambda)$ is zero, and conversely. Thus, the net transmittance, $T(\lambda) = T_{\lambda_R}(\lambda)T_{\lambda_B}(\lambda)$ vanishes for all visible wavelengths. See Fig. 13.15.

That is, the two filters remove all light!

To further simplify our discussion, we will use the following symbols to refer to various general colors:

W: white	**G**: green	**M**: magenta	**BN**: brown
R: red	**B**: blue	**C**: cyan	**BK**: black
O: orange	**V**: violet	**P**: purple	**Y**: yellow

In order to indicate the effect that a filter has on a beam of light, we will use the symbols in Fig. 13.16. Their significance should be self-explanatory.

We can use a prism to determine the transmittance of a filter as shown below for a *typical yellow filter*. See Fig. 13.17. First, let us recall that yellow is a spectral color. As such, we might expect to see one outgoing yellow beam. Instead, we would see two outgoing beams, green and red.

This analysis reveals that the filter does not transmit spectral yellow. Instead, it transmits the two colors, red and green. We will label such a filter with an asterisked Y: **Y***, to distinguish it from a spectral yellow filter. From the above, we learn that the eye sees yellow when a mixture of red and green light is incident on the eye. We will explain this phenomenon in Chap. 14, where we will discuss a theory of Color Vision.

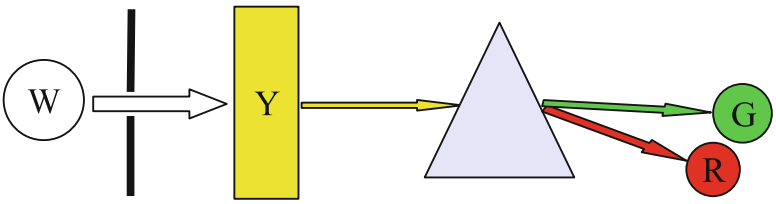


Fig. 13.17 *Yellow* beam and a prism

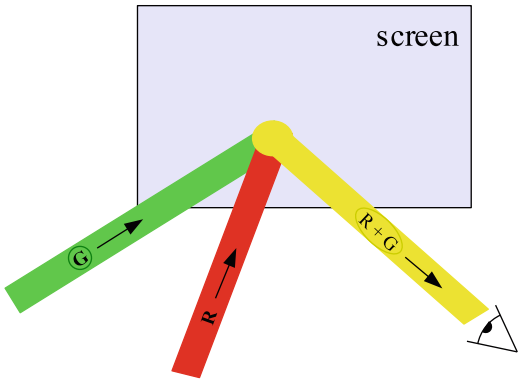


Fig. 13.18 Addition of *red* and *green* to give *yellow*: $R \bigoplus G \equiv Y$

The most significant conclusion of these observations, a conclusion that has widespread validity, is that there is

NO one-to-one correspondence between the multitude of spectral intensities $I(\lambda)$ and multitude of color sensations.

In fact, every color will and can be produced by an infinite number of spectral intensities.

Now suppose that we cast a circular beam of light onto a white screen that scatters light diffusely. Such an image is called a **color patch**. If we cast two color beams onto the screen, and have the patches overlap, our eyes will see the light from both beams that has scattered diffusely from the screen into our eyes. Our eyes will then receive the *sum* of the two light intensities. In particular, overlapping patches from a red filter and a green filter will look like a *yellow patch*! See Fig. 13.18.

The sum of red light and green light appears yellow as with the production of yellow with the yellow filter Y*. The process reflects a synthesis, from a mixture of spectral green and spectral red, of a color that is *essentially* indistinguishable[2] from the sensation produced by a spectral yellow. This synthesis is the reverse of the analysis carried out by the prism on the light transmitted by the yellow filter.

[2]The word *essentially* is inserted here because light from a single monochromatic source will always be more saturated than light from a mixture of monochromatic sources, as we will learn in Chap. 14.

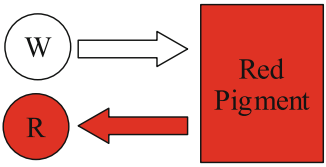


Fig. 13.19 Action of *red* pigment

Most people are surprised at the above observation of mixing red and green to produce yellow. They expect the same result as is produced by mixing red and green **paints**. In that case, **red** added to **green** equals a brown. In the next section, we will discuss the behavior of **pigments**, which are the basic ingredients of paints.

13.3 Pigments

What determines the color of opaque objects? Clearly, the transmittance of a filter is mirrored by the ability of the filter to absorb selectively. Similarly, an opaque object absorbs selectively and has a corresponding wavelength-dependent *reflectance*. **Pigments** are materials that are extremely selectively absorptive. They can color a material when present in very dilute concentrations. They are most commonly used in paints. Here are examples with a simplified account:

1. RED pigment absorbs all but red and reflects red. See Fig. 13.19.
2. GREEN pigment absorbs all but green and reflects green.
3. BLUE pigment absorbs all but blue and reflects blue.
4. Y* pigment absorbs B and reflects R and G.
5. Y* pigment mixed with B pigment absorbs B, R, and G – so absorbs all!

Thus, Y* pigment mixed with blue pigment should appear black according to our simple description. Most often, such a mixture appears green.

Paints behave like the pigments described above. How they are prepared so as to have the desired absorption–reflection characteristics is fascinating.[3]

13.4 Summary Comments on Filters and Pigments

1. Both filters and pigments function by a process of *selective absorption*.
2. With a filter, we observe *transmitted light*.
3. With pigment, we observe *reflected light*.

[3]The following are suggested for further reading: *Light and Color in Nature and Art*, by S. Williamson and H. Cummins, (J. Wiley and Sons, N.Y., 1983), *Light and Color*, by R. D. Overheim and D. L. Wagner (J. Wiley and Sons, N.Y., 1982), and *Seeing the Light*, by D. Falk, D. Brill, and D. Stork (Harper and Row, N.Y., 1986).

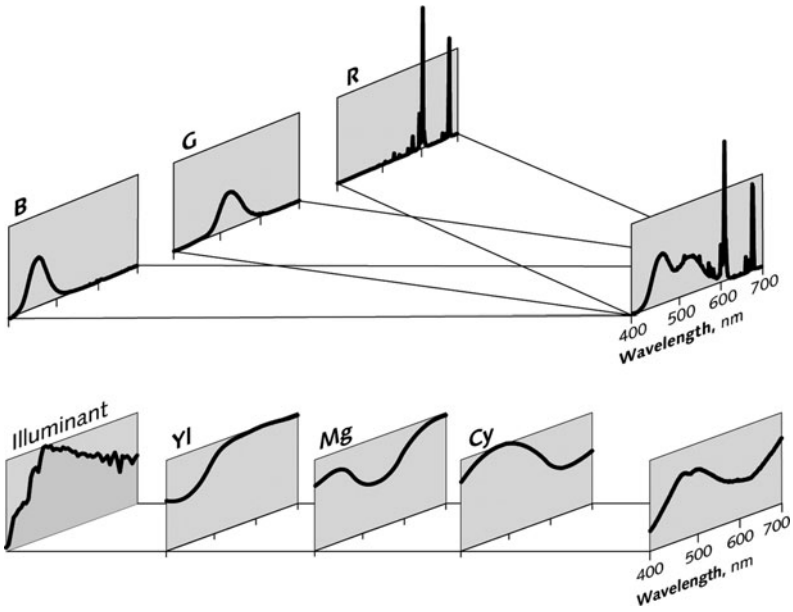


Fig. 13.20 Comparison of addition and subtraction of spectral intensities (source: Charles Poynton and Garrett Johnson, *Color science and color appearance models for CG, HDTV, and D-Cinema*; ACM Digital Library, © 2004 http://dl.acm.org/citation.cfm?id=1103903, reprinted with permission)

4. When the color patches of light beams are *overlapped*, spectral intensities are added and we have **additive mixing**.
5. When color filters are *stacked* or pigments are *mixed*, spectral absorptions are added and we have **subtractive mixing**.

Figure 13.20 illustrates very nicely the difference between the two processes, addition and subtraction. Results for both types of mixing are summarized in Fig. 13.21.

13.5 Terms

- Additive mixing
- Bandwidth of a spectrum
- Color filter
- Color
- Additive mixing
- Bandwidth of a spectrum
- Color patch
- Equal energy spectrum
- Hue
- Monochrometer
- Monochromatic light
- Pigment
- Saturation
- Spectral color
- Spectral intensity
- Subtractive mixing
- Transmittance

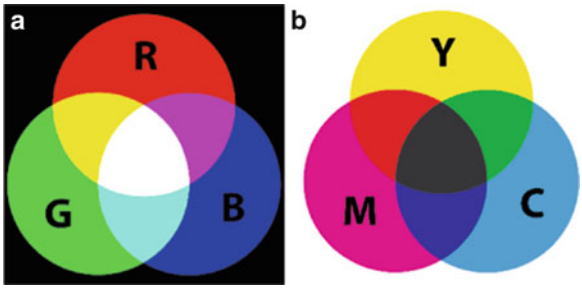


Fig. 13.21 Mixing of colors. (**a**) Additive mixing. (source: http://upload.wikimedia.org/wikipedia/commons/thumb/c/c2/AdditiveColor.svg/1000px-AdditiveColor.svg.png); (**b**) Subtractive mixing. (source: http://upload.wikimedia.org/wikipedia/commons/thumb/1/19/SubtractiveColor.svg/1000px-SubtractiveColor.svg.png)

13.6 Important Equations

Transmitted intensity:

$$I'(\lambda) = T(\lambda)I(\lambda). \tag{13.5}$$

Transmittance for two stacked filters:

$$T(\lambda) = T_1(\lambda)T_2(\lambda). \tag{13.6}$$

13.7 Problems for Chap. 13

1. What are the physical characteristics of a light source that are, respectively, related to the following:
   (a) Hue
   (b) Saturation
   (c) Brightness
2. When white light is passed through two identical stacked filters having a strong degree of monochromaticity, which of the three attributes of the outgoing beam *change considerably in comparison* with passage through *one* of the filters?
   (a) Hue
   (b) Saturation
   (c) Brightness
3. In Fig. 13.22 are sketched the transmittances of two filters, $T_1(\lambda)$ and $T_2(\lambda)$, respectively.
   (a) What colors do these filters individually produce?

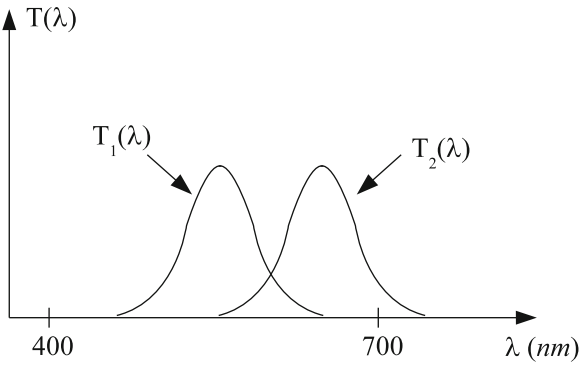


Fig. 13.22 Two-maxima spectrum

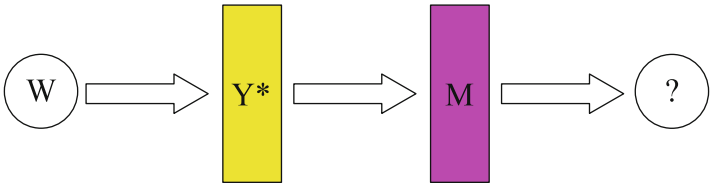


Fig. 13.23 *White* light enters a *yellow* and *magenta* filter

(b) If the filters are stacked, sketch the resulting transmittance. What is the corresponding color?
(c) Suppose a white beam is passed through each of the filters and the beams are projected onto a screen so as to produce two overlapping patches. What is the color of the patch? Sketch the spectral intensity of the reflected light.

4. (a) In mixing blue and yellow (Y*) lights, one can produce _____ light.
   (b) In mixing blue and yellow (Y*) pigments, one can produce _____ pigment.
   (c) Explain why your answers to parts (a) and (b) are different.
5. Complete Fig. 13.23.

Chapter 14
Theory of Color Vision[1]

In Chap. 13, we saw that one can characterize a light source in terms of its spectral intensity. We now turn to the question as to the relationship between the **objective** characteristics of a light and the **subjective** perception of the light. We have already identified three attributes of one's visual perception – **hue**, **saturation**, and **brightness**. The first two are the attributes that, together, are referred to as the **color**. Thus, our goal is to establish the relationship between the spectral intensity and these three attributes. We will first summarize this relationship as it was determined through years of testing many individuals. Then we will discuss how this relationship can be understood in terms of the biophysical behavior of the visual apparatus, the rods of the retina.

Before we begin, some important comments are in order. What does it mean to say that I see the color **blue** and that you agree with me? Do we experience the same sensations? Does the color blue look the same to you as it does to me? Some thought reveals that no one can tell what another person's actual sensation and experience is in connection with the color blue. We could say the same thing about any sensation or feeling that we have. For example, I know what I myself experience when I say that I feel cold or I feel sad. But I cannot tell what any other person's experience is like when they refer to these experiences with the same words. If someone says that they feel cold, I might observe behaviors that are commonly displayed by people who feel cold – shivering, for example. Observing such behaviors support and/or give me confirmation that a person is experiencing a cold feeling and lead me to assume that their experience is the same as mine.

If I cannot tell what another person's experience is like, based on how they describe in standard verbal terms, what can I be essentially sure of? One possibility is the following: people have a multitude of experiences. These experiences are **mapped** onto a language of words. *Ideally*, this mapping is in **one-to-one**

[1] In addition to the elementary book,*Light and Color* by **Overheim and Wagner** (op, cit. in Chap. 13), the reader is referred to the advanced texts: **T. N. Cornsweet**'s *Visual Perception* (Academic Press, N.Y., 1970) and **Y. Le Grand's** *Light, Color, and Vision* (Dover, N.Y., 1957).

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3_14,

correspondence, as with the object and image of a lens. In this case, a one-to-one verbal mapping has the property that

no given experience can be expressed using more than one specific verbal description, AND, no specific verbal description can be used to describe more than one given experience.

As far as color perception is concerned, we can enumerate a multitudinous set of color sensations that have been shown to be in **one-to-one correspondence** with a corresponding set of physical characteristics of light sources. Thus, all people with normal vision and a rich exposure to the visual world will use the word **red** to refer to a certain set of visual sensations. Furthermore, they can distinguish the sensation of "red" from other colors like "green" or "blue." More generally, they have established, from experiences in childhood and onward, a vocabulary to describe their sensations of color in response to various inputs of light. This correspondence is shared by all people with normal vision. On the other hand, there are people with various kinds of **color blindness**. Most common is the inability to distinguish red from green. That is, two light sources, which produce the sensations of red and green in people with normal vision, appear to have the same color. Color blind people have a different mapping than people with normal vision. Our first goal is to describe the mapping for people with normal vision. We will then discuss some of the characteristics of color blindness and give a physiological explanation for the phenomenon.

The ultimate purpose of this chapter is to show how any color can be specified in terms of two numbers that correspond to a choice of three spectral primaries. Such a reliable specification of color is essential for the communication of color among interested individuals and for the reproduction of color by mixing various color sources. Examples are paints, color monitors, color printers, and fabrics. In fact, it is possible for someone to "discover" a color among the infinite number of possible colors, a color that happens to be extremely pleasing or exciting to most people.[2] Such a color can then be patented!

14.1 A Simplified Version of the Three-Primary Theory

The observation that any color might be produced by adding together **three primary colors** is well known.[3] This idea is the basis for color television and was used by the **pointillist** painters, such as, Seurat, to produce a rich color painting by using

[2]It should be recognized that any particular individual is limited in their ability to discriminate one color from another. There are **Just Noticeable Differences in Color** in analogy with **Just Noticeable Differences in Frequency**. Therefore, a particular individual can discriminate among only a finite number of distinct colors.

[3]We will discuss the modern developments in the field of color. The history of the science of color vision started with Isaac Newton, who in the 1600s proposed that there are seven primaries that can be mixed in appropriate proportions to produce any color sensation. The basis of the proposal was Newton's studies of the decomposition of white light by a prism into its rainbow of

dots of three colors alone. We will begin this chapter by discussing a very simplified version of the three-primary theory of color vision. This version recognizes only six hues. Later, we will refine it with the **chromaticity diagram** , which reflects much more accurately the true nature of color vision, treating the full range of hues and degrees of saturation. It turns out that the ideal three-primary theory is quite good in reflecting human color vision. Finally, we will see how the three-primary theory is connected with the existence of three types of cones, the **red cone** , **green cone**, and **blue cone**. We will also see how these three cones can provide us with the three basic attributes of our visual sense: *hue*, *saturation*, and *brightness*. Recently, in 1986, genes for these three cones were identified at Stanford University by Jeremy Nathans, thus establishing a physiological basis for the primary color theory.

According to the Three-Primary Theory of Color Perception, there exists a set of three spectral colors which have the following property:

Any visual sensation can be produced
by an appropriate mixture of these sources.

We will henceforth label these spectral sources as R, G, and B. We must keep in mind that the *choice of these primaries is not unique*, and that the names and symbols merely indicate that the choices of primaries that work well tend to be in the respective regions of color, that is, red, green, and blue.

To introduce the ideas of the Three-Primary theory, we will neglect saturation and consider only six hues: R, G, B, Y, C, and M. The three non-primary hues can be produced from mixtures of the primaries as follows:

Let the symbol $\oplus$ represent the addition of two light sources. Thus, for example,

$$\text{Yellow} = \text{Red} + \text{Green}$$
$$\text{Alternatively we will write } Y = R \oplus G$$
$$\text{Cyan} = \text{Blue} + \text{Green}$$
$$\text{Alternatively we will write } C = B \oplus G$$
$$\text{Magenta} = \text{Blue} + \text{Red} \quad \text{Alternatively we will write } M = B \oplus R$$
$$\text{White} = \text{Blue} + \text{Green} + \text{Red} \quad \text{Alternatively we will write } W = B \oplus G \oplus R$$

R, G, and B are referred to as **additive primaries**, in which they can produce the remaining basic hues, Y, C, and M, along with white (W).

Within the framework of this simplified set of hues,

- A C-filter filters out R
- A Y-filter filters out B
- A M-filter filters out G

---

colors. He identified the color of an object as an attribute of the response of the eye to various wavelengths of light that are reflected off the object as opposed to the idea that the color "resides" in the object itself. The proposal that there are three primaries is due to Thomas Young (1807). Many other scientists helped develop the basic principles of color mixing – in particular, James C. Maxwell, who provided a theoretical basis for electromagnetic waves, as discussed in Chap. 5. In 1860, Maxwell produced the first, albeit crude, set of **color-matching functions**, which will be discussed in detail in this chapter. For an excellent history of studies of color vision, see Deane Judd in the publication: NATIONAL BUREAU OF STANDARDS: VOL. 55, p. 1313, (1966).

Suppose that we pass white light through these filters, C, Y, and M. Furthermore, let the symbol $\ominus$ represent the removal of the component that follows the symbol from the component that precedes it. Then we can write:

$$C = W \ominus R$$
$$Y = W \ominus B$$
$$M = W \ominus G.$$

When stacked in series, C, Y, and M filters act as **subtractive primaries**, in which they can produce the remaining basic hues, R, G, and B, along with black (BK): Let us introduce the symbol $\odot$ to indicate that two filters are in series. Thus, $C \odot Y$ represents C and Y in series. Then $C \odot Y = W \odot R \odot B = G$.

Homework: Explain why $B \odot Y = BK$, $C \odot M = B$, $Y \odot M = R$, and $C \odot Y \odot M = BK$.

The three subtractive primaries form **complementary color pairs** with the additive primaries, in which the addition of a pair of such colors produces WHITE:

$$B \oplus Y = W$$
$$G \oplus M = W$$
$$R \oplus C = W$$

Next, we want to move on to consider the full set of colors perceived with normal vision. This set is infinite in number.

14.2 Exploration of Color Mixing with a Computer

It is quite an experience to see how a mixture of primaries can produce a vast set of colors. You can do so in a simple way by using Paintbrush, which is available with either a PC or a MAC. Typically this program comes with any purchased PC. In the case of a MAC, you can obtain a free download from the apple.com website.

Your first step is to gain access to a window that displays a box whose color is determined by a set of the three numbers corresponding to the admixture of the primaries of your monitor. I suggest that you choose the setting that provides 8-bit Truecolor. The number eight means that there are $2^8 = 256$ possible values for the intensity of each primary, the values being given by 0, 1, 2, 3,... 255. The total number of possibilities is then $256 \times 256 \times 256 = 16,777,216$. Here is how you can access the palette of colors: In the PC version, you should click on colors/edit colors. In the MAC version you should click on Tools/Font/Colors. I will refer to the MAC version in what follows. In Fig. 14.1, you can see two of the windows that should appear.

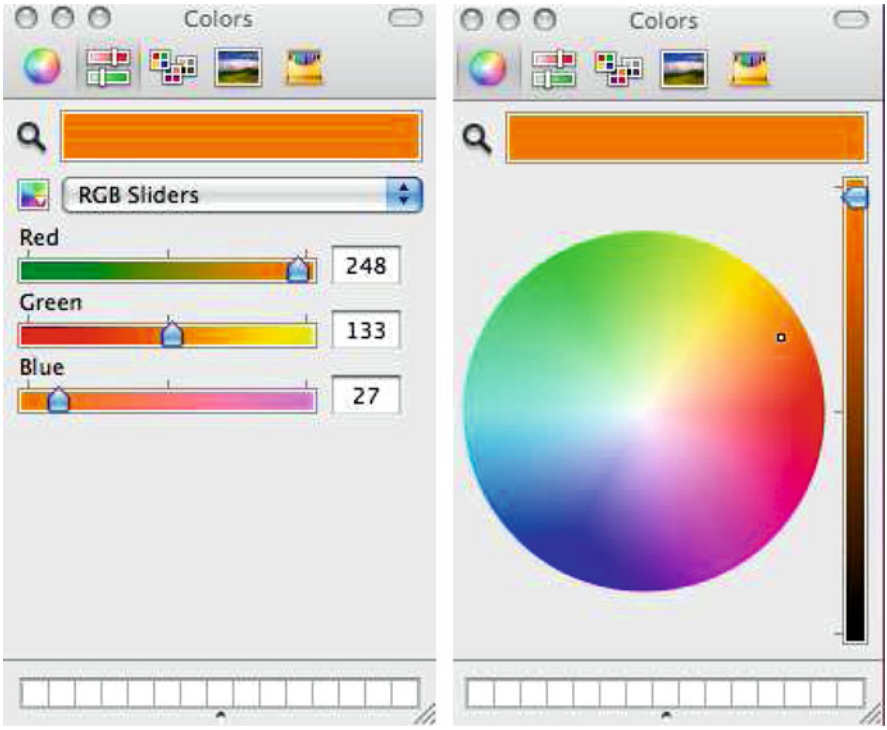


Fig. 14.1 Color mixing with paintbrush in a Mac computer – RGB sliders on the *Left*, color wheel and intensity slider on the *right*

At the very top you will see a small color wheel to the far left, followed by three color sliders.

Choose RGB sliders. Here you can choose the sets of three numbers that are associated with the color in a computer. I will use the bold letters **R**, **G**, and **B** here to refer to these values. *Be forewarned that the unbolded letters R, G, and B will refer to another set of numbers used later on in this chapter – the so-called tristimulus values. The two sets of three numbers are both used to specify the color but are not the same! Unfortunately, you will find both sets of symbols used interchangeably in the literature, so that you will have to make sure that you are certain about which set the symbols represent.* Another important fact to keep in mind is that the actual color you will see on a monitor or produced by a printer by a given set of monitor color coordinates **R**, **G**, and **B** will vary considerably. As a result, there is incredibly extensive literature on the problems of matching colors.

Experiment by moving the sliders and observing the changing values **RGB** and the corresponding color in the box at the top. Next, choose a set RGB and switch to the **color wheel**. In this window, you will see a circle on the left filled with colors. Note the small "pointer circle" someplace within the colored circle. It characterizes

Table 14.1 Preucil circle of colors

Ordering	Hue region	Formula
$R \geq G \geq B$	Red–yellow	$h_{\text{Preucil circle}} = 60° \cdot \frac{G-B}{R-B}$
$G > R \geq B$	Yellow–green	$h_{\text{Preucil circle}} = 60° \cdot \left(2 - \frac{R-B}{G-B}\right)$
$G \geq B > R$	Green–cyan	$h_{\text{Preucil circle}} = 60° \cdot \left(2 + \frac{B-R}{G-R}\right)$
$B > G > R$	Cyan–blue	$h_{\text{Preucil circle}} = 60° \cdot \left(4 - \frac{G-R}{B-R}\right)$
$B > R \geq G$	Blue–magenta	$h_{\text{Preucil circle}} = 60° \cdot \left(4 + \frac{R-G}{B-G}\right)$
$R \geq B > G$	Magenta–red	$h_{\text{Preucil circle}} = 60° \cdot \left(6 - \frac{B-G}{R-G}\right)$

the color. **Color** refers here to **hue** and **saturation**. The position of the pointer circle is characterized by its distance from the center of the color circle and by its angle with respect to a direction to the right. The two numbers are the **polar coordinates** of its position. The angle determines the hue, ranging from red to magenta, as the angle moves around from 0° to 360°. Thus, RED is at 0°, GREEN is at 120°, and BLUE is at 240°.

The angle is determined by the **RGB** values according to the formula introduced by Frank Preucil.[4] Table 14.1 provides the equations needed to calculate the angle. Next you can control the degree of **saturation**. By moving the pointer circle radially inward, you will see how the color becomes more pale: red becomes a pink, blue becomes pale blue, and so on. Lowering the saturation decreases your ability to distinguish among various hues. Ultimately, you will reach white at the center.

By moving the vertical slider at the right up and down, you can change the overall intensity while maintaining the color. Doing so will not change the hue. However, as you lower the intensity all the **RGB** coordinates will be reduced. The question is whether they are reduced in the same proportion. Check this out by switching back and forth between the circle window and the RGB slider window. A good place to start is with RGB set initially to full brightness white: **R=G=B=255**. As you reduce the intensity, you should find the **RGB** values reduced in proportion, so that they are always equal.

In Fig. 14.1, we see the two windows with an orange patch having **RGB** coordinates $\{248, 133, 27\}$, corresponding to 243 units of red primary, 128 units of green primary, and 2 units of blue primary. The actual colors seen will vary from monitor to monitor because of the varying mapping systems between computer color coordinates and intensities. If you now move the intensity slider downward toward decreasing intensity, you would find that the orange looks brown. Thus we learn that brown is not a specific color in the sense that we are using the term; rather, it is the color orange with low intensity. I am reminded of how a slice of bread that starts out with a very pale yellow color turns brown as it is toasted. Toasting increases the level of absorption of light so that the intensity of reflected light decreases. When the toast is fully burned, it absorbs a great fraction of the light and looks black.

[4]Frank Preucil, *Color Hue and Ink Transfer Their Relation to Perfect Reproduction*, TAGA Proceedings, p 102–110 (1953).

14.3 Introduction to the Chromaticity Diagram

Recall that our perception of a light source has three distinct characteristics, hue, saturation, and brightness. Hue and saturation together constitute what we will call the **color** of the light – technically referred to as the **chromaticity** of the light source. Thus,

$$\text{Chromaticity} = \text{Hue} + \text{Saturation}$$

In Sect. 14.1, we ignored brightness and saturation and recognized only six hues. Here we neglect brightness and consider the full range of chromaticities.

The results of tests on people with normal vision – those who are not classified as color blind – are summarized in what is known as a **Chromaticity Diagram**. It reflects the observation that two numbers alone are sufficient to specify the chromaticity of a light source:

For any spectral intensity $I(\lambda)$, one can calculate the values of these two numbers, referred to as **color coordinates**. They are given the symbols r and g.

We can easily appreciate why the chromaticity diagram can be of great interest and use to the student of color vision, purely for academic reasons as well as for medical reasons. However, the diagram is also invaluable to artists, to stage designers, as well as to designers of cloth for clothing. Imagine how useful it is to be able to refer to any color precisely by telephone or mail by merely specifying two numbers, say 0.2023, 0.4285.

Before we continue, it is important to keep in mind the following: Our discussion will assume normal color vision. Color blindness will be discussed later in this chapter in a dedicated section.

We cannot tell how the visual sensation of a particular chromaticity varies from person to person. We are taught from an early age to make associations between the set of words we use for various colors and the colors of objects that we observe.[5] As far as I know, children with normal vision do not have problems with learning how to differentiate colors. They seem to agree about what names to give to the colors they ascribe to objects.

However – this is the central fortunate observation – experiments have shown that when people with normal vision are asked to look at two color patches of light and asked whether they are the same, there is a strong agreement in their responses. If a pair of color sources appear different for one observer, they appear different for the other observer, and conversely. Because of this observation, in the late 1920s and early 1930s, pioneers in color specification carried out extensive experiments to quantify color perception. The two who stood out were W. David Wright and John Guild.[6]

---

[5] It should be clear from our study of color in Chap. 13 that physicists are far from inclined to get involved with the age old philosophical question as to whether a color resides in a colored object. Or, whether the color of an object is merely perceived.

[6] **Wright, William David** (1928). "A re-determination of the trichromatic coefficients of the spectral colours". Transactions of the Optical Society 30: 141–164. **Guild, John** (1931). "The

14.4 Metamers

It is found that an infinite number of different spectral intensities can produce the same color sensation. Any two spectral intensities producing the same color sensation are said to be **metamers**. Alternatively, two spectral intensities that have the same pair of color coordinates r and g are **metamers**. An example of such a pair of metamers is shown in Fig. 14.2 According to the figure on the right, we expect the common color to be bluish. One would hardly guess so from the figure on the left! [The figure was produced using the applet on the following wonderful website: http://www.cs.brown.edu/exploratories/freeSoftware/repository/edu/brown/cs/exploratories/applets/spectrum/metamers_guide.html].

As we will see, the fact that *two* numbers are sufficient to label a metamer is connected with the fact that **chromaticity** is specified by two characteristics – hue and saturation. In fact, they are specified by a pair of numbers. We have an infinite

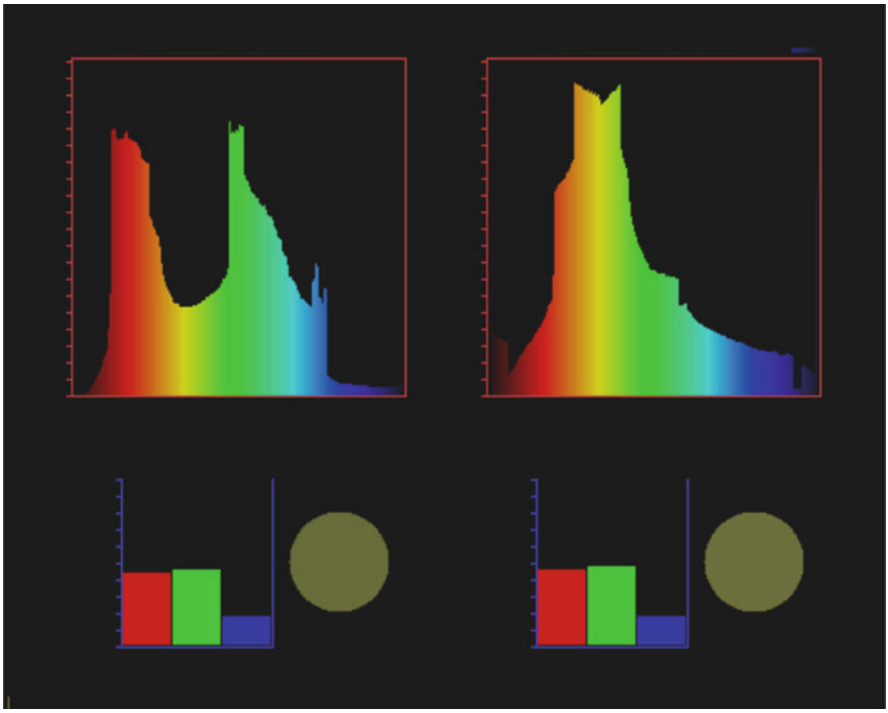


Fig. 14.2 Metamers: Two distinct spectra that produce the same color sensation

colorimetric properties of the spectrum". Philosophical Transactions of the Royal Society of London (Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, Vol. 230) A230: 149–187.

number of spectral intensities as well as an infinite number of pairs of numbers that specify the chromaticity. Mathematicians would say that the infinite set of all possible spectral intensities are **mapped** onto an infinite set of pairs of numbers.[7,8]

14.5 A Crude Chromaticity Diagram

The existence of metamers for the mathematician implies that the mapping between spectral intensity and perceived color is not one to one. It has been found that

1. Most, but **not all**, chromaticities can be produced by an appropriate mixture of three primary sources of light.
2. The choice of the three primary sources is arbitrary in which all choices leave out some set of unmatchable chromaticities.
3. However, the fraction of chromaticities which are matchable is maximized if the three primary sources are monochromatic.

Interestingly, we will see that

A set of monochromatic primaries cannot be mixed to match any other monochromatic color.

We will assume here that the primaries chosen are monochromatic.

Let us suppose that **three** primaries can be mixed to match the visual sensation of any given spectral intensity $I(\lambda)$. The mixture is specified by the intensities of each of the three primaries, namely, I_R, I_G, and I_B. We will now see how these three functions can determine the three parameters of color perception: hue, saturation, and brightness.

We can understand how this might be possible as follows:

– The *total* intensity, $I = I_R + I_G + I_B$, associated with $I(\lambda)$ is a measure of *brightness*.

[7] In the late 1800s, the mathematician Georg Cantor pioneered the study of levels of infinity and introduced a clearly defined method of comparing these levels. The lowest order of infinity is the number of integers, given the symbol $\aleph_0$. The next order of infinities is the set of real numbers, given the symbol $\mathcal{C}$. It can be shown that $\mathcal{C} = 10^{\aleph_0}$. As surprising as it may seem, Cantor was able to show using his method of comparing infinities that $\mathcal{C}$ is also the number of points in a finite area. This infinity is the infinite number of chromaticities. The number of spectral intensities is the number of ways you can draw a continuous graph along a finite axis. This number is an even higher infinity than $\mathcal{C}$ and can be shown to be equal to $\aleph_0^{\mathcal{C}}$. See the Wikipedia article (1-8-2011): http://en.wikipedia.org/wiki/Georg_Cantor.

[8] Here is a wonderful website that enables you to play around with pairs of spectral intensities, each independently and see their respective color patches. You can then produce metamers galore. http://www.cs.brown.edu/exploratories/freeSoftware/repository/edu/brown/cs/exploratories/applets/spectrum/metamers_guide.html.

– The three *fractions*

$$r' = \frac{I_R}{I_R + I_G + I_B}$$
$$g' = \frac{I_G}{I_R + I_G + I_B}$$
$$b' = \frac{I_B}{I_R + I_G + I_B}. \tag{14.1}$$

can characterize the *hue* and *saturation*, that is, the chromaticity.

We must have $r'+g'+b' = 1$. Therefore, only **two** of these three fractions are independent. For example, if you know r' and g', b' is determined to be (1-r'-b'). The convention is to specify r' and g'. They determine the two characteristics, hue and saturation.

Expressed succinctly, associated with any spectral intensity $I(\lambda)$ is a pair of numbers, the color coordinates (r', g'), which characterize fully the color sensation of the spectral intensity.

Suppose that the three intensities have the values of 1, 2, and 4, respectively, in some arbitrary units. The choice is irrelevant as far as chromaticity is concerned since only the ratios are relevant.[9] The total intensity is $1 + 2 + 4 = 7$ units and determines the brightness.

The three fractions are $r' = 1/7$, $g' = 2/7$, and $b' = 4/7$, which add up to unity. The first two fractions, here $1/7$ and $2/7$, are used to specify the chromaticity. The third is automatically determined: $b' = 1 - 1/7 - 2/7 = 4/7$.

Since two numbers specify a point in a plane, we can specify the chromaticity by a point in a plane. Furthermore, since the sum of the two numbers cannot exceed one, it can be shown that the point must lie within or on the boundary of the triangle shown in Fig. 14.3.

Note that at the corners of the triangle, we have the corresponding pure, fully saturated primaries, R, G, and B, respectively.

It happens that to match white , one needs to add a mixture of primaries such that the intensity of the R-primary is much greater than that of the other two primaries, G and B. As a result, white (W) will lie extremely close to the R-corner at $(1, 0)$, as shown in Fig. 14.3. The chromaticity diagram then has the undesirable feature of having the region of rapid variation of hues strongly concentrated in this corner, with very little variation occurring elsewhere. This feature is removed in the actual chromaticity diagram by using a different unity of intensity for each primary. This change in the diagram will be discussed in the next section.

[9] If a given spectral intensity is increased uniformly for all wavelengths by the same factor, it has been found from studies of human subjects that the color coordinates do not change. This observation amounts to saying that color is independent of brightness or, analogously, that pitch is independent of loudness.

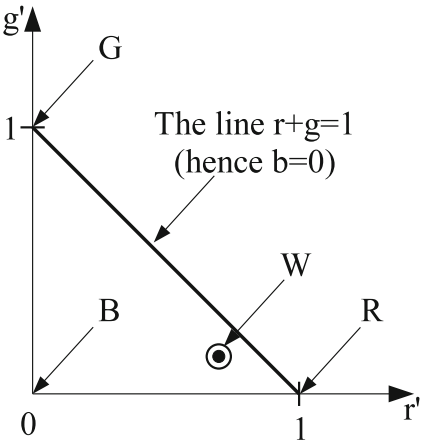


Fig. 14.3 The color coordinates are restricted to the triangle in the diagram

Ideally, monochromatic colors would lie along the two line segments, joining "B" to "G" and "G" to "R," respectively. The line segment from "B" to "R" is called the **line of purples**. **None of these purple colors is represented by monochromatic light**. In more common terms: these colors are not part of the rainbow of colors. **Magenta** is such an example.

Unfortunately, experiments reveal that it is impossible to match all colors with a sum of three primaries. Any such addition cannot be fully saturated. In the next section, we will describe the diagram and the information it contains more fully. In the section following, we will show how one calculates the color coordinates for a given spectral intensity.

14.6 A Chromaticity Diagram of Practical Use

The chromaticity diagram we will now discuss is a detailed version of the diagram described in the previous section. It is based on the experimental results obtained by Wright and Guild using the following three monochromatic primaries:

$$\lambda_B = 436\,\text{nm},\ \lambda_G = 546\,\text{nm},\ \text{and}\ \lambda_R = 700\,\text{nm}.$$

Here is why these wavelengths were chosen. The wavelengths of the first and second primaries are quite specific: They correspond to two highly monochromatic spectral lines produced by an electrical discharge excitation of mercury vapor. On the other hand, at the time of the experiments, an intense monochromatic source of a spectral line in the red region was not available. At the same time, it fortunately happens that color vision is not extremely discriminating in the red region. Thus, an intense source of a red primary, albeit with a not too small spread of wavelengths, could

be obtained from an intense light source passing through a red filter with a peak transmittance near 700 nm. In spite of this primary not being highly monochromatic, the primary is labeled with this wavelength.

For any given spectral intensity $I(\lambda)$, there is a method – to be discussed later in this section – for calculating the corresponding **color coordinates**, also called **chromaticity coordinates**. These two numbers specify the **chromaticity** of that spectral intensity. The set of all possible color coordinates corresponds to the entire set of possible colors that a person with normal vision can perceive.

> **Note**
>
> – All spectral intensities that yield the same set of color coordinates appear to have the same color and are therefore **metamers**.
> – Had a different set of primaries been used, the color coordinates corresponding to a given spectral intensity would change. However, if the color coordinates are the same for two metamers using one choice of primaries, the color coordinates will be the same for any other choice of primaries. The reason is that the color coordinates for one choice of primaries are in a one-to-one correspondence with the color coordinates with any other choice of primaries. How the two sets are related will be discussed qualitatively later in this chapter.
> – A set of primaries need not all be monochromatic. We will see that choosing monochromatic primaries is highly useful, as will be explained later in this chapter. In Appendix I, we show in mathematical detail how you can calculate one set of coordinates from another.

14.6.1 The Units for the Admixture of the Three Primaries

In the simplified discussion of Sect. 14.1, the chromaticity was expressed in terms of the intensities of the primaries, where the intensities were all expressed in the same units, say 1 W/m^2. It happens to be more convenient, though not necessary, to express the amount of each primary present using different units for each primary. The units are chosen so that[10]

$$1 \text{ Red unit} \oplus 1 \text{ Green unit} \oplus 1 \text{ Blue unit} = 1 \text{ unit of white.}$$

[10]There is an arbitrariness as to the choice of white. Some standard sources say sunlight at noon will do. Most commonly, **equal energy** white is chosen, corresponding to $I(\lambda) =$ **constant**.

In order to produce equal-energy white (W) with the Wright-Guild primaries, we need the following ratios of intensities:

$$1.000 \text{ for } R : 0.019 \text{ for } G : 0.014 \text{ for } B.$$

Thus, to produce equal energy W we can mix

$$1.00\,\mathrm{W/m^2} \text{ of } R \ \oplus\ 0.019\,\mathrm{W/m^2} \text{ of } G \ \oplus\ 0.014\,\mathrm{W/m^2} \text{ of } B.$$

Let us now introduce the following symbols:

$$1 \text{ Red unit} \equiv u_{\mathrm{R}} = 244\,\mathrm{W/m^2}$$

$$1 \text{ Green unit} \equiv u_{\mathrm{G}} = 0.019 \times 244 = 4.63\,\mathrm{W/m^2}$$

$$1 \text{ Blue unit} \equiv u_{\mathrm{B}} = 0.014 \times 244 = 3.41\,\mathrm{W/m^2}.$$

Then

- An intensity of $100\,\mathrm{W/m^2}$ of R is equal to $100/u_{\mathrm{R}} = 100/244 = 0.41$ Red units.
- An intensity of $100\,\mathrm{W/m^2}$ of G is equal to $100/u_{\mathrm{G}} = 100/4.63 = 21.6$ Green units.
- An intensity of $100\,\mathrm{W/m^2}$ of B is equal to $100/u_{\mathrm{B}} = 100/3.41 = 29.3$ Blue units.

It is important to realize that only the ratios of the chosen unit intensities matters. This is because we are matching only the color and not the intensity. We have chosen to use unit intensities that are greater than those of Williamson and Cummins by a factor of 244 to simplify their relationship to the so-called color-matching functions to be introduced later.

14.6.2 Tristimulus Values

There exists a prescription – *to be described in detail later* – for calculating the number of units of each of the three primaries which, when added, will match a given spectral intensity. These three numbers are called **Tristimulus Values** and will be labeled R, G, and B, respectively.

In particular, if R, G, and B are equal, such as $R = 2$, $G = 2$, and $B = 2$, we have a *white chromaticity*. The corresponding intensities of the three primaries would be (for the primaries 436, 546, and 700 nm):

$$I_{\mathrm{R}} = 2 \times 244 = 488\,\mathrm{W/m^2}$$

$$I_{\mathrm{G}} = 2 \times 4.63 = 9.26\,\mathrm{W/m^2}$$

$$I_{\mathrm{B}} = 2 \times 3.41 = 6.82\,\mathrm{W/m^2}.$$

14.6.3 Color Coordinates

We define the color coordinates, as follows:

$$r = \frac{R}{R+G+B}$$
$$g = \frac{G}{R+G+B}$$
$$b = \frac{B}{R+G+B}. \tag{14.2}$$

Note that $r + g + b = 1$ so that $b = 1 - r - g$; thus, r and g determine b.

> **Note**
>
> For equal energy white, $I(\lambda) = constant$, $B = G = R$. Then, $b = g = r = 1/3$. W will then be at the **center** of the chromaticity diagram, not in the corner close to R, as in Fig. 14.3.

In Fig. 14.4, we exhibit a **schematic** representation of the chromaticity diagram based on the primaries. All chromaticities are represented by color coordinates (r, g) within the **horseshoe-shaped perimeter of the chromaticity diagram**. The monochromatic primaries have color coordinates at the respective corners of the triangle: Thus R is at (1,0), G is at (0,1), and B is at the origin (0,0).

The reader should note that the actual diagram that corresponds to these primaries differs considerably from this schematic diagram. It is shown in Fig. 14.5.

Note how large a fraction of colors is unmatchable in the green region. Also shown is the boundary of the CIE-1931 XYZ chromaticity space, which will be discussed in Sect. 14.9.

14.6.4 On the Significance of the Chromaticity Diagram

The following are important characteristics of the chromaticity diagram:

1. All monochromatic (spectral) sources are represented by color coordinates which lie along the curved, upper part of the horseshoe perimeter.
2. The straight line at the lower boundary, called the **line of purples**, represents mixtures of the two monochromatic sources 400 and 700 nm.
3. Any point that lies *outside* the triangle formed by the points $(0, 0)$, $(1, 0)$, and $(0, 1)$, will involve a negative color coordinate. Examples of such chromaticities are:

$$r = 0.60,\ g = 0.42,\ b = 1 - r - g = -0.02$$
$$r = 0.60,\ g = -0.02,\ b = 1 - r - g = +0.42$$
$$r = -0.13,\ g = 0.46,\ b = 1 - r - g = +0.67.$$

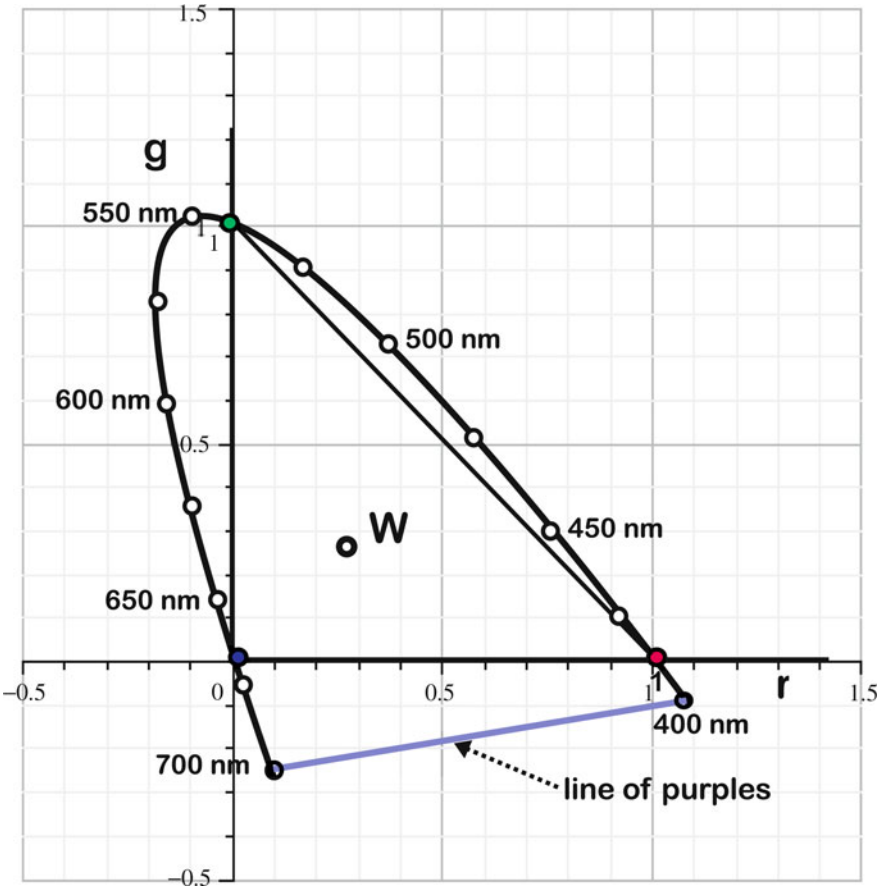


Fig. 14.4 Schematic chromaticity diagram

What is the meaning of a negative color coordinate, given that there is no meaning to a negative intensity? First of all, it implies that any chromaticity with a negative color coordinate cannot be produced by a mixture of the three primaries.

Definition: The **gamut** of colors of a given set of primaries is defined as that set of colors that can be matched by a mixture of those primaries. Thus, the gamut of colors for the above primaries have color coordinates that lie within the above triangle.

For a full answer as to how one interprets a negative color coordinate, let us consider the particular example above, with $r = -0.13$, $g = 0.46$, and $b = +0.67$. This point lies on the boundary of spectral colors and corresponds to monochromatic $\lambda = 480\,\text{nm}$, which is a monochromatic cyan.

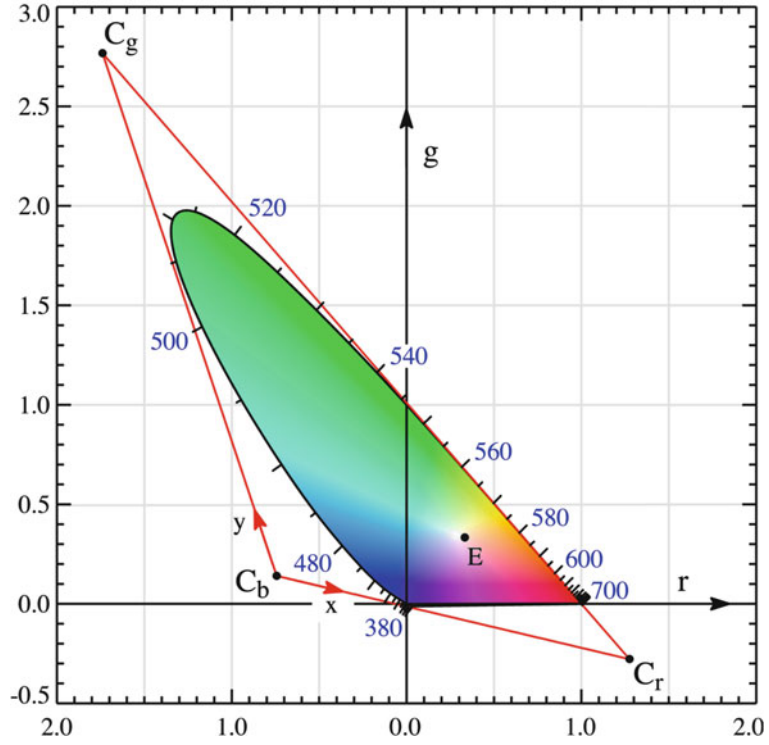


Fig. 14.5 Wright-guild chromaticity diagram along with the CIE 1931 chromaticity space

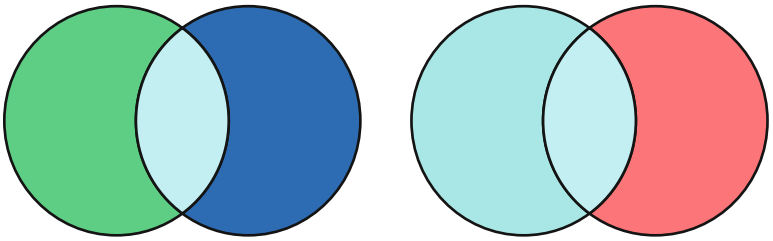


Fig. 14.6 Matching a color having a negative color coordinate

(Note that all monochromatic spectra have one negative color coordinate, **except** for a **primary** that happens to be monochromatic.)

These numbers mean that:

$$0.46 \text{ units of G } \oplus 0.67 \text{ units of B} \equiv$$
$$1 \text{ unit of monochromatic } (\lambda = 480 \text{ nm}) \oplus 0.13 \text{ units of R.}$$

The matching is indicated in Fig. 14.6.

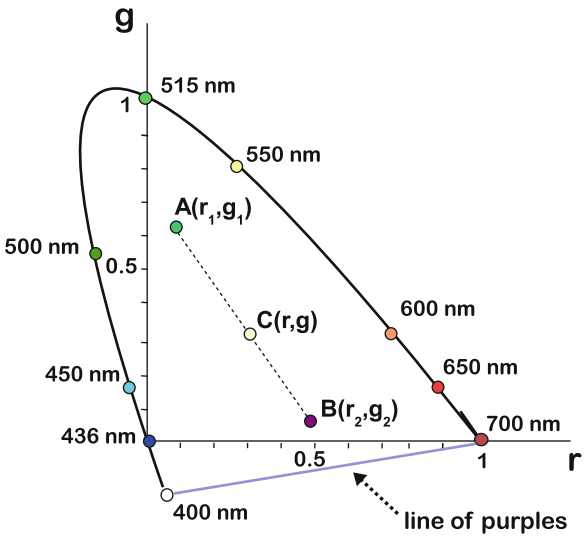


Fig. 14.7 Color coordinates for a mixture of two sources

We cannot produce monochromatic cyan by mixing green primary with blue primary. Instead, such a mixture matches a monochromatic cyan that has been **desaturated** by the addition of some red primary.

4. Pure, **equal energy white** is at the center of the horseshoe, with the color coordinates (1/3, 1/3). Recall that this is so because of the particular choice of the unit intensities of the primaries. (**White** is usually labeled with the letter **E**, for **equal energy**. We have labeled it with the letter "W".)
5. The closer a point is to the perimeter of the horseshoe, the more saturated is the color.
6. **Mixing two incoherent sources of light.**
   Suppose that we have two **incoherent sources**, with spectral intensities $I(\lambda)^{(1)}$ and $I(\lambda)^{(2)}$. **Incoherence** means that the two sources consists of **wave packets** that have a random distribution of phase relations. Therefore, the total intensity of the light is a sum of the intensities of the individual sources.

   They have sets of tristimulus values (R_1, G_1, B_1) and (R_2, G_2, B_2) and **color coordinates** (r_1, g_1) and (r_2, g_2), respectively. Now suppose that the eye receives these two sources. What will be the color coordinates (r, g) of the resulting color? The result is quite simple. The coordinates (r, g) lie along the line joining the two sets of coordinates (r_1, g_1) and (r_2, g_2), with the position dependent upon the relative strengths of the two sources. See Fig. 14.7.

   Since the sources are incoherent, the net **spectral intensity** received by the eye is the sum of the two spectral intensities.

$$I(\lambda) = I_1(\lambda) + I_2(\lambda). \tag{14.3}$$

The *strengths* of the two sources, which were alluded to above, are measured by the sum

$$S_1 = R_1 + G_1 + B_1 \text{ and } S_2 = R_2 + G_2 + B_2.$$

They are qualitatively related to the **brightness** of the sources.

In the next section, we will see how the tristimulus values and color coordinates are obtained from a given spectral intensity using a specific relation. Because the relation is linear, we have the simple result for the tristimulus values and strength of the mixture:

$$R = R_1 + R_2, \quad G = G_1 + G_2, \quad B = B_1 + B_2, \quad \text{and} \quad S = S_1 + S_2. \tag{14.4}$$

The **color coordinates** will then be

$$r = \frac{R}{S}, \quad g = \frac{G}{S}, \quad \text{and} \quad b = \frac{B}{S}. \tag{14.5}$$

Given that $r_1 = R_1/S_1$ and $g_1 = G_1/S_1$, we obtain

$$\begin{aligned} r &= \frac{R_1 + R_2}{S} = \frac{S_1}{S} r_1 + \frac{S_2}{S} r_2 \\ g &= \frac{G_1 + G_2}{S} = \frac{S_1}{S} g_1 + \frac{S_2}{S} g_2. \end{aligned} \tag{14.6}$$

The color coordinates (r, g) can be shown to lie along the line joining the two sets of color coordinates, (r_1, g_1) and (r_2, g_2). The position along the line segment depends upon the weights S_1/S and S_2/S.

It can also be shown that

$$\frac{\overline{AC}}{\overline{AB}} = \frac{S_2}{S} \quad \text{and} \quad \frac{\overline{BC}}{\overline{AB}} = \frac{S_1}{S}. \tag{14.7}$$

Thus, if the strengths are equal, with $S_1 = S_2$, then $\overline{AC} = \overline{AB}/2$ and (r, g) is at the midpoint between (r_1, g_1) and (r_2, g_2). Generally, the greater the admixture of, say, $I(\lambda)^{(2)}$, the closer will the point C be to the point B.

The result is similar to a seesaw (see Sect. 9.4): In order to balance two unequal weights, one places the fulcrum at a position such that the ratio of the distances from the fulcrum to the two weights is **inversely proportional** to the ratio of the two weights. The formulas apply, with the strengths replaced by the weights.

7. **Complementary Colors**
Any two points within the horseshoe which are on a line segment passing through W and which are on opposite sides of W are **complements** of

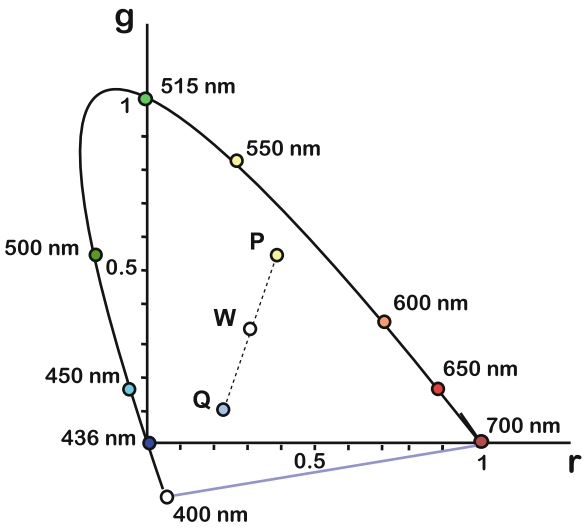


Fig. 14.8 Complements exhibited in a chromaticity diagram

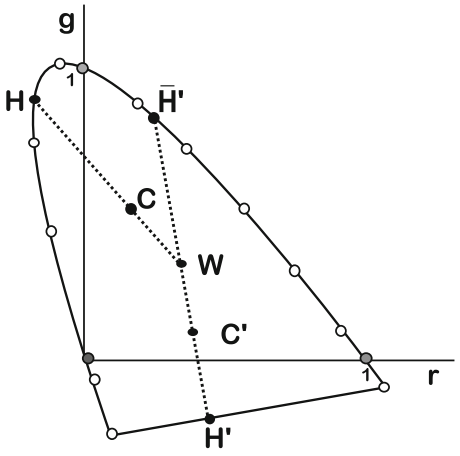


Fig. 14.9 Determination of the hue and saturation of a color

each other. This idea is illustrated in Fig. 14.8. Here, points P and Q are complements because an appropriate mixture of P and Q produces W.

8. It is clear that except for the three monochromatic primaries themselves, no monochromatic color coordinates lie along a line that joins two primaries.
   As we previously stated:
   The only fully saturated (i.e., monochromatic) chromaticities that can be produced by the full addition (without negative color coordinates) of the three primaries are the primaries themselves!
9. **A numerical characterization of color – HUE and PURITY**
   The **hue** and degree of **saturation** of a color C are given *numerical values* as follows, using Fig. 14.9.

Draw a line from W, through C – at (r, g) – to the perimeter of the horseshoe curve.

Then the **hue** is defined in terms of the wavelength of point H. This wavelength is referred to as the **dominant wavelength** of C.

The point C' does not have a dominant wavelength. Its hue is defined in terms of its complement, as follows. We extend the line from C' to the perimeter of spectral colors, reaching the point $\overline{H'}$. The point $\overline{H'}$ is the complement of H' since

$$\overline{H'} + H' \equiv W. \tag{14.8}$$

$\overline{H'}$ is referred to as the **complementary hue** of C'. The hue of C' is defined as the complement of the hue whose wavelength lies at $\overline{H'}$.

The **degree of saturation** is expressed as the **purity** of the color. It is defined as follows:

$$p \equiv \%\ \text{purity} = \frac{\overline{CW}}{\overline{HW}} \times 100\%. \tag{14.9}$$

Note, in particular, that:

(a) $p = 0$ at W.
(b) $p = 100\%$ for a monochromatic and hence a fully saturated color.
(c) Special case: Point C' is associated with the point H' on the line of purples. Then the purity is calculated using the point H' in (14.9). The hue is defined as discussed above.

10. What happens to the hue and purity of a filter if white light passes through two identical filters in sequence – as discussed in Sect. 13.2 of Chap. 13? First we expect the hue to be close in value to the hue of a single filter. Second, since the saturation increases, the purity should increase. Suppose that we express the purity as a fraction less than unity instead of as a percentage. Then, very crudely, the new purity will be approximately the square root of the purity of the single filter. Thus, a purity of 64% will lead to a purity of approximately $\sqrt{0.64} = 0.80$, or 80%.

14.7 The Calculation of Color Coordinates

In this section, we show how one calculates the color coordinates for a given spectral intensity. The method is based on a **table of color-matching functions** shown in Table 14.2. Here,we display the **color-matching functions**, $\overline{r}(\lambda)$, $\overline{g}(\lambda)$, and $\overline{b}(\lambda)$ for the Wright–Guild primaries $\lambda_B = 435.8\,\text{nm}$, $\lambda_G = 546.1\,\text{nm}$, and $\lambda_R = 700.0\,\text{nm}$.

Let us isolate the color-matching functions for the primaries themselves. See Table 14.3. Note that the array of numbers is "diagonal." That is to say, the only nonzero numbers lie along a diagonal. The nonzero values for a given primary correspond to the wavelengths of the corresponding primaries.

The unit intensities were chosen so that they are the inverses of the corresponding color-matching functions in Table 14.3. See Problem 14.18.

Table 14.2 Table of color-matching functions for the Wright–Guild primaries: 435.8 nm, 546.1 nm, 700 nm (source: C.I.E. Publication No. 15, *Colorimetry*, 1971)

λ (nm)	$\overline{r}(\lambda)$	$\overline{g}(\lambda)$	$\overline{b}(\lambda)$
380	0.00003	−0.00001	0.00117
390	0.0001	−0.00004	0.00359
400	0.0003	−0.00014	0.01214
410	0.00084	−0.00041	0.03707
420	0.00211	−0.0011	0.11541
430	0.00218	−0.00119	0.24169
440	−0.00261	0.00149	0.31228
450	−0.01213	0.00678	0.3167
460	−0.02608	0.01485	0.29821
470	−0.03933	0.02538	0.22991
480	−0.04939	0.03914	0.14494
490	−0.05814	0.05689	0.08257
500	−0.07173	0.08536	0.04776
510	−0.08901	0.1286	0.02698
515	−0.09398	0.153839	0.018589
520	−0.09264	0.17468	0.01221
530	−0.07101	0.20317	0.00549
540	−0.03152	0.21466	0.00146
550	0.02279	0.21178	−0.00058
560	0.0906	0.19702	−0.0013
570	0.16768	0.17087	−0.00135
580	0.24526	0.1361	−0.00108
590	0.30928	0.09754	0.00079
600	0.34429	0.06246	−0.00049
610	0.33971	0.03557	−0.0003
620	0.29708	0.01828	−0.00015
630	0.22677	0.00833	−0.00008
640	0.15968	0.00334	−0.00003
650	0.10167	0.00116	−0.00001
660	0.05932	0.00037	0
670	0.03149	0.00011	0
680	0.01687	0.00003	0
690	0.00819	0	0
700	0.0041	0	0

Table 14.3 Values of the color matching functions for the spectral primaries themselves (435.8, 546.1, and 700.0 nm)

λ (nm)	$\overline{r}(\lambda)$	$\overline{g}(\lambda)$	$\overline{b}(\lambda)$
435.8	0	0	0.293
546.1	0	0.215	0
700	0.00410	0	0

Comments

With

- R Red-units $\oplus$ G Green-units $\oplus$ B Blue-units matches the **chromaticity** of $I(\lambda)$ but not necessarily the brightness of $I(\lambda)$.
- Referring to the second, third, and fourth columns of the table, we note that

$$\text{Sum of } \overline{r}(\lambda) \;\cong\; \text{Sum of } \overline{g}(\lambda) \;\cong\; \text{Sum of } \overline{b}(\lambda).$$

  The reason that this is so is that if $I(\lambda)$ were the same (a constant) for all λ, corresponding to equal energy W, we must obtain $R = G = B$ (A table with more λ's listed will lead closer to equality).
- **How is the boundary of the chromaticity determined?** The color-matching functions themselves have a special significance: Consider a **monochromatic** source of light, with wavelength λ. Then the corresponding color-matching functions, $\overline{r}(\lambda)$, $\overline{g}(\lambda)$, and $\overline{b}(\lambda)$, determine the color coordinates $(r(\lambda), g(\lambda))$ of the monochromatic source of wavelength λ, since $I(\lambda)$ is nonzero only for the particular wavelength λ of interest. Thus, for the wavelength λ

$$r(\lambda) = \frac{\overline{r}(\lambda)}{\overline{r}(\lambda) + \overline{g}(\lambda) + \overline{b}(\lambda)}$$

$$g(\lambda) = \frac{\overline{g}(\lambda)}{\overline{r}(\lambda) + \overline{g}(\lambda) + \overline{b}(\lambda)}. \tag{14.10}$$

  As we move from one wavelength to another, **the color coordinates** $\{r(\lambda), g(\lambda)\}$ **produce the curve that forms the boundary of the horseshoe of the chromaticity diagram.**

  THUS, the three color coordinates for a given wavelength are nothing but the tristimulus values for a monochromatic spectral intensity with the corresponding wavelength. Now let us recognize that a single source, with a given spectral intensity, is equivalent to adding an infinite number of independent monochromatic sources involving generally all wavelengths and with variable intensity, according to the given spectral intensity. Then, the calculation of the tristimulus values amounts to adding the corresponding tristimulus values for all wavelengths as superpositions of the primaries and weighted by the spectral intensity.

  Notice that for any set, one of the three color coordinates is zero or negative. Except for the primaries themselves, the zeros in the table are actually slightly negative, yet too small to exhibit fully in the table. This property reflects the fact that the primaries cannot possibly be added to produce a match with a monochromatic source.

Consider the values of the three color-matching functions for a given wavelength λ: $\overline{r}(\lambda)$, $\overline{g}(\lambda)$, and $\overline{b}(\lambda)$. They have the property that the addition

$$\overline{r}(\lambda)\text{units of R} \oplus \overline{g}(\lambda)\text{units of G} \oplus \overline{b}(\lambda)\text{units of B}.$$

matches (i.e., produces the same color sensation as) a monochromatic source with wavelength λ.

It is very important to note our use of the symbol $\oplus$ denoting physical addition of sources rather than the plus sign that denotes **numerical** addition of intensities. For, **it is NOT necessarily true that**

$$\begin{aligned}&\overline{r}(\lambda)\text{Red-units} \oplus \overline{g}(\lambda)\text{Green-units} \oplus \overline{b}(\lambda)\text{Blue-units}\\ &\quad = \text{ intensity of the matched monochromatic source!}.\end{aligned}$$

This last comment does not hold for the monochromatic primaries themselves. Why?

Note in particular, that for $\lambda = \lambda_R$, $\overline{g}(\lambda_R) = 0$ and $\overline{b}(\lambda_R) = 0$. Thus, the tristimulus value and the color coordinate of the red primary is given by

$$R = \overline{r}(\lambda_R)\text{Red-units} = (0.00410)(244\,\text{W/m}^2) = 1.00.$$

Similar relations hold for the other two primaries.

$$G = \overline{g}(\lambda_G)\text{G-units} = 0.215 \times 4.64 = 1.00$$

$$B = \overline{b}(\lambda_B)\text{B-units} = 0.293 \times 3.42 = 1.00.$$

These numbers should be exactly equal. They are not so because of round off errors.

– In the problem set for this chapter, you will complete the above table for butter, thereby obtaining the tristimulus values, R, G, and B, for butter. You will then calculate the color coordinates using (14.2).

14.7.1 Color Coordinates of Butter

Let us see how we use this table to produce the color coordinates of butter. Study Table 14.4. In the column labeled $I(\lambda)$ is the spectral intensity for the light reflected by butter upon which equal energy WHITE light is incident (also based on Williamson and Cummins). (I have reduced the number columns for simplification at the expense of accuracy.) The three columns to the right of this one have obvious labels: For example, $\overline{r}(\lambda)\ I(\lambda)$, is the product of the second column and the fifth

Table 14.4 Analysis of butter (source: based on Williamson and Cummins, op. cit., Sects. 3–5)

λ(nm)	$\overline{r}(\lambda)$	$\overline{g}(\lambda)$	$\overline{b}(\lambda)$	$I(\lambda)$	$\overline{r}(\lambda)I(\lambda)$	$\overline{g}(\lambda)I(\lambda)$	$\overline{b}(\lambda)I(\lambda)$
400	0.00030	−0.00014	0.01214	56	0.01680	−0.00784	0.67984
420	0.00211	−0.00110	0.11541	46	0.09706	−0.05060	5.30886
440	−0.00261	+0.00149	0.31228	36	−0.09396	+0.05364	11.24208
460	−0.02608	0.01485	0.29821	32	−0.83456	0.47520	
480	−0.04939	0.03914	0.14494	34	−1.67926	1.33076	
500	−0.07173	0.08536	0.04776	35	−2.51055	2.98760	
520	−0.09264	0.17468	0.01221	49	−4.53936	8.55932	
540	−0.03152	0.21466	0.00146	63	−1.98576	13.52358	
560	+0.09060	0.19702	−0.00130	72	+6.52320	14.18544	
580	0.24526	0.13610	−0.00108	75	18.39450	10.20750	
600	0.34429	0.06246	−0.00049	78	26.85462	4.87188	
620	0.29708	0.01828	−0.00015	78	23.17224	1.42584	
640	0.15968	0.00334	−0.00003	78	12.45504	0.26052	
660	0.05932	0.00037	0	78	4.62696	0.02886	
680	0.01687	0.00003	0	77	1.29899	0.00231	
700	0.00410	0	0	77	0.31570	0	
total	0.94564	0.94654	0.94136		R = 82.1117	G = 57.8520	

column. On the bottom line are the sums of the numbers in the respective columns. In symbolic form, we have

$$\begin{aligned} R &= \text{SUM over } \lambda \text{ of products } \overline{r}(\lambda)\ I(\lambda) \\ G &= \text{SUM over } \lambda \text{ of products } \overline{g}\ (\lambda)\ I(\lambda) \\ B &= \text{SUM over } \lambda \text{ of products } \overline{b}\ (\lambda)\ I(\lambda). \end{aligned} \tag{14.11}$$

14.8 Using a Different Set of Primaries

In the Color Vision experiments, one often uses three primaries that are not monochromatic. Examples would be primaries from color filters or from the pixels used in a computer or a color TV monitor.

How can we determine the color coordinates corresponding to this new set of primaries? The new set of color coordinates will certainly be different from the first set. Must we repeat the same tests on individuals that were used to create the above table? Or, can we use the above table to create a new table that can be used to determine the color coordinates for the new set of primaries just as we did for the original primaries?

It turns out that the original table of color-matching functions contains all the information we need to know to deal with the new primaries: The spectral intensities

Table 14.5 Values of the color-matching functions for the wavelengths of the primaries of stiles and burch based on the primaries of Wright–Guild (Table 14.2)

λ (nm)	$\overline{r}(\lambda)$	$\overline{g}(\lambda)$	$\overline{b}(\lambda)$
444.44	0.01832	0.01036	0.30849
526.32	−0.07897	0.19269	0.00796
645.16	0.12975	0.00228	−0.00002

of the new primaries determine a **transformation between two sets of primaries** that allows us to calculate each member of the new table of color-matching functions. This transformation consists of **nine numbers**. If the new primaries are also spectral, these nine numbers can be calculated from **nine numbers** taken from the color-matching table: the three color-matching functions of each of the three new primaries taken from the table of the old primaries. See Table 14.5.[11] We have chosen the set of primaries that were used by Stiles and Burch to produce a table of color-matching functions for these primaries by studying the vision of a group of people with normal vision, as did Wright and Guild.[12] All of the primaries are monochromatic. $\lambda_{B'} = 444.44\,\text{nm}$, $\lambda_{G'} = 526.32\,\text{nm}$, and $\lambda_{R'} = 645.16\,\text{nm}$.

The details are presented in Appendix I. I found that both the Wright–Guild primaries and Stiles–Burch primaries lead to a chromaticity diagram with a large region requiring negative red coordinates. By playing around with other choices of the negative monochromatic primary, I arrived at a set that does extremely well in producing small regions of negative color coordinates. The wavelengths are:

$$\lambda_{B''} = 436\,\text{nm},\ \lambda_{G''} = 515\,\text{nm},\ \text{and}\ \lambda_{R''} = 700\,\text{nm}.$$

The horseshoe perimeter of the chromaticity diagram for these primaries is shown in Fig. 14.10.

14.8.1 General Features of a Different Set of Primaries

In Fig. 14.11, we exhibit the position of the color coordinates of the new set of primaries – here not spectral primaries! – based on the chromaticity diagram of the first set of primaries. Note that, to be general, we have chosen primaries whose coordinates are not on the perimeter of the horseshoe; therefore, the primaries are not monochromatic.

[11] The values in this table were obtained by interpolation, using Table 14.2.

[12] See Stiles, Walter Stanley & Birch, Jennifer M. (1958), *N.P.L. colour matching investigation: final report*. Optica Acta 6: 1–26. See also the website: http://cvrl.ioo.ucl.ac.uk/database/text/cmfs/sbrgb2.htm.

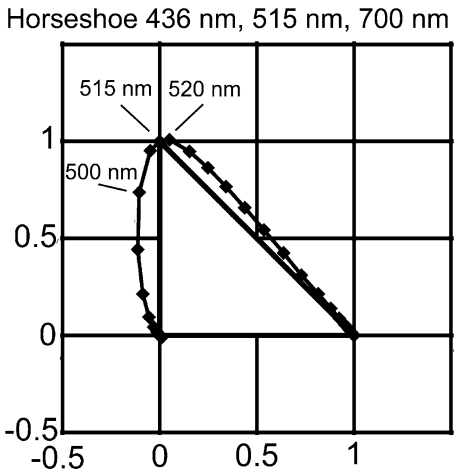


Fig. 14.10 Horseshoe perimeter for the primaries 436, 515, and 700 nm

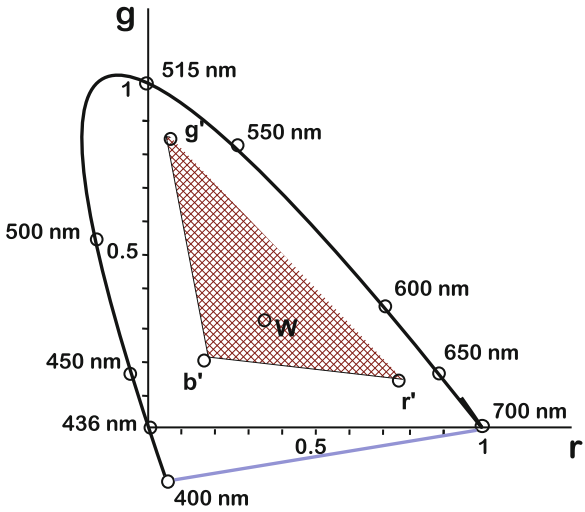


Fig. 14.11 Gamut of colors for the new primaries in the chromaticity diagram of the original spectral primaries

Note that the **gamut** of colors for the new primaries is defined by triangle $\overline{R'G'B'}$. It is clear that by *not* using spectral primaries, we reduce the number of colors in the gamut of colors. Furthermore, if the new primaries are spectral, the gamut would change: some colors matchable by the old primaries would not be matchable by the new primaries, and vice versa.

Next, we lay out the new color coordinates for all colors of interest in the r'–g' plane. Some thought will make it clear that the positions of the color coordinates of the new primaries with respect to the new primaries themselves will lie at the vertices of the triangle, as in Fig. 14.12.

There is a nice way to understand and appreciate the significance of the above transformation. Now imagine the original color coordinates laid out with respect to

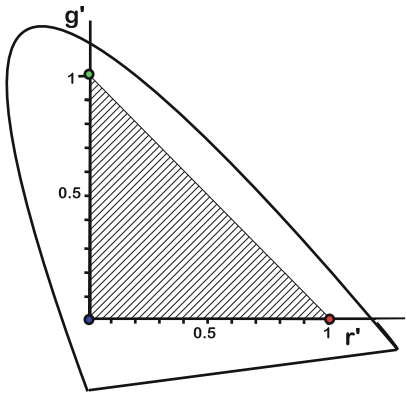


Fig. 14.12 Gamut of colors for the new primaries – the *shaded triangle* – within the full gamut of colors bounded by the horseshoe perimeter

a set of r–g axes drawn on a piece of elastic material. The above transformation amounts to distorting the elastic material by a combination of rotation, and stretching and/or compressing. In particular, the color coordinates of the new primaries would move to their respective corners of the triangle.

It can be shown that the only chromaticities that can be matched by the full addition (i.e., with no subtractions) of primaries are located within the triangle formed by the points R', G', and B'. One strives to choose primaries that at least have W included, as shown in the figure and as saturated as possible. With all other factors being equal, it is best to choose primaries that are monochromatic.

14.9 The Standard Chromaticity Diagram of the C. I. E.[13]

In principle, any set of primaries can be used to produce a table of color-matching functions that maps any color onto a set of color coordinates. To the extent that the experiments used in testing the color vision of a group of people have been carried out with care, all reflect color vision reliably and contain the same information! Any set can be used to establish a standard table for labeling a color. All tables of color-matching functions map onto one another in a one-to-one correspondence. It was decided early on with the advent of the results of Wright and Guild to establish a universal standard table of color-matching functions, the C.I.E. table, that is independent of any set of primaries that one might choose to use. The table maps any spectral intensity onto a set of **CIE tristimulus values** – labeled X, Y, and Z.

[13]Commission Internationale de l'Éclairage. See Wikipedia (1-6-2011): http://en.wikipedia.org/wiki/CIE_1931_color_space.

In place of the color-matching functions $\overline{r}(\lambda)$, $\overline{g}(\lambda)$, and $\overline{b}(\lambda)$, we have the color-matching functions $x(\lambda)$, $y(\lambda)$, and $z(\lambda)$.

Note

- The mapping was chosen so that all tristimulus values are positive.
- The tristimulus value Y has the specific significance of providing the **luminance** of the spectral intensity. Luminance is based on the variation of brightness with wavelength with fixed intensity. It corresponds to the phon level of sound.[14]
  In line with the definition of luminance, the color-matching function $\overline{y}(\lambda)$ is chosen to be proportional to the relative sensitivity of the eye with respect to wavelength.
- On the other hand, the tristimulus values do not represent any linear measure of the intensity of any primary. They cannot since they are not specific to any set of primaries.
- The color coordinates that replace the set $r = R/(R + G + B)$, $g = G/(R + G + B)$, and $b = G/(R + G + B)$ are $x = X/(X + Y + Z)$, $y = Y/(X + Y + Z)$, and $z = Z/(X + Y + Z)$. Obviously, only two color coordinates specify a color since z = 1-x-y. And none (x, y, or z) are negative.
- The coordinates $x = y = z = 1/3$ represents equal energy white.
- When both the RGB values and the XYZ values are determined by a given spectral intensity, there is a straightforward transformation to obtain one set from the other without having to specify the spectral intensity from which they were calculated. (Remember that we have metamerization so that neither set of tristimulus values determines a specific spectral intensity!)

The result is the standard C.I.E. chromaticity diagram. See Fig. 14.13. The C.I.E. Color Matching Functions are found in Table 14.6.

The first such set of C.I.E. color-matching functions and chromaticity diagram were obtained from the Wright–Guild data. However, the results of later improved testing has led to modified C.I.E. standards.

[14]Recall from Chap. 10 that the loudness in phons is not directly related to the perceived loudness; the latter is measured by the **sone** level, which we learned is proportional to $I'^{0.3}$ where I' is a scaled intensity that takes into account the equal loudness curves. Similarly, the perceived brightness for a given wavelength is not proportional to the intensity. The actual perceived brightness with respect to intensity is expressed by the **lightness** L^* wherein approximately, $L^* \propto Y^{\beta}$, where β is an exponent. Some sources claim that $\beta = 0.3$. (See (1-12-2011): http://en.wikipedia.org/wiki/Lightness_(color)). However, others point out that the value varies depending upon whether the eye has adapted to the level of light intensity and that it can vary from about 0.4 for the dark adapted eye to about 0.5 for the light adapted eye.

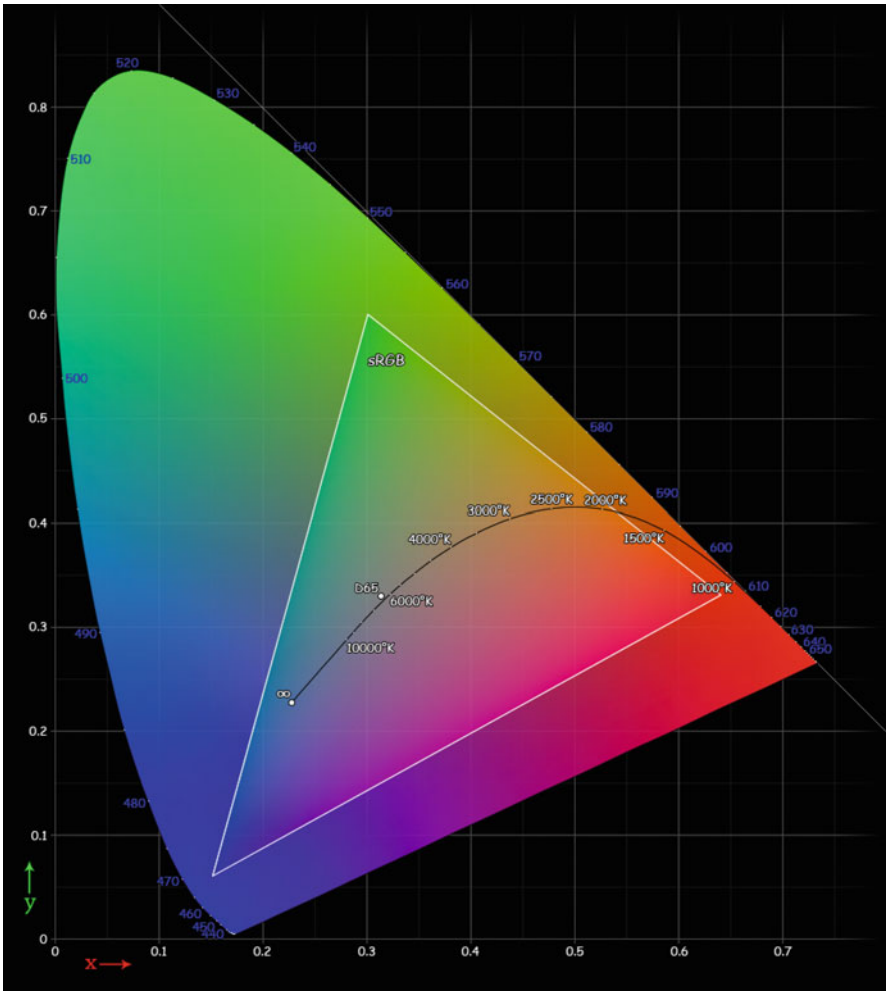


Fig. 14.13 Standard CIE chromaticity diagram (source: http://upload.wikimedia.org/wikipedia/commons/6/60/Cie_Chart_with_sRGB_gamut_by_spigget.png)

Note that the horseshoe region is shifted and distorted, but lies entirely within the first quadrant. In the same figure, we see a black curve representing the colors for the spectra of black body radiation at various temperatures, labeled from 2,000 to 25,000 K. (See Sect. 6.7 on black body radiation for more details.) Note that the temperature of the sun – 6,000 K – is close to white in color.

We also see a highlighted triangle. At the corners are the color coordinates of what are referred to in the figure as the **primary illuminants**. Those shown in the figure are the sRGB primaries, which are the standard used in many color monitors. We note that the gamut of colors (the triangle) produceable with the sRGB primaries is limited. The region that is most omitted is toward the "green corner." It happens

Table 14.6 Table of color-matching functions – C.I.E. 1964 (source: http://www-cvrl.ucsd.edu/cmfs.htm)

λ (nm)	x_λ	y_λ	z_λ	$\lambda(nm)$	x_λ	y_λ	z_λ
390	0.0023616	0.0002534	0.0104822	565	0.793832	0.98238	0
395	0.0072423	0.000769	0.032344	570	0.878655	0.955552	0
400	0.0191097	0.0020044	0.0860109	575	0.951162	0.915175	0
405	0.0434	0.004509	0.19712	580	1.01416	0.868934	0
410	0.084736	0.008756	0.389366	585	1.0743	0.825623	0
415	0.140638	0.014456	0.65676	590	1.11852	0.777405	0
420	0.204492	0.021391	0.972542	595	1.1343	0.720353	0
425	0.264737	0.029497	1.2825	600	1.12399	0.658341	0
430	0.314679	0.038676	1.55348	605	1.08917	0.593878	0
435	0.357719	0.049602	1.7985	610	1.03048	0.527963	0
440	0.383734	0.062077	1.96728	615	0.95074	0.461834	0
445	0.386726	0.074704	2.0273	620	0.856297	0.398057	0
450	0.370702	0.089456	1.9948	625	0.75493	0.339554	0
455	0.342957	0.106256	1.9007	630	0.647467	0.283493	0
460	0.302273	0.128201	1.74537	635	0.53511	0.228254	0
465	0.254085	0.152761	1.5549	640	0.431567	0.179828	0
470	0.195618	0.18519	1.31756	645	0.34369	0.140211	0
475	0.132349	0.21994	1.0302	650	0.268329	0.107633	0
480	0.080507	0.253589	0.772125	655	0.2043	0.081187	0
485	0.041072	0.297665	0.57006	660	0.152568	0.060281	0
490	0.016172	0.339133	0.415254	665	0.11221	0.044096	0
495	0.005132	0.395379	0.302356	670	0.0812606	0.0318004	0
500	0.003816	0.460777	0.218502	675	0.05793	0.0226017	0
505	0.015444	0.53136	0.159249	680	0.0408508	0.0159051	0
510	0.037465	0.606741	0.112044	685	0.028623	0.0111303	0
515	0.071358	0.68566	0.082248	690	0.0199413	0.0077488	0
520	0.117749	0.761757	0.060709	695	0.013842	0.0053751	0
525	0.172953	0.82333	0.04305	700	0.00957688	0.00371774	0
530	0.236491	0.875211	0.030451	705	0.0066052	0.00256456	0
535	0.304213	0.92381	0.020584	710	0.00455263	0.00176847	0
540	0.376772	0.96199	0.013676	715	0.0031447	0.00122239	0
545	0.451584	0.98220	0.007918	720	0.00217496	0.00084619	0
550	0.529826	0.99176	0.003988	725	0.0015057	0.00058644	0
555	0.616053	0.99911	0.001091	730	0.00104476	0.00040741	0
560	0.705224	0.99734	0	735	0.00072745	0.000284041	0

that in this region, the eye is poor at seeing color differences with respect to changes in color coordinates. The **white** for sRGB primaries has coordinates labeled by "D65" and is close to a black body radiation of temperature 6,500 K. Note that the coordinates are not $x = y = 1/3$. Instead, they are $x = 0.3127$ and $y = 0.3291$, which are supposed to correspond to the color of the sky in Europe at midday.

We will now see how Fig. 14.13 is mapped onto Fig. 14.5. The corners of the right triangle (incompletely shown) in Fig. 14.13, with coordinates, $x = 1, y = 0$, $x = 0, y = 1$, and $x = 0, y = 0$, correspond to the points labeled C_r, C_g, and C_b,

respectively, in Fig. 14.5. The line from C_b to C_r corresponds to the line segment along the x-axis, from $x = 0$ to $x = 1$, along which $y = 0$ and has therefore $Y = 0$. The line $Y = 0$ corresponds to **zero luminosity** and is referred to by the term **alychne**.

14.10 From Computer RGB Values to Color**

We have noted that a computer stores colors using the **RGB** coordinates. In this section, we will discuss how these coordinates are used to determine the XYZ coordinates of a color pixel. In turn, the XYZ coordinates determine the visual chromaticity and intensity produced by the pixel. Our focus will be on color monitors. However, the essential issues raised apply to color printers as well. We will assume the use of 24-bit color.

We begin by presenting the parameters that characterize a particular color monitor.

- The chromaticity of each of the three primaries – this is expressed in terms of their {xy} coordinates, resulting in six numbers that are used to carry out a transformation between the two sets of coordinates. For the so-called sRGB primaries, they are $\{x_r = 0.64, y_r = 0.33\}$, $\{x_g = 0.30, y_g = 0.60\}$, and $\{x_b = 0.15, y_b = 0.06\}$.
- The chromaticity – hence the {xy} coordinates – of the color corresponding to $\mathbf{R} = \mathbf{G} = \mathbf{B}$, which provide us with two more numbers for the transformation. The intensities of light for each primary are related to the tristimulus values so as to produce the chosen white. Typically, the chosen white is not equal energy white. For the so-called sRGB primaries we have $\{x_w = 0.3127, y_w = 0.3291\}$; this white, with the name **D65**, corresponds to the color of blackbody radiation at a temperature of 6500 K and is reported to be close to the color of a clear midday sky in Western Europe.
- The last parameter – the γ **value** – reflects the fact that the tristimulus values {RGB} are not necessarily proportional to the computer's respective {**RGB**} values. The history of this situation is complex. A good approximation to the relationship is what is referred to as a simple power law:

$$R = \left(\frac{\mathbf{R}}{255}\right)^{\gamma}, \quad G = \left(\frac{\mathbf{G}}{255}\right)^{\gamma}, \quad B = \left(\frac{\mathbf{B}}{255}\right)^{\gamma}. \tag{14.12}$$

  Note that this scaling of the tristimulus values results in a range from zero to unity, so that maximum luminance or brightness is achieved with a value of unity. Note too that if γ were unity, we would have proportionality; however, typically its value lies between 1 and 3. A reason for the nonlinearity is that the cathode ray tubes (CRTs) that produce color have a light intensity I that is far from linear with respect to the strength of the electric signal – the so-called voltage

V – that is responsible for the light. Approximately, $I \propto V^{\gamma}$. The {RGB} values are approximately proportional to the voltage, thus resulting in (14.12).[15]

Common values of gamma are 1.8 or 2.2 . For PCs, the value is usually $\gamma = 2.2$. MAC monitors have fluctuated between both values.

Note

There are important consequences of nonlinear mapping from {**RGB**} to RGB.

- Suppose that all **RGB** values are multiplied by a constant c. Then we can see from (14.12) that all values of RGB are multiplied by the constant c^{γ}. For example, let **R** change to 2**R**. Then

$$R = \left(\frac{\mathbf{R}}{255}\right)^{\gamma} \rightarrow \left(\frac{\mathbf{2R}}{255}\right)^{\gamma} = 2^{\gamma}\left(\frac{\mathbf{R}}{255}\right)^{\gamma} = 2^{\gamma} R. \qquad (14.13)$$

  We then have no change in the chromaticity because all three RGB values change by the same factor.[16]
- The effect of nonlinearity is dramatic when we do not add {**RGB**} values in the same proportion. For example, suppose that we add $\mathbf{R_2} = \mathbf{150}$ to $\mathbf{R_1} = \mathbf{100}$ and double the **G** and **B** values from 100 to 200. We will have $\mathbf{R} = \mathbf{100} + \mathbf{150} = \mathbf{250}$, so that the resulting value of R will be

$$R = \left(\frac{\mathbf{250}}{255}\right)^{\gamma}. \qquad (14.14)$$

  The result is that R will change by a factor different than that of G and B, so that the chromaticity will change: R changes by a factor 2.5^{γ} vs. a factor of 2^{γ} for G and B. This result is not surprising; we expect the chromaticity to change. On the other hand, more importantly, while $\mathbf{R_1} + \mathbf{R_2} = \mathbf{R}$, $R_1 + R_2 \neq R$:

$$R_1 + R_2 = \left(\frac{\mathbf{100}}{255}\right)^{\gamma} + \left(\frac{\mathbf{150}}{255}\right)^{\gamma} \neq \left(\frac{\mathbf{250}}{255}\right)^{\gamma} = R. \qquad (14.15)$$

  See the last problem at the end of the chapter for an interesting effect of gamma nonlinearity.

[15]The subject of gamma and its related **gamma correction** is very complex. As a result, it is extremely difficult to find resources that are reliable. Articles abound with contradictory information. For what I consider a very reliable reference I highly recommend Charles Poynton, *Video and HDTV*, [Morgan Kaufmann Publishers and Elsevier Science, San Francisco, 2003].

[16]Tests of some monitors have revealed that their three **RGB** values do not have the same value of gamma. In this case, the chromaticity will change.

Suppose that you want to see the color associated with a given spectral intensity on a color monitor. A straightforward way to do so is to find the tristimulus values for the associated primaries and chosen white point, and then obtain the computer **RGB** values.

In Table 14.7, we present the color-matching functions for sRGB with a D65 white point.[17] Once you obtain the sRGB tristimulus values, you can proceed to calculate the color coordinates {r,g,b}, as we have seen earlier in the chapter. Finally, we can use the equations inverse to those of (14.12) to obtain the **RGB** values. They are given by

$$\mathbf{R} = 255\, r^{1/\gamma} \quad \mathbf{G} = 255\, g^{1/\gamma} \quad \mathbf{B} = 255\, b^{1/\gamma}. \tag{14.16}$$

14.11 How Many Colors Are There?

The answer to this question depends upon how we define colors. Whatever the choice, we have to be able to assign a numerical value. We cannot be vague here. In the final analysis, we need to make use of our knowledge of the chromaticity diagram as well as color vision studies. Thus, it is very important to avoid making rapid conclusions. Recently (April, 2010) the Sharp Electronics Corporation announced to the public that it is offering a four primary **color monitor**. This monitor will certainly increase the range of colors that the monitor will be able to display over that of a three primary color monitor. There are various reports as to how many more colors will be displayable. Most of the blogs are full of nonsense because the authors do not know enough about color vision.[18]

So let us start to examine the question very slowly. Here are some possible choices of what we can mean by countable colors.

- The number of spectral intensities. We have pointed out that this number is a high level of infinity – labeled by mathematicians as $\aleph_0^C$.
- Number of chromaticities (sets of color coordinates, each ranging continuously from 0 to 1). This number is also infinite but at a lower level of infinity. The mathematical infinity for the number of points in a finite area (the horseshoe of the chromaticity diagram) is $\aleph_0$.
- The number of colors that are visually distinguishable – estimated at about 26,000.

[17]The table was produced by using transformation matrices between the CIE table of color-matching functions and the RGB coordinates for the sRGB primaries.

[18]For example, in the website (1-12-2011): http://www.gizmag.com/sharp-4-primary-color-tvs-enables-trillion-colors/13823/
we read: "By adding yellow to the colors red, green and blue, the televisions are capable of rendering nearly all the colors a human eye can discern."

Table 14.7 Table of color-matching functions for sRGB primaries with D65 white point

λ (nm)	$\overline{r}(\lambda)$	$\overline{g}(\lambda)$	$\overline{b}(\lambda)$	λ (nm)	$\overline{r}\lambda$	$\overline{g}\lambda$	$\overline{b}\lambda$
390	0.0020379	−0.0013774	0.0111593	565	1.0624985	1.0735629	−0.1562685
395	0.0061633	−0.0042311	0.0344334	570	1.3786552	0.9410231	−0.1460794
400	0.0159679	−0.0111828	0.0915671	575	1.6757260	0.7950021	−0.1338111
405	0.0354432	−0.0254042	0.2098490	580	1.9509934	0.6471963	−0.1208752
410	0.0670300	−0.0495022	0.4144850	585	2.2124935	0.5076572	−0.1086960
415	0.1061226	−0.0818657	0.6990658	590	2.4299409	0.3743422	−0.0964009
420	0.1449626	−0.1176064	1.0349829	595	2.5687956	0.2520187	−0.0838849
425	0.1732094	−0.1478947	1.3643045	600	2.6307181	0.1456766	−0.0718077
430	0.1858490	−0.1678059	1.6516346	605	2.6169719	0.0584916	−0.0605933
435	0.1863771	−0.1788303	1.9107849	610	2.5280954	−0.0082826	−0.0504098
440	0.1673589	−0.1736197	2.0880869	615	2.3713247	−0.0550566	−0.0413530
445	0.1277173	−0.1503345	2.1491184	620	2.1632857	−0.0831681	−0.0335935
450	0.0693082	−0.1084812	2.1108656	625	1.9246978	−0.0946749	−0.0272949
455	0.0004766	−0.0539885	2.0064321	630	1.6625984	−0.0956921	−0.0218334
460	−0.0876709	0.0201495	1.8355095	635	1.3833738	−0.0904241	−0.0168117
465	−0.1866384	0.1050043	1.6264932	640	1.1222411	−0.0809174	−0.0126898
470	−0.3076486	0.2126340	1.3657585	645	0.8983389	−0.0700685	−0.0094939
475	−0.4228504	0.3271911	1.0514122	650	0.7041793	−0.0581450	−0.0070380
480	−0.5139261	0.4298260	0.7688802	655	0.5373194	−0.0457007	−0.0052031
485	−0.6087477	0.5423271	0.5441134	660	0.4017969	−0.0347817	−0.0038145
490	−0.6760153	0.6378142	0.3706395	665	0.2958794	−0.0260298	−0.0027567
495	−0.7419776	0.7493351	0.2392183	670	0.2144757	−0.0191002	−0.0019692
500	−0.8049760	0.8698089	0.1371703	675	0.1530033	−0.0137450	−0.0013898
505	−0.8462604	0.9884878	0.0607874	680	0.1079449	−0.0097546	−0.0009733

(continued)

Table 14.7 (continued)

λ (nm)	$\overline{r}(\lambda)$	$\overline{g}(\lambda)$	$\overline{b}(\lambda)$	λ (nm)	$\overline{r}\lambda$	$\overline{g}\lambda$	$\overline{b}\lambda$
510	−0.8672447	1.1065961	−0.0032616	685	0.0756554	−0.0068610	−0.0006791
515	−0.8638713	1.2205595	−0.0489710	690	0.0527167	−0.0047904	−0.0004720
520	−0.8197702	1.3174593	−0.0846822	695	0.0365982	−0.0033320	−0.0003269
525	−0.7267116	1.3787319	−0.1128393	700	0.0253230	−0.0023074	−0.0002259
530	−0.5942649	1.4139555	−0.1332074	705	0.0174647	−0.0015906	−0.0001559
535	−0.4445743	1.4390806	−0.1497857	710	0.0120362	−0.0010948	−0.0001076
540	−0.2646612	1.4400910	−0.1608415	715	0.0083127	−0.0007546	−7.452E-05
545	−0.0503985	1.4052614	−0.1668914	720	0.0057481	−0.0005205	−5.169E-05
550	0.1904443	1.3472022	−0.1686456	725	0.0039784	−0.0003592	−3.592E-05
555	0.4600521	1.2772972	−0.1684127	730	0.0027597	−0.0002483	−2.502E-05
560	0.7523205	1.1875067	−0.1642469	735	0.0019210	−0.0001722	−1.75E-05

– The number of sets of **RGB** of color coordinates stored in a computer and associated with colors on a monitor. The settings are variable on a given monitor. For a three color, **24-bit** monitor setting, we have $2^8 = 256$ (8-bits) different values for each color, **R**, **G**, and **B** , ranging from 0, 1, 2, . . . , 255. The total number of combinations is then $256 \times 256 \times 256 = 2^{24} = 16,777,216$.[19] Another option is 36-bit color, which translates to a total of $2^{12} = 4096$ settings for each of **R**, **G**, and **B**, giving a total of $2^{36} \sim$70-billion total number of settings.
This number corresponds to the number of different possible tristimulus values produced by a monitor – but not the number of chromaticities.
– The number of different chromaticities produced by a given color monitor. This number can be calculated too. See the sample problem below.

Sample Problem 14-1

Why is the number of colors as defined in the context of color monitors not the same as the number of colors we would deduce using our definition of color?

Solution

We defined **color** in terms of hue and saturation but did not include brightness or intensity. On the other hand, the 16,777,216 different values of **RGB** of a monitor determine hue, saturation, and luminance (which is an objective measure of brightness). For our future discussion, to avoid any mistake as to usage, we will refer to **hs colors** as determined by hue and saturation; on the other hand, we will refer to **hsb colors** as determined by hue, saturation, and brightness.

A crude estimate of the number of color coordinates we obtain from the **RGB** values of a monitor is to treat them as tristimulus values. (For technical reasons, the tristimulus values are not proportional to the monitor's **RGB** values.)

The number of different hs colors according to our definition of color would be the number of distinct fractions **r=R/(R+G+B)** and **g=G/(R+G+B)**. We expect this number to be certainly less than 16,777,216, but still on the order of a few million. Bruce Boghosian estimated this number at about 14 million and then computed it to be exactly 13,936,094.

[19] Adding a fourth color with 8-bits will increase this number by a factor of 256, so that we would have over four billion different combinations! In a recent (May, 2010) website of the SHARP Corporation, it was claimed that their monitor would produce trillions of colors. It is incomprehensible to understand how they can arrive at such a number. See SHARP website: http://www.sharpusa.com/AboutSharp/NewsAndEvents/PressReleases/2010/January/2010_01_06_Booth_Overview.aspx.

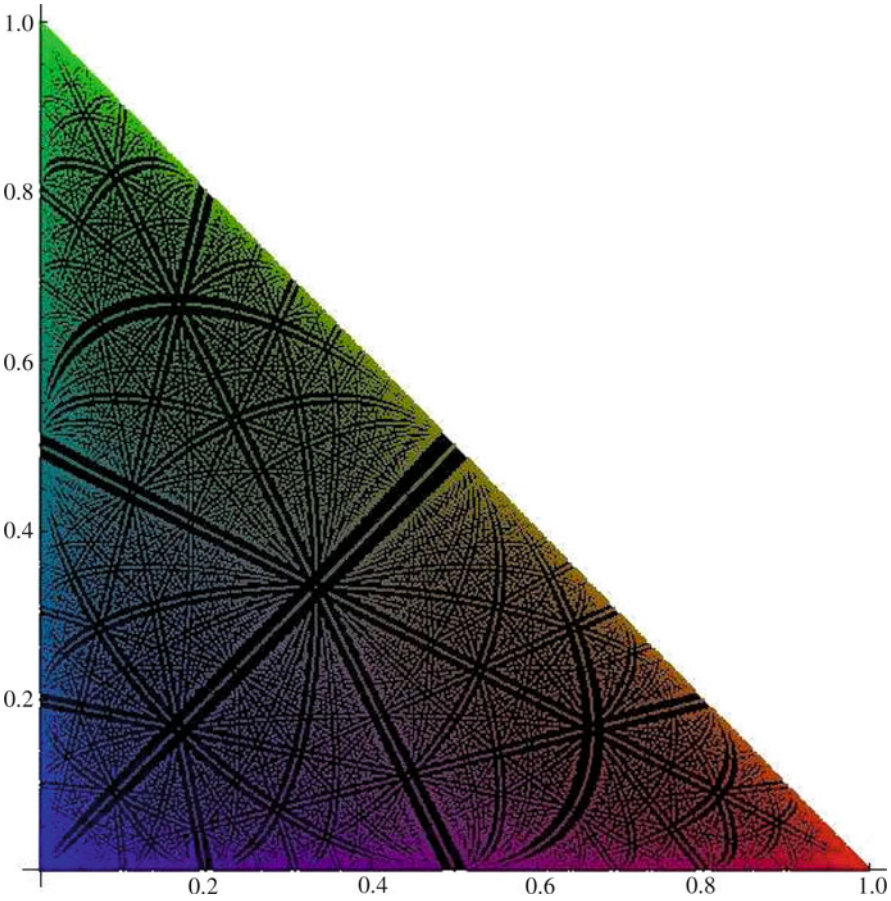


Fig. 14.14 Colors produced by three primaries on a 15-bit color monitor with **RGB** parameters having values from zero to 31 each – instead of the common zero to 255 each

In Fig. 14.14, we see the color coordinates produced by all possible combinations of **R** = 0 to 31, **G**=**0** to 31, and **B**=0 to 31 in a chromaticity diagram. The monitor then has 15-bit color. The triangle represents the gamut of colors by the primaries of the monitor. There are $32 \times 32 \times 32 = 2^{15} = 32,768$ colored dots placed within the black triangular background of the figure. Each colored dot represents a possible producible color. Note that there are "avenues" of regions where colors cannot be produced!

Finally, we arrive at the ultimate way to count colors – what matters as far as color vision is concerned – what we see!

While the number of hsb colors and hs colors produced by a 24-bit monitor is in the millions, the question remains as to whether we can **distinguish** among all these

colors. In concrete terms, we can have the two sets of color coordinates (r,g) in the chromaticity diagram, one for each of two colors, so close together that we cannot tell the colors apart.

Consider two points in the CIE Chromaticity Diagram that correspond to the spectral intensity of blackbody radiation at two different temperatures, 2,000 and 3,000 K. (See Sect. 6.7.) For this author, the two colors are barely distinguishable in the figure. I have concluded that the two colors in the figure do not represent the actual colors represented by the color coordinates. You can check whether this is so for you by varying the color coordinates **(R,G,B)**, each ranging from 0 to 255, on your color monitor and determining how well you can discriminate colors.

For any point in the diagram, we can draw a small ellipse such that all points within the ellipse produce colors that are not distinguishable. We see a number of such ellipses in Fig. 14.15 below drawn on the standard CIE Chromaticity Diagram. The ellipses are drawn ten times their actual size.[20]

It is possible to draw a set of ellipses that fills the horseshoe with minimal overlap. The total number of these ellipses will be a good measure of the **number of distinguishable colors**. It is often stated that the number of distinguishable colors is on the order of ten million.[21] However, this number refers to the number of hsb colors, which takes into account varying brightness, and not hs colors. According to recent studies, the number of distinguishable hs colors is on the order of 26,000.[22, 23]

In Fig. 14.16, we see a chromaticity diagram based on the monochromatic primaries 415, 515, and 700 nm.[24] I will refer to this diagram as the **lRGB gamut**.

The red triangle has corners at the points (0,0), (1,0), and (0,1). Thus all the hs colors that can be matched with these monochromatic primaries have coordinates that lie within this triangle. Only a small fraction of all colors lies outside this triangular gamut. The monochromatic colors lie on the blue perimeter. Also displayed are the color coordinates for a standard set of primaries known as **sRGB**. They lie at the corners of the black triangle. If you look closely you will note that the gamuts are displayed on a grid. There are 100 horizontal lines and 100 vertical lines passing through the red triangle. The number of squares in the triangle is therefore $100^2/2$, or 5,000. Since there are about 26,000 distinguishable hs colors, a single square is associated with about five distinguishable hs colors.

---

[20] Note that the ellipses are largest in the green region, indicating that the eye does not discriminate changes in chromaticity as well. On the other hand, the ellipses are much smaller towards the blue region. We can see this variation in discrimination in the CIE chromaticity diagram of Fig. 14.13.

[21] See D. B. Judd and G. Wyszecki (1975), *Color in Business, Science and Industry, Wiley Series in Pure and Applied Optics (3rd ed.). New York: Wiley-Interscience. p. 388.*

[22] J. M. Linhares, et al., J Optical Society of America, volume **25**, p. 2918 (2008).

[23] This number is just under 20 times the number of distinguishable pitches of pure tones, which has been found to be about 1,400. See Wikipedia (1-7-2011): http://en.wikipedia.org/wiki/Pitch_(music).

[24] In Appendix I, I show how this set of primaries is close to producing the largest possible gamut of colors.

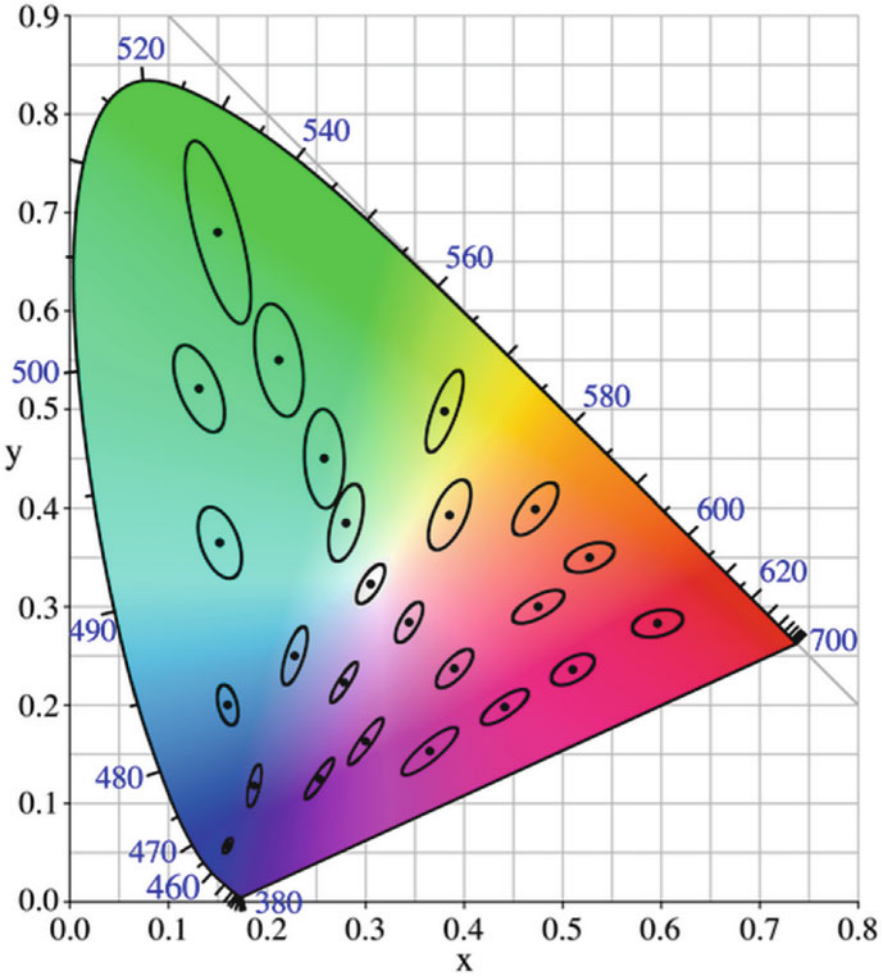


Fig. 14.15 Ellipses representing the area of colors that are indistinguishable from the color at the center of each respective ellipse – drawn ten times to actual scale (source: http://en.wikipedia.org/wiki/File:CIExy1931_MacAdam.png, attributed to David MacAdam)

Current advertisements of color monitors brag about their gamut of colors – all points lying within the black triangle – in terms of the sRGB standard. A company might claim that their monitor has a gamut of 117% of the sRGB standard. From the fact that the human eye can distinguish about ten million hsb colors, it is reasonable to assume that the 16.8 million monitor values of RGB in the sRGB gamut can produce essentially all the **distinguishable colors** that lie within the sRGB gamut.

We note that the sRGB standard is quite poor in the green area. If only the green primary at the top were moved along the line from R to G all the way up toward the perimeter, the gamut would be a good fraction of the optimum lRGB gamut!

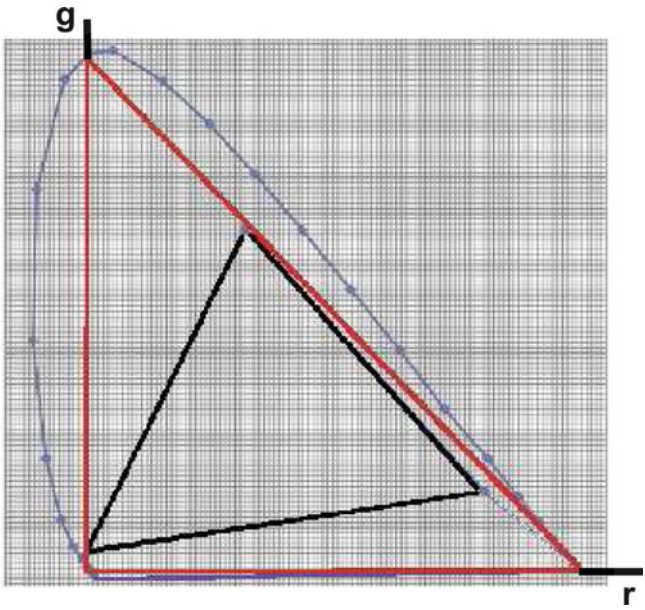


Fig. 14.16 sRGB coordinates (at the corners of the *black triangle* on a Chromaticity Diagram with monochromatic primaries 436,515, and 700 nm. The *blue dots* along the horseshoe perimeter are monochromatic colors

So how would the addition of a fourth primary in a color monitor increase the set of distinguishable colors? The answer is dealt with in Problem 14.23a.It should be clear that the number of distinguishable hsb colors is not increased by even a factor of 256.

14.11.1 Limitations of a Broadened Gamut of a Monitor

We recognize that Sharp's four-color monitor will broaden the gamut of colors that can be produced. However, can it **faithfully reproduce** a broader gamut of colors? Suppose that an image or video shown by the monitor was produced by a camera with a smaller gamut? How would the monitor handle the input? What might be the gain?[25] To answer these questions, we need to know how the camera handles a color that lies outside its gamut. The typical response of the camera is to **replace** a color outside its gamut by another that has the same hue but lies at the boundary of the

[25]I am grateful to Raymond Soneira for communication on this subject. You are invited to see his extremely informative website (1-27-2011): http://www.displaymate.com/eval.html.

gamut. As a result, the output will be a color that has no relation to how saturated the original color was. Sharp's monitor cannot reproduce a replication of the original; all it can do is to increase the saturation of the color by an amount not determined by the original color. Our conclusion is that the Sharp monitor can enrich the color of the image but **not faithfully**.

14.12 A Simple Physiological Basis for Color Vision

We mentioned that there are three different types of cones, or receptors of light.[26] The genes for the three cones have been identified in the laboratory.[27] They were originally hypothesized to exist as a means of understanding the results of color matching that are summarized in the chromaticity diagram.

We suppose that the three types of cones are distinguishable by their different spectral response curves, that is, the curves that describe how the rate at which each the individual type of cone produces nerve impulses depends upon the wavelength of a monochromatic source. These response curves are proportional to the respective **absorption spectra** of the respective **pigments** in the cones: Thus, R-cones have R-pigment, G-cones have G-pigment, and B-cones have B-pigment.

The absorption spectrum tells us how the fraction of monochromatic incident light intensity that is absorbed depends upon the wavelength λ. The presumed spectra are shown in Fig. 14.17.[28] The dashed curve is the absorption of the rods. "S" stands for "short wavelength," "M" stands for "medium wavelength," and "L" stands for "long wavelength."

Note that all the curves have a maximum value of 100. This is because the curves represent the variation of each cone with respect to wavelength; the actual *relative* absorption of one cone to another is quite different. Absorption by the blue cone is far weaker than the other two.

Consider, for example, the absorption spectrum of the L-cones. The curve has a peak at $\lambda = 564\,\text{nm}$. Notice how broad the peak is; the width is on the order of 150 nm, corresponding to about $150/440 = 0.34$, or 34% of the wavelength at the peak. Recall, however, that the width of a spectral line produced by a single atomic or molecular transition is very narrow. For atomic transitions, the width is on the order of **one part in ten million** of the frequency of the photon emitted.[29]

---

[26]For more details, see http://en.wikipedia.org/wiki/Retina and http://webvision.med.utah.edu/sretina.html.

[27]See the article by Jeremy Nathans, who first identified the genes: Scientific American, volume 260, pp. 42–49 (1989).

[28]The figure is based on Bowmaker J.K. and Dartnall H.J.A., "Visual pigments of rods and cones in a human retina." J. Physiol. 298: pp501–511 (1980).

[29]If we plot the absorption spectrum as a function of the frequency, we would obtain a peak for the L-cone that has a width in frequency that is about 34% of the frequency at the peak.

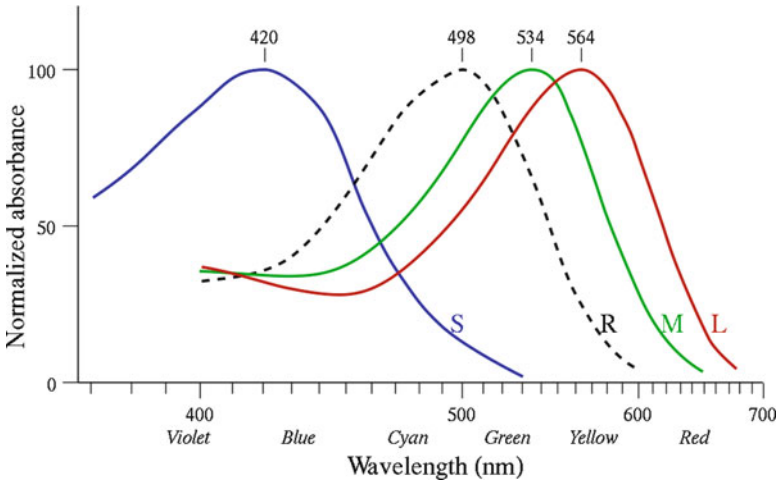


Fig. 14.17 Absorption by the three color pigments in the retina (source: http://en.wikipedia.org/wiki/photoreceptor_cell)

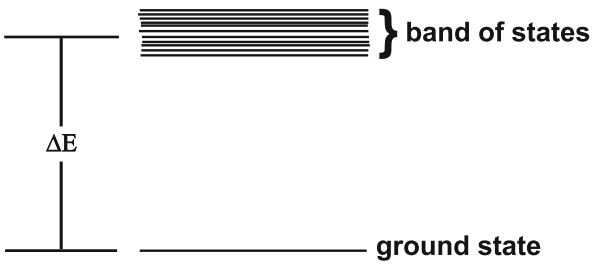


Fig. 14.18 Absorption of a photon by a molecule in a cone

The reason that the absorption peak is so broad in the case of pigment is that the energy level diagram consists of a band of excited states. See Fig. 14.18. There are many excited states that correspond to a large number of transitions. Each transition has its own narrow absorption peak. Since these peaks overlap, the sum total of the peaks produces a single broad peak in the absorption spectrum.

When photons of light impinge upon a pigment molecule that is in its ground state (see Chapter 6, THE ATOM AS A SOURCE OF LIGHT, for a review of this subject), it has the possibility of exciting the molecule into one of the many quantum states that lies in the energy band. Now any system that is excited from its ground state by absorption of photons has the process of photon emission available to it as means of returning to its ground state. However, additional processes are available to it for a return to the ground state.[30]

[30]See a full discussion of this subject in the section "Complex Scenarios of Absorption and Emission" in Chap. 6.

For example, a molecule in a gas can return to its ground state from an excited state by transferring its excitation energy to another molecule during a collision with that molecule. However, in the case of a molecule of cone pigment, in the process of returning to its ground state, excitation energy is used to produce a nerve impulse.

In the figure, ΔE is the difference between the energy of the ground state and the middle of the band of energy levels. It is equal to hf, where h is Planck's constant and f is the frequency ($=c/\lambda$). This frequency corresponds to the wavelength $\lambda \sim 430\,\text{nm}$, where $\lambda = c/f$.

Sample Problem 14-2

Find the photon frequency and ΔE corresponding to $\lambda = 430\,\text{nm}$.

Solution

We have

$$f = \frac{c}{\lambda} = \frac{3 \times 10^8}{430 \times 10^{-9}} = 7.0 \times 10^{14}\,\text{Hz}$$

$$\Delta E = hf = (6.6 \times 10^{-34}\,\text{J-s})(7.0 \times 10^{14}\,\text{Hz}) = 4.6 \times 10^{-19}\,\text{J}. \tag{14.17}$$

The **response of the cone**, in emitting nerve impulses, is proportional to its absorption spectrum. In practical terms:

$$\text{nerve impulse rate of a cone at } \lambda \propto \text{spectral absorption at } \lambda \times \text{intensity } I(\lambda) \text{ incident upon the cone}$$

We are now in a position to present a simplified theory of a physiological basis for accounting for the chromaticity diagram. We introduce the following symbols:

Let

$$N_\text{R}(\lambda) \equiv \text{nerve impulse rate of R-cone.}$$

and

$$S_\text{R}(\lambda) \equiv \text{response function of R-cone for monochromatic } \lambda.$$

The response functions $S_\text{R}(\lambda)$, $S_\text{G}(\lambda)$, and $S_\text{B}(\lambda)$ are proportional to the absorption spectra shown in Fig. 14.17.

Then

$$N_\text{R}(\lambda) = S_\text{R}(\lambda) \times I(\lambda). \tag{14.18}$$

Similarly for G and B we write:

$$N_{\mathrm{G}}(\lambda) = S_{\mathrm{G}}(\lambda) \times I(\lambda)$$
$$N_{\mathrm{B}}(\lambda) = S_{\mathrm{B}}(\lambda) \times I(\lambda). \tag{14.19}$$

Now suppose that we have a general non-monochromatic source with spectral intensity $I(\lambda)$. The spectral intensity tells us what the intensity is for each of the component wavelengths. For each component wavelength, we can compute the numbers $N_{\mathrm{R}}(\lambda)$, $N_{\mathrm{G}}(\lambda)$, and $N_{\mathrm{B}}(\lambda)$. We next add up the set of these numbers for each primary. Thus obtaining the total nerve impulse rate emitted by each of the three cones:

$$N_{\mathrm{R}} = \text{ Sum of } N_{\mathrm{R}}(\lambda)'s$$
$$N_{\mathrm{G}} = \text{ Sum of } N_{\mathrm{G}}(\lambda)'s$$
$$N_{\mathrm{B}} = \text{ Sum of } N_{\mathrm{B}}(\lambda)'s. \tag{14.20}$$

This is analogous to the mathematical process we carry out to obtain the tristimulus values R, G, and B. The response functions replace the color-matching functions.

According to the simple theory of color vision, the signals N_{R}, N_{G}, and N_{B} are analyzable as distinct signals, so that they can be processed and interpreted by the brain to produce the following perceptions:

$$\text{Brightness} \propto N_{\mathrm{R}} + N_{\mathrm{G}} + N_{\mathrm{B}}. \tag{14.21}$$

Hue and saturation are determined by the fractions

$$n_{\mathrm{R}} = \frac{N_{\mathrm{R}}}{N_{\mathrm{R}} + N_{\mathrm{G}} + N_{\mathrm{B}}}$$
$$n_{\mathrm{G}} = \frac{N_{\mathrm{G}}}{N_{\mathrm{R}} + N_{\mathrm{G}} + N_{\mathrm{B}}}$$
$$n_{\mathrm{B}} = \frac{N_{\mathrm{B}}}{N_{\mathrm{R}} + N_{\mathrm{G}} + N_{\mathrm{B}}}. \tag{14.22}$$

These three fractions reflect the color coordinates of the chromaticity diagrams. We do not expect the color coordinates, r, g, and b to be proportional to the respective fractions. However, we do expect that they will increase together: if one increases, so should the other.[31]

[31] In mathematics, we say that r is a **monotonically increasing** function of n_{R}, and so on.

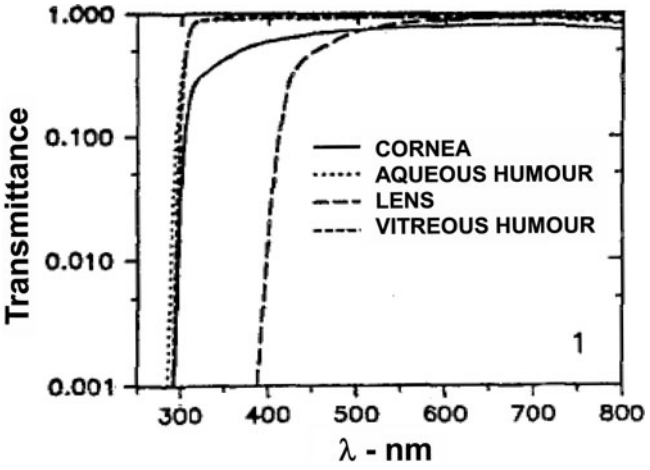


Fig. 14.19 Transmittance of parts of the eye on the way to the retina (note the log scale of the transmittance, which indicates an even sharper drop than is shown) (source: Based upon a figure in the article by W. Ambach, et al., *Documenta Ophthalmologica*, Volume 88, pp. 165–173, !994)

Comment: It is important to keep in mind that testing of color vision involves light that must pass through various components of the eye on its way to the retina. All these components – the cornea, aqueous humour, lens, and vitreous humour – have a transmittance that falls off toward the ultraviolet, low wavelength range. Thus, we have

$$I_{\text{retina}}(\lambda) = T_{\text{eye}}(\lambda) \cdot I_{\text{incident}}(\lambda),$$

where $T_{\text{eye}}(\lambda)$ is the total transmittance of all these parts. The responses of the cones is to the ultimate light incident upon the cones, I_{retina}.

In Fig. 14.19, we see the transmittance of the individual parts.[32]

NOTES: The natural filtering out of UV radiation is beneficial in protecting the retina. With age, filtering increases but unfortunately moves into the visible region. Finally, cataract surgery, which involves removal of the lens, reduces UV filtering, and increases potential damage to the retina. We can see that the lens serves to remove wavelengths from about 300 to 400 nm.

[32]Omitted is the absorption of the macula, which contains the fovea. See the Wikipedia site (1-26-2011): http://en.wikipedia.org/wiki/Macular_degeneration, wherein it is pointed out that while "the macula comprises only 2.1% of the area of the retina . . . almost half of the visual cortex (in the brain) is devoted to processing macular information."

14.13 Color Blindness

We will briefly discuss the simplest kind of color blindness, that of the **dichromats**.[33] Such color-blind individuals are missing one cone. If the green cone is missing, the condition is called **deuteranopia** and the individual is referred to as a **deuteranope**. If the red cone is missing, the condition is called **protanopia** and the individual is referred to as a **protanope**. And finally, for the rarest dichromacy, **tritanopia**, the blue cone is missing. Having only two cones, the dichromat cannot perceive the full range of colors associated with normal vision. They can still perceive the three characteristics – hue, saturation, and brightness; however, they can perceive only two hues.

It is reasonable to wonder how anyone could tell what colors they perceive. How can we know if a person is color blind? How can we compare the mappings of their sensations with the mappings under normal vision? The answer is that there are individuals who have color blindness in one eye and normal vision in the other, referred to as **unilateral dichromats**: These individuals can make a mapping of the vision of their color blind eye onto the vision of their normal eye.

Below are some of its interesting features that are revealed from testing unilateral dichromats:

1. Their chromaticity diagram is reduced to a **line**.
2. There being but two pigments, a point on the line represents the fractional response to light of a single pigment, say the B-pigment.
3. Equal energy white is toward the middle of the line segment. This white sensation can be produced by a single monochromatic source that has a wavelength of about 495 nm!
4. **Protanopes** and **deuteranopes** perceive a line of colors ranging from blue to yellow. See Fig. 14.20. They cannot distinguish among various shades of

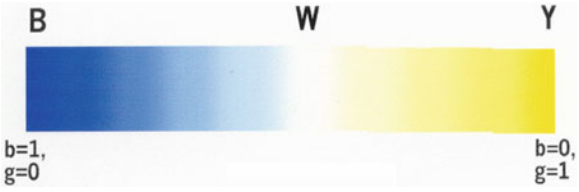


Fig. 14.20 Chromaticity diagram for Deuteranopes – the *Blue-White-Yellow Line*

[33] See the following website for a wonderful resource on color blindness (1-12-2011): http://en.wikipedia.org/wiki/Color_blindness It includes a fascinating set of figures that displays how the rainbow of colors appears for various types of color blindness. It also discusses anomalous dichromacy, wherein there are three cones, but one of them is defective. See also the website (1-12-2011): http://en.wikipedia.org/wiki/Evolution_of_color_vision_in_primates for material on the evolution of color vision in primates through mutation. Most interesting is the article's claim that a remote ancestor of the primates was a **tetranope**, in having four different types of cones.

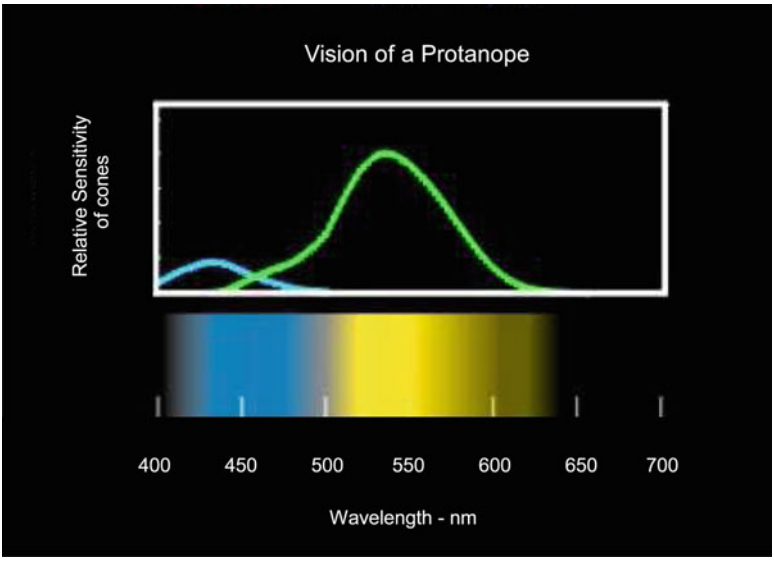


Fig. 14.21 Protanope: spectral sensitivities of the blue and green cones and the image of the rainbow of colors. (source: Dr. Jay Neitz and Dr. Maureen Neitz; http://www.neitzvision.com/content/excuseme.html)

red, yellow, and green.[34] The reader is referred to the references listed at the beginning of the chapter in footnote 1 for further details.

In Fig. 14.21, we see at the top the spectral sensitivities of the blue and green cone of a protanope. At the bottom is shown how a protanope would see a rainbow of colors. Note the gray band close to 500 nm; at the center of the band is a low intensity white.

You can test your own vision using the Ishihara Test for color blindness, shown in Fig. 14.22. The dots in the pattern and in the background are metamers for some observers but not for others. Individuals with normal vision see the number **26**. Deuteranopes see only the number **2**, while protanopes see only the number **6**.

14.14 After-Images

Suppose that we produce a color patch of blue light over a white background in Fig. 14.23. Thus, when the blue light is removed, we will have white light everywhere.

[34] See the websites http://www.neitzvision.com/content/home.html and http://www.handprint.com/HP/WCL/color1.html#dichromat for details.

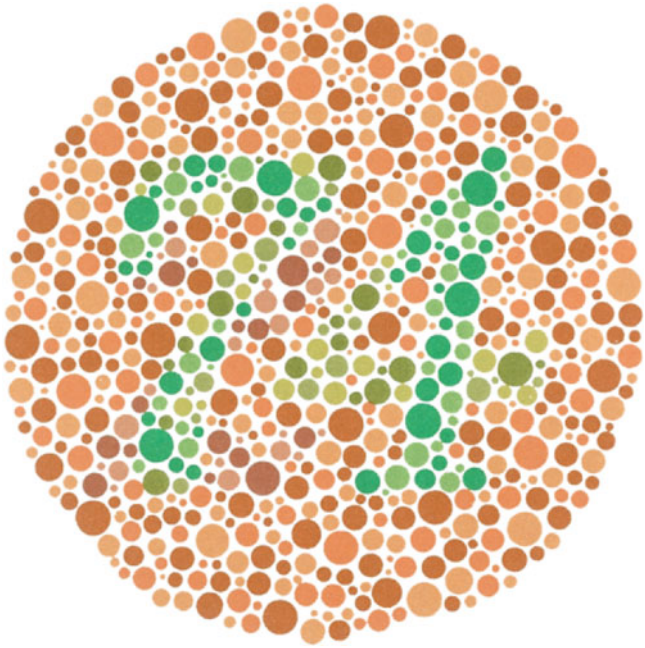
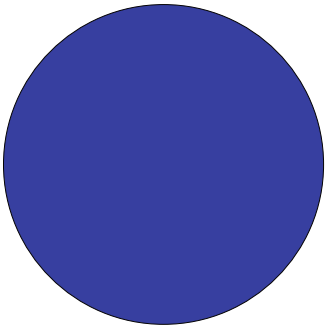


Fig. 14.22 Ishihara test for color blindness (source: http://en.wikipedia.org/wiki/File:Ishihara_9.png)

Fig. 14.23 *Blue* light on a *white* background



If you stare at the patch, without moving your eyes much, for about 30 s and then stare at a blank area, there will appear a patch of yellow light in place of the blue patch. The *yellow* patch is referred to as the **after-image** of the blue patch. Our model of color vision can account for the phenomenon as follows.

When a cone pigment absorbs light and emits nerve signals, the pigment becomes "fatigued" – that is, it has a reduced ability to respond further to light by emitting nerve impulses. Such a process of fatigue also occurs in rod pigment, whose ground state color of purple is bleached (i.e., turned to white). The **recovery time** – that is, the time needed to regain full sensitivity – is only about 1-1/2 min for cones, in contrast to about 25–30 min for rods.

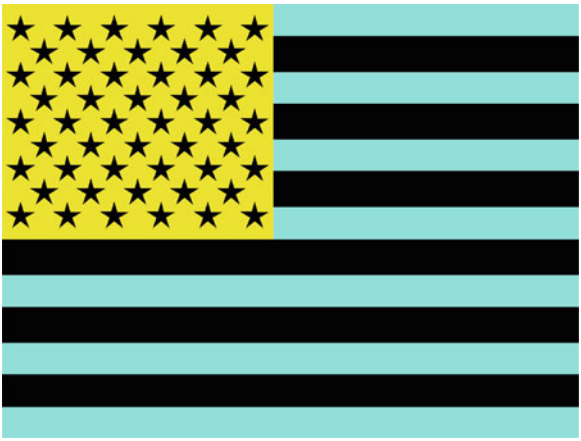


Fig. 14.24 Complement of the US flag

What happens in the above experiment is that the blue patch fatigues the blue receptor cones that lie on the image area of the retina. Upon removing the blue light, that retinal area suddenly *receives* white light. Since the rate of emission of nerve impulses from the blue cones is less than normal, the bluish component of incident white light is reduced, leaving its **complement**, yellow.

Interestingly, if you look carefully at a blue patch, you may notice a yellow halo around its boundary. A detailed explanation appears to be lacking. Nevertheless, this phenomenon is one of many that indicate that there is interference between the nerve impulses emitted by neighboring cones. A related phenomenon is the appearance of Mach bands, which are described and discussed in Chap. 10 in the context of pitch perception of sound. A second source of the halo may be the after-images produced by the erratic, so-called saccadic movements of the eye. (See **R. L. Gregory**, op. cit.)

14.14.1 Questions for Consideration

1. What after-images would result from removing the following patches from a white background?

   (a) A yellow patch?
   (b) A green patch?
   (c) A black patch?

2. Stare at the complementary image of the US flag in Fig. 14.24. Identify the pairs complementary colors. For this exercise, I suggest that you choose a specific star and focus your attention on it. Doing so will help you keep your eyes fixed and improve on the effect of fatigue.

14.15 Terms

- Additive primary
- After-image
- Blue cone
- Chromaticity
- Chromaticity diagram
- Green cone
- Hue
- Metamer
- Primary colors
- Purity
- Color coordinate
- color-matching functions
- Color-blindness
- Complementary colors
- Deuteranope
- Deuteranopia
- Dominant wavelength
- Recovery time
- Red cone
- Saccadic movement
- Saturation
- Subtractive primary
- Tristimulus values
- Unilateral dichromat

14.16 Important Equations

Color coordinates in terms of the intensities of the primaries used to produce the color:

$$r' = \frac{I_\mathrm{R}}{I_\mathrm{R} + I_\mathrm{G} + I_\mathrm{B}}$$
$$g' = \frac{I_\mathrm{G}}{I_\mathrm{R} + I_\mathrm{G} + I_\mathrm{B}}$$
$$b' = \frac{I_\mathrm{B}}{I_\mathrm{R} + I_\mathrm{G} + I_\mathrm{B}}. \tag{14.23}$$

Color coordinates based on the tristimulus values R, G, and B that are calculated from the spectral intensity:

$$r = \frac{R}{S}$$
$$g = \frac{G}{S}$$
$$b = \frac{B}{S}, \tag{14.24}$$

where

$$S = R + G + B. \tag{14.25}$$

Color coordinates for a light source that is a mixture of two incoherent light sources:

$$r = \frac{R_1 + R_2}{S} = \frac{S_1}{S} r_1 + \frac{S_2}{S} r_2$$
$$g = \frac{G_1 + G_2}{S} = \frac{S_1}{S} g_1 + \frac{S_2}{S} g_2, \tag{14.26}$$

where for the first source

$$r_1 = \frac{R_1}{S_1}$$
$$g_1 = \frac{G_1}{S_1}$$
$$b_1 = \frac{B_1}{S_1}$$
$$S_1 = R_1 + G_1 + B_1 \tag{14.27}$$

and similarly for the second source.

Color coordinates for monochromatic light of wavelength λ:

$$r(\lambda) = \frac{\overline{r}(\lambda)}{\overline{r}(\lambda) + \overline{g}(\lambda) + \overline{b}(\lambda)}$$
$$g(\lambda) = \frac{\overline{g}(\lambda)}{\overline{r}(\lambda) + \overline{g}(\lambda) + \overline{b}(\lambda)}. \tag{14.28}$$

14.17 Problems on Chap. 14

1. Discuss briefly how an infant might be taught to distinguish among the colors of objects. How does the infant get to appreciate what the significance of color is as distinct from among the other various attributes that an object can have?
2. (a) The complement of magenta is _____.
   (b) The complement of cyan is _____.
3. The chromaticity diagram shows (choose one)
   (a) The relative response of each type of cone to various wavelengths of light
   (b) A comparison of how rods and cones react to light
   (c) A way of plotting all colors in terms of two variables
   (d) All the colors that are complementary pairs
   (e) Which three colors are the additive primaries
4. (a) What color is complementary to blue?
   (b) What is **subtractive mixing** of colors? What are the primaries of subtractive mixing?

(c) What are **metamers**?
(d) What kind of color mixing do you suppose color television uses?
(e) What are the two attributes of color?
(f) What colors have zero saturation?

5. When we view a magenta light (choose one)

(a) Blue and red cones in the retina are stimulated.
(b) Blue and green cones in the retina are stimulated.
(c) Red and green cones in the retina are stimulated.
(d) Magenta cones in the retina are stimulated.
(e) Since green is the complementary color to magenta, only green cones in the retina are stimulated.

6. A **deuteranope** (choose one)

(a) Sees all objects as shades of red and green.
(b) Cannot distinguish between red and green.
(c) Has lost all red and green vision and sees the world in shades of blue.
(d) Is nearsighted for red light and farsighted for green light, or vice versa.
(e) Reverses reds and greens.

7. There is evidence that some women are **tetrachromats**, meaning that they have four different color cones instead of three. How could you test for **tetrachromacy**?
8. (a) On a schematic chromaticity diagram such as Fig. 14.4, draw the triangle within which lie all colors that have a purple hue. We can refer to these colors as the set of **purple colors**.
HINT: The point W is at one of the corners of the triangle.

(b) Does this definition of purple colors depend upon the choice of primaries?

9. There are animals that have more than three color receptors; that is, they have more than three different cones. Animals with four different cones, such as doves, are called **tetrachromats**. Interestingly, there is evidence that some *women* have four receptors – those, in particular, who have the recessive gene for a dichromat. See the following website for more information:
http://www.freerepublic.com/forum/a3a24199b1ef8.htm

Given what we have discussed about color perception for both **dichromats** and **trichromats**, discuss what changes you might expect for **tetrachromats**. Consider the range of perception of colors, in particular. How many primaries might one need to match any perceived color? How about the ability to discriminate between two spectral intensities? Might one expect improvement?

10. Might two people be found to need two different sets of color coordinates to match a given spectral intensity?
11. In practice, the primaries of various devices, such as color monitors in producing light that leaves the monitor or of color printers that print color images, are not spectral. Furthermore, the actual sets of primaries varies from monitor to monitor and from printer to printer.

(a) Comment on the consequences for the reliable reproduction of color images.
(b) How might we deal with this situation? What we would we have to know about each of the sets of primaries used in the various devices in relation to a standard such as the C.I.E. table of color matching functions?
(c) Name some other devices involving color reproduction that would require analysis in order to determine the relation between the image produced and the input of light.

12. Two monochromatic sources, D and E, are projected onto a screen, so as to appear *equally bright*. A photocell indicates that their intensities are **100 units and 10 units, respectively**. Qualitatively compare the eye *sensitivities* to the *two sources*.
13. Suppose that for a certain set of primaries,

$$0.1\,\mathrm{W/m^2}\text{ of B} \oplus 0.1\,\mathrm{W/m^2}\text{ of G} \oplus 1.0\,\mathrm{W/m^2}\text{ of R}$$

produces a match with W (white). Suppose also, that we choose $1\,\mathrm{W/m^2}$ to be the *unit intensity* for all *three primaries*.

(a) What would be the color coordinates of W?
(b) Why would such a choice *not* be practical?

14. Complete Table 14.4 of color-matching functions so as to determine the color coordinates of butter.
15. Use the schematic chromaticity diagram in Fig. 14.4 for the following problem: Four units of R, two units of G, and two units of B are mixed together.

(a) Find the color coordinates of the mixture.
(b) Find the dominant wavelength and purity of the mixture.
(c) Describe the color.
(d) Find the dominant wavelength of the complementary hue of this chromaticity.

16. Consider a table of color-matching functions such as the one we are using that is based on the spectral primaries $\lambda_B = 435.8\,\mathrm{nm}$, $\lambda_G = 546.1\,\mathrm{nm}$, and $\lambda_R = 700.0\,\mathrm{nm}$.

(a) What would be the effect on the tristimulus values of a given spectral intensity if the values of all of the color-matching functions were doubled?
(b) What would be the effect on the color coordinates of a given spectral intensity if the values of all of the color-matching functions were doubled?
(c) What would be the effect on the color coordinates of a given spectral intensity if the values of all of the color-matching functions were multiplied by any given number beside two?

17. For this problem you will need to open the Java applet on Metamers: metamers_java_jnlp.jnlp. You can find it in the Powerpoint for Chapter 14 or the following website from which the applet was downloaded: http://www.cs.brown.edu/exploratories/freeSoftware/repository/edu/brown/cs/exploratories/applets/spectrum/metamers_java_browser.html.

Note that the graph is a plot of the spectral intensity vs. frequency and not wavelength.

(a) Produce with the mouse two different spectral intensities that are orange in hue and are metamers. Print out the result of your applet and hand it in as part of the homework set.

(b) Recall that in Chap. 8 we learned that the origin of the blueness of the sky is that scattering of sunlight is inversely proportional to the fourth power of the wavelength. As a consequence, the spectral intensity at 400 nm is $(7/4)^4 \sim 9$ times that of the intensity at 700 nm. In the applet, produce with your mouse a spectral intensity that increases from the red end to the blue end by about nine-fold and thus produce the resulting color patch. While we realize that the spectral intensity increases as the fourth power of the frequency, you need not be fussy about the precise shape of the spectral intensity. Print out your result as part of the homework set.

18. Suppose that the two spectral intensities, $I_1(\lambda)$ and $I_2(\lambda)$, are metamers and we add the spectral intensity $I_3(\lambda)$ to each, resulting in two new spectral intensities

$$I_1'(\lambda) = I_1(\lambda) + I_3(\lambda) \quad \text{and} \quad I_2'(\lambda) = I_2(\lambda) + I_3(\lambda). \tag{14.29}$$

Are these two new spectral intensities metamers? To answer this question, consider how the resulting color coordinates would compare. Remember how these numbers are obtained by adding the color-matching functions together.

19. Check that the color-matching functions for the primaries, as shown in Table 14.3, are the inverses of the corresponding unit intensities. For example,

$$\overline{r}(\lambda_R) = \frac{1}{u_R}. \tag{14.30}$$

Show that this must be so for a mixture of unit intensities to produce equal tristimulus values.

20. The goal of this problem is to produce the chromaticity diagram for the primaries corresponding to the color-matching functions in Table 14.4.

> **Note**
>
> You will need this diagram to analyze the results of the Color Lab!!

(a) Use Table 14.4 to determine the color coordinates $r(\lambda)$, $g(\lambda)$ for monochromatic sources of wavelength λ. See (14.6).
HINT: Note that for monochromatic a 400 nm source, $I(\lambda)$ vanishes for all but 400nm. Setting $I_{400\,\mathrm{nm}} = I$, we obtain $R = 0.00030\,I$, $G = -0.00014\,I$, and $B = 0.01214\,I$, from which one can obtain r_{400}, g_{400}, and b_{400}.

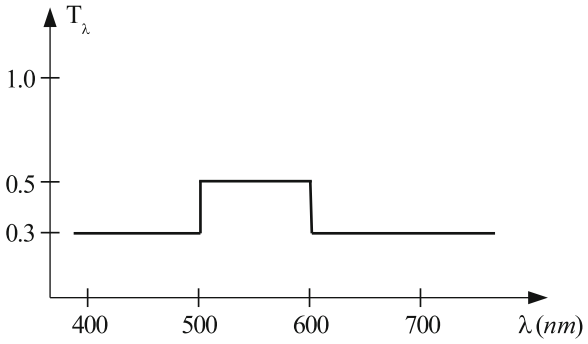


Fig. 14.25 Spectral intensity for Problem 14.21

(b) Plot the coordinates $(r(\lambda), g(\lambda))$ on graph paper and connect the points, thereby obtaining the horseshoe-shaped perimeter of the chromaticity diagram corresponding to Table 14.4. Label the points on the perimeter with the corresponding wavelengths.

21. (a) Determine the tristimulus values and the color coordinates of the spectral intensity depicted in Fig. 14.25.
    (b) Determine the dominant wavelength and the purity using Fig. 14.13 and (14.9), respectively.
22. On a Chromaticity diagram, draw the region within the horseshoe that has hues that must be expressed in terms of their complement.
23. (a) Consider Fig. 14.16, which displays the primaries of sRGB in an lRGB chromaticity diagram. Suppose you wanted to add a fourth primary to these three primaries. We would have a four-primary color set for producing colors. Note that it is perfectly fine to have four primaries as sources of light even though a normal eye has only three different cones. Approximately where would you put the color coordinates so that the gamut of colors that lies within the resulting **quadrangle** encompasses the largest gamut of colors? It should be clear that the color of this fourth primary would not be yellow! The SHARP Corporation claims that in its four-color monitor, it has added a Yellow primary. Perhaps the other three primaries are far from sRGB.
    (b) The added fourth point will define a second triangle. Explain why the additional primary adds only colors that lie within this triangle.
    (c) Explain why for any point within the quadrangular gamut, there is more than one way to produce a color by mixing the four primaries. The factor of 256 to the monitor-count of colors will thus be redundant.
    (d) An additional redundancy is produced by the fact that the density of points in the gamut will be much greater than the density in the three-primary gamut. Give an argument as to why the number of distinguishable colors should be increased by only about 50% at best.
24. Suppose that the spectral intensity of Problem 14.21 is produced by white light passing through a filter. That is, the above spectral intensity is the *transmittance*

Fig. 14.26 Color photograph of Alim Khan – using three black and white filtered photographs (source: http://en.wikipedia.org/wiki/File:Mohammed_Alim_Khan_cropped.png. The original photographer was Sergey Mikhaylovich Prokudin-Gorsky.)

of the filter. (See Sect. 13.2.) In order to increase the saturation of the light beam, two identical such filters are *stacked* (i.e., are placed back to back).

(a) What is the resulting transmittance of the stacked filters? Exhibit your answer as a graph of transmittance vs. wavelength.
(b) Determine the tristimulus values, the color coordinates, the dominant wavelength, and the purity of the spectral intensity. Indicate roughly on a copy of Fig. 14.4 or a rough *sketch* thereof, how the color has been modified by the stacking process. Has the stacking resulted in an increase in the saturation?

25. Suppose you are adding two color patches so as to produce various shades of red, from white to pink. You have one source of red light that is close to being monochromatic and you want to use the *minimum intensity* possible of the second source. Which hue should that second source have? Choose the best answer below. *Explain your answer.*

(a) White
(b) Blue
(c) Cyan

26. Before the advent of color film photography, photographers learned to reproduce color by the following trick: The scene of the photograph was photographed three times in **black and white**, each with a different filter – red, green, and blue. The three negatives were then overlapped so as to produce a single color photograph. We see one of these photographs of Mohammed Alim Khan in Fig. 14.26. Here are the three black and white photographs of filtered light in Fig. 14.27.

Explain how this process works.

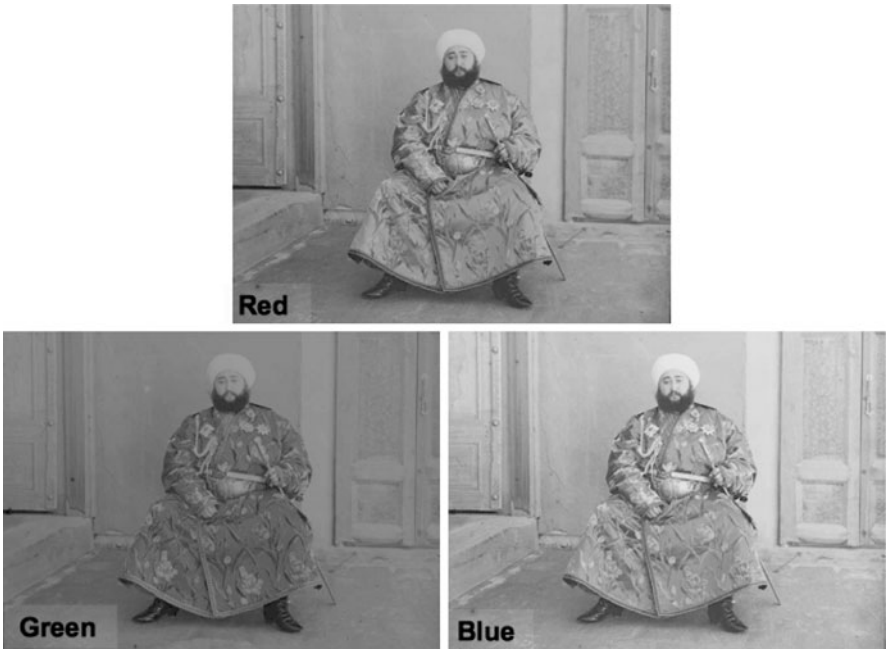


Fig. 14.27 Three *black* and *white* color filtered photographs that produce the color photograph of Alim Khan (source: Anandajoti Bhikkhu)

27. What color is the after-image of a bright green light? Why?
28. In Sect. 14.14.1, it was suggested that you focus on a specific star of the US flag to enhance the effects of fatigue. Comment on why focusing might be of help in producing a clear after-image.
29. In Fig. 14.28, we see two colored discs.

    The upper disc consists of a checkerboard of two colors. The lower disc has a uniform color. If one were to stand far enough away from the figure , the upper disc will appear uniformly colored and indistinguishable from the lower one. The basis for this phenomenon is that the distance between neighboring squares in the image of the upper disc on the retina is comparable to or less than the limit of resolution of the retina. If the cones behave independently of one another, the limit of resolution should be about equal to the distance between neighboring rods. In fact, as we have noted, rods communicate with each other, so that the distance between cones can be expected to be a bit smaller than the limit of resolution.

    Exercise:

    (a) Measure the distance between neighboring squares in the upper disc.
    (b) Look at the figure from a distance and determine the **minimum distance** at which the upper disc has a uniform color.

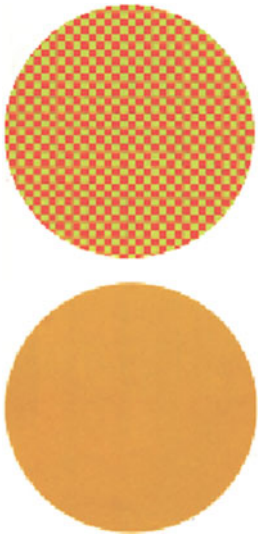


Fig. 14.28 Two discs that appear alike at a great distance (source: Courtesy, Tom N. Cornsweet, *Visual Perception*, (Academic Press, 1970))

(c) Assuming that the distance from your lens to your retina is 2.5 cm, determine the distance between neighboring squares in the image on your retina. Is the value reasonable?

(d) When the upper disc appears uniform in color, it is of interest to compare the intensity of light that a cone receives from a set of squares in comparison to the intensity of light it receives from an equal area of the uniformly colored disc.

Explain why it should be expected that the effective color coordinates should be the respective sums: If we let r_1, g_1 and r_2, g_2 be the color coordinates of the respective squares, then the color coordinates of the color perceived at a distance should be $\{(r_1 + r_2), (g_1 + g_2)\}$.

Start by explaining why the intensity of light incident on the cones is effectively $(I_1 + I_2)/2$.

(e) Use the results of (14.15) in Sect. 14.10 to show that if the color patches match at a distance when viewed with one monitor, they will not necessarily match when viewed with another monitor. The same holds true for color printing.

(f) Finally, consider the question of viewing the figures from a printed copy. If the patches match when viewed with one light source reflected off the page, will there necessarily be a match if you use a different light source having a different spectral intensity?

30. **From spectral intensity to monitor color – the Blue Sky**

The purpose of this problem is to give you experience in producing the color on your monitor that corresponds to a given spectral intensity. You will focus on the spectral intensity of sunlight that is scattered by a clear sky. We learned in

chapter 8 that theory predicts that the spectral intensity is inversely proportional to the fourth power of the wavelength. Thus, we write $I(\lambda) \propto 1/\lambda^4$. The goal is to produce the corresponding color on your monitor, as discussed in Sect. 14.10. You will need to choose a value for gamma. If you cannot determine gamma for your monitor, use the value of 2.2 to complete the problem.

EXCEL is a great tool for carrying out this calculation. You can produce an Excel file of any table in this book by a number of ways. One way is to copy the table as a pdf file. Then you can convert the pdf to an Excel file with various programs. For a MAC, you can use "AnyBizSoft PDF to Excel for Mac." For Windows you can use Nitro's "PDF to Excel Converter." Alternatively, you can download any table from this book from my website: https://wikis.uit.tufts.edu/confluence/display/physics/gunther.

Note that the color on your monitor is likely to have a low brightness compared to what we are used to seeing in the sky. You can resolve this problem by dividing each of the color coordinates {r, g, b} by **b**. You will obtain a new set of numbers – {r/b, g/b, 1} – which you will substitute into (14.16). As a result, we will have $\mathbf{B} = 255$.

Appendix A
Symbols

– $\mathring{A}$ Ångstrom
– α Attenuation per distance in dB per km
– γ Gamma (used to characterize ultra high frequency electromagnetic radiation)
– λ Wavelength
– $\mathcal{T}$ Tension of string
– μ Linear mass density
– ϕ Loudness in phons
– ρ Mass density
– σ Optical activity (angle of rotation of axis of polarization per distance through medium)
– θ Angle
– θ_c Critical angle for the absence of refraction
– θ_{inc} Angle of incidence of a ray of light
– θ_{rfl} Angle of reflection of a ray of light
– θ_{rfr} Angle of refraction of a ray of light
– a Acceleration
– a_0 Bohr radius ($\sim 0.53\mathring{A}$)
– A Area *or* amplitude of a wave or of oscillation
– b Blue color coordinate
– $b(\lambda)$ Blue color coordinate for monochromatic light of wavelength λ
– B Bulk modulus *or* magnetic field *or* blue tristimulus value of a spectral intensity
– B_{ind} Induced magnetic field
– c Speed of light in vacuum
– C Musical interval in cents
– d Distance from a point source *or* distance between two sources of a wave
– dB Decibel
– d_{ie} Image distance for an eye (distance from the center of the effective lens of the eye and the retina)
– d_{im} Diameter of an image as a result of diffraction

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

– d_i Image distance for a lens
– d_{min} Minimum diameter of an image as a result of diffraction
– d_{np} Near point of vision
– d_{o} Object distance for a lens
– E Energy *or* electric field
– EM Electromagnetic
– E_{f} Final energy of a quantum system
– E_{ind} Induced electric field
– E_i Initial energy of a quantum system
– E_{ph} Energy of a photon
– f Frequency *or* focal length of a lens
– f_{B} Beat frequency
– f_n Frequency of the n^{th} mode
– F Force
– g Green color coordinate
– $g(\lambda)$ Green color coordinate for monochromatic light of wavelength λ
– G Green tristimulus value of a spectral intensity
– h Planck's constant ($\sim 4.15 \times 10^{-15} eV/Hz$)
– h_{i} Height of image
– h_{o} Height of object
– I Intensity *or* electric current
– $I(\lambda)$ Spectral intensity
– I_B Intensity of blue primary
– I_G Intensity of green primary
– I_R Intensity of red primary
– k Spring constant
– KE Kinetic energy
– L Horizontal distance from source(s) of a wave
– ℓ A length variable, not specific
– m Mass
– M Magnification of an object by a lens
– max d_{o} Maximum object for the eye
– max f Maximum focal length of the eye
– min f Minimum focal length of the eye
– n Index of refraction
– N_B Signal to the brain from the blue cones
– N_G Signal to the brain from the green cones
– N_R Signal to the brain from the red cones
– p Pressure
– P Power
– PE Potential energy
– q Electric charge
– r Radius *or* red color coordinate
– $r(\lambda)$ Red color coordinate for monochromatic light of wavelength λ
– R Reflectance or red tristimulus value of a spectral intensity

– RT Reverberation time
– s Loudness in sones
– S Sum of tristimulus values ($S = R + G + B$)
– SL Sound level in dB's
– t Time *or* time interval
– T Period *or* transmittance
– $T(\lambda)$ Transmittance of a filter as a function of wavelength
– v Speed or velocity
– V Volume
– x Displacement
– y Displacement of an SHO
– Z Impedance

Appendix B
Powers of Ten: Prefixes

- 10^3 one-thousand **kilo** as in **kilogram** (kg) or kilometer (km) or kilohertz (kHz)
- 10^6 one-million **mega** as in **megahertz** (MHz) (frequency for WGBH FM radio waves is 89.7 MHz)
- 10^9 one-billion **giga** as in **gigahertz** (GHz) (the frequency of microwaves in microwave ovens is 2.5 GHz)
- 10^{-2} one-hundredth **centi** as in **centimeter** (cm)
- 10^{-3} one-thousandth **milli** as in **millimeter** (mm)
- 10^{-6} one-millionth **micro** as in **micrometer** (μm); $1\mu m \equiv 1$**micron** (monochromatic red light has a wavelength of $\sim 0.7\,\mu m$)
- 10^{-9} one-billionth **nano** as in **nanometer** (nm) $= 10$ **Ångstroms** $= 10\,\mathring{A}$ (the size of an atom is typically a few tenths of a nanometer)
- 10^{-12} one-trillionth **pico** as in **picogram** (pg) or **picosecond** (ps) (a bacterium has a mass of about 250 pg; there exist chemicals such that merely 1 pg can be fatal(!); fiber optics signals can be made as short as a ps in duration)

Note that a cube, 1 cm on a side, has a volume of $1\,\text{cm}^3 = (10\,\text{mm})^3 = 1{,}000\,\text{mm}^3$.

Problem: Suppose that an elemental device storing each BIT of a computer hard drive has a volume of $1\,\text{nm}^3$, with the bits stored in a compact way. Suppose, too, that the total volume occupied by the bits is $1\,\text{cm}^3$. How many bits are stored by this hard drive?

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

Appendix C
Conversion of Units and Special Constants

Constants
$\pi = 3.14159...$
$e = 2.7183...$
$c = 2.998... \times 10^8$ m/s speed of light in vacuum
$h = 4.14 \times 10^{-15}$ eV per Hz$= 6.63 \times 10^{-34}$ J per Hz Planck's constant

Length
1 Ångstrom ($\mathring{A}$) $= 10^{-8}$ centimeter (cm)
1 micron (μ) $= 10^{-6}$ meter (m) $= 10^{-4}$ cm
1 cm $= 0.39370$ inch (in)
1 in $= 2.540$ cm
1 foot (ft) $= 30.480$ cm
1 mile (mi) $= 5280$ ft $= 1.61$ kilometer (km)

Time
1 day (d) $= 86{,}400$ s
1 year (yr) $= 3.15 \times 10^7$ s

Speed
1 mph $= 0.448$ m/s
1 m/s $= 2.23$ mph

Area
1 sq-in $= 6.4516$ sq-cm
1 sq-ft $= 929.03$ sq-cm

Volume
1 liter (lit) $= 1{,}000$ cu-cm
1 gallon $= 3.785$ lit

Angle
1 radian (rad) $= 57.3$ degrees (deg)
1 deg $= 60$ minutes
1 minute (min) $= 60$ seconds (s)

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

Force
1 Newton (N) = 0.224 pound (lb)

Weight equivalents (*symbol* $\doteq$) on the Earth's surface
(one unit is for a mass, the other is for a force; as a result, a one kilogram mass will weigh less on the moon.)
1 lb $\doteq$ 454 gram (g)
1 kilogram (kg) $\doteq$ 2.2 lb
1 ounce (oz) $\doteq$ 28.350 gram (g)

Pressure
1 atmosphere (atm) = 1.0×10^5 Pascals (Pa) = 14.7 lb/sq-in

Energy
1 joule (J) = 10,000,000 ergs
1 electron-volt (eV) = 1.6×10^{-19} J
1 calorie (cal) = 4.19 joule
1 Calorie = 1 kilocalorie (kcal) = 1000 cal = 1 food calorie
1 foot-pound (ft-lb) = 1.3549 J
1 British Thermal Unit (Btu) = 252.00 cal = 778 ft-lb

Power
1 horsepower (hp) = 746 watts (W)

Appendix D
References for The Physics of Music and Color

- Backus, J.: The Acoustical Foundations of Music. W.W. Norton, Inc., New York (1977) ML3805.B295
- Benade, A.H.: Foundations of Musical Acoustics. Oxford University Press, New York (1976) ML3804.B328
- Berg, R., Stork, D.: The Physics of Sound. Prentiss-Hall, Englewood Cliffs, N.J. (1982) QC225.15.B47
- Committee on Colorimetry: The Science of Color. Optical Society of America (1970) QC495.06
- Cornsweet, T.N.: Visual Perception. Academic, New York (1970) QP481.C67
- Deutsch, D. Phantom Worlds and Other Curiosities, CD with audio illusions. Philomel Records, Inc, La Jolla, CA, (1995)
- Falk, D., Brill, D., Stork, D.: Seeing the Light. Harper $ Row, Publishers, New York (1986) QC358.F36
- Feynman, R.: Surely You're Joking Mr. Feynman. W.W. Norton, Inc., New York (1985)
- French, A.P.: Vibrations and Waves. M.I.T., Cambridge, MA (1971) QC235.F74
- Gilbert, P.U.P.A., Haeberli, W.: Physics and the Arts. Academic, New York (2008) QC220.G55
- Gregory, R.L.: Eye and Brain, 5th edn. Princeton University Press, Princeton (1997) BF241.G7
- Hall, D.: Musical Acoustics. Brooks/Cole Publishing, Wadsworth, Inc. Belmont, CA (1991) ML3805.H153
- Helmholtz, H.: On the Sensations of Tone. Dover Publications, Mineola, NY (1954) ML3820.H47
- Isacoff, S.: Temperament. Random House, Inc., New York (2003) ML3809 .I83 2001
- Judd, D.B., Wyszecki, G.: Color in Business, Science and Industry. Wiley Series in Pure and Applied Optics, 3rd edn. Wiley-interscience, New York (1975)
- Le Grand, Y.: Light, Color, and Vision. Dover, New York (1957)
- Levitin, D.: This is Your Brain on Music. Penguin Group, New York (2006) ML3830.L38 2006

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

- Loy, G.: Musimathics, volumes 1 and 2. M.I.T., Cambridge, MA (2006) QA19. M87 L69
- Moles, A.: Information Theory and Esthetic Perception. University of Illinois Press, Urbana, IL (1968) BH221.F84 M63
- Moravcsik, M.: Musical Sound of Tones and Tunes. Paragon House Publishers, New York (1987) QC225.15.M65
- Overheim, R., Wagner, D.: Light and Color. Wiley, New York (1982) QC495.088
- John, R.P.: The Science of Musical Sound. Scientific American Library, New York (1983)
- Pirenne, M.H.: Vision and the Eye. Associated Book Publishers, London (1987) QP475.P57
- Poynton, C.: Video and HDTV. Morgan Kaufmann Publishers and Elsevier Science, San Francisco (2003)
- Rigden, J.: Physics and The Sound of Music, 2nd edn. Wiley, New York (1985) ML3805.R56
- Roederer, J.: Int. to Physics and Psychophysics of Music, 4th edn. Springer, New York (2008) ML3805.R74
- Rossing, T.: The Science of Sound. Addison-Wesley, Reading, MA (1982) QC225.15.R67
- Rossing, T., Chiaverina, C.J.: Light Science. Springer, New York (1999) QC358. R68 1999
- Rossing, T.D. et al. (eds.): Springer Handbook of Acoustics. Springer, New York (2007)
- Rossing, T.D.: Resource Letter: MA-2, Musical Acoustics, American Journal of Physics, Volume 55, p. 589 (1987)
- Schwartz, S.: Visual Perception: A Clinical Orientation, 4th edn. McGraw-Hill Medical (2009)
- Scott, D.: The Physics of Vibrations and Waves. Merrill Publishing Company, Columbus, OH (1986) QC225.7 .S48 1998
- Seashore, C.E.: Psychology of Music, 1st edn. Dover Publications, New York (1967) ML3830.S32 P8
- Sethares, W.: Tuning, Timbre, Spectrum, Scale, 2nd edn. Springer, London (2005) QC225.7 .S48 2005
- Taylor, C.A.: Exploring Music: the Science and Technology. Institute of Physics Publishing, Bristol, England (1994) ML3807.T4E9
- Tobias, J.V.: Foundations of Modern Auditory Theory. Academic, New York (1970) QP461.T6
- Tongren, M.C.: Overtone Singing, Revised Second Edition. Fusica, Amsterdam (2004)
- von Bekesy, G.: Experiments in Hearing. McGraw-Hill, New York (1960) QP461.V64
- Walker, J.: The Flying Circus of Physics. Wiley, New York (2006) QC32 .W2
- Waldman, G.: Introduction to Light, Revised edition. Dover Publications, New York (2002) QC355.2.W34

– White, H., White, D.: Physics and Music. Saunders College/Holt, Rhinehart and Winston (1980) ML3805.W44
– Williamson, S., Cummins, H.: Light and Color. Wiley, New York (1983) QC355.2.W56
– Yost, W.: Fundamentals of Hearing, 5th edn. Emerald Group Publishiing, Ltd., Bingley (2006)

Appendix E
A Crude Derivation of the Frequency of a Simple Harmonic Oscillator**

We start with the three Laws of Dynamics that Isaac Newton (1642–1727) proposed to account for the observed motion of the planets about the sun, the motion of the moon about earth, and the motion of projectiles (like bullets or baseballs) just above the earth's surface. To these three laws, he had to add his law for the gravitational force.

Newton's First Law: If an object experiences no net force, its velocity will remain constant, be it zero or otherwise.

Newton's Second Law: If an object does experience a net force, its velocity will change, being reflected by a rate of change of velocity with respect to time – the **acceleration** – that is given by

$$a = \frac{F}{m} \quad \text{or} \quad F = ma. \tag{E.1}$$

Note that acceleration is to velocity as velocity is to position:

$$\text{Velocity} = \frac{\text{Change in position}}{\text{Time interval}}$$

$$\text{Acceleration} = \frac{\text{Change in velocity}}{\text{Time interval}}. \tag{E.2}$$

Newton's Third Law: When an object exerts a force on a second object, the second object automatically must be exerting a force on the first object with a force of equal magnitude but opposite in direction.

Thus, if I am pushing on a wall with a force of 450 N ($\sim$100 lbs), the wall is pushing back on me with a force of 450 N. Likewise, if the mass of an SHO is pulling on the spring with a force F downward, the spring is exerting a force F on the object upward.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

We can now combine Hooke's Law with Newton's second and third laws as follows: Because the force on the massive object of an SHO is opposite to the direction of the displacement, we insert a minus sign and write Hooke's Law as

$$F = -ky \quad \text{Hooke's Law.} \tag{E.3}$$

In this equation, F is the force on the object. It is called the **restoring force** of the SHO because it tends to bring the object back toward the equilibrium position. Since $F = ma$,

$$ma = -ky \quad \text{or} \quad a = -\frac{k}{m}y. \tag{E.4}$$

This equation can be analyzed mathematically. The analysis reveals that once the position and velocity is given at some time, the initial time, the motion is determined forever after. This characteristic of Newtonian dynamics is referred to as **determinism**. In particular, the equation can be shown to lead to the sinusoidal behavior of the SHO.[1]

Now we return to our expression for the period T. During the first quarter cycle, the total displacement is A, while the speed changes from zero to a maximum value v_{m}. The average speed over $1/4$ of a cycle is given by

$$\langle v \rangle = \frac{\text{Displacement}}{\text{Time interval}} = \frac{A}{(T/4)} = \frac{4A}{T}. \tag{E.5}$$

Since the maximum speed is smaller that the average speed, we will use the estimate:

$$v_{\text{m}} \sim 2\langle v \rangle. \tag{E.6}$$

Thus,

$$v_{\text{m}} = \frac{8A}{T}. \tag{E.7}$$

Next, the average acceleration over $1/4$ of a cycle is given by

$$\langle a \rangle = \frac{\text{change of velocity}}{\text{time interval}} = \frac{v_{\text{m}}}{(T/4)} = \frac{4v_{\text{m}}}{T}. \tag{E.8}$$

The maximum acceleration can be obtained from (E.4) by setting $y = A$.

[1]The reader might be interested in carrying out the exercise in the appendix to this chapter, entitled "Numerical Integration of the Equation of Motion of an SHO." In simple terms, the initial displacement y determines the initial acceleration a through (E.4). The initial acceleration determines the change in the velocity v from its initial value and hence its value soon after. The initial velocity determines the change in the displacement and hence the displacement soon after. This cycle is repeated on and on to yield the displacement, velocity, and acceleration for all future times.

We ignore the minus sign since we are interested only in magnitudes. We then obtain

$$a_{\mathrm{m}} = \frac{k}{m} A. \tag{E.9}$$

We estimate that $a_{\mathrm{m}} \sim 2\langle a \rangle$. Thus, we obtain

$$\frac{k}{m} A = a_{\mathrm{m}} \sim 2\langle a \rangle = 8\frac{v_{\mathrm{m}}}{T} = \frac{64A}{T^2}. \tag{E.10}$$

We finally arrive at the approximate relation

$$T^2 \sim 64\frac{m}{k}. \tag{E.11}$$

so that

$$T^2 \sim 64\frac{m}{k} T \sim 8\sqrt{\frac{m}{k}}. \tag{E.12}$$

This result compares very favorably with the exact relation $T = 2\pi\sqrt{m/k}$.

The two expressions differ only by the numerical prefactor, 2π vs. 8, respectively. Their ratio is $2\pi/8 \sim 0.8$. Most important is the agreement between the two expressions with respect to the mass m and the spring constant k.

Appendix F
Numerical Integration of Newton's Equation for an SHO**

This mathematical procedure for analyzing Newton's equation of motion (the Second Law) for an SHO shows us dramatically one of the most important characteristics of classical physics, namely, that nature is such that if we are given all the information about a system at some instant of time, the behavior of the system in the future is fully determined. In the case of the SHO, the initial position x and velocity v of the mass of the SHO determine the future behavior of x and of v. We call this property **determinism**.

We seek to show how the equation

$$a = -\frac{k}{m}y \tag{F.1}$$

generates a sinusoidal function of time. For simplicity, we will set $\mathrm{k/m} = 0.1/\mathrm{s}^2$, so that

$$a = -0.1y, \tag{F.2}$$

where y is in meters and a is in m/s per second. This value of k/m can be shown to correspond to a period of oscillation equal to $2\pi\sqrt{10} \cong 19.9\,\mathrm{s}$.

We will drop the units in what follows. We will assume that initially, when the time $t = 0$,

$$y_0 = 0 \text{ and } v_0 = 1.00 \qquad \text{e.g., } v = 1\,\mathrm{m/s}. \tag{F.3}$$

We have put a subscript 0 next to the letters, so as to refer to $t = 0$.

Now suppose we want to know y and v after 1 s. Recall that

$$\text{Velocity} = \frac{\text{Change of Displacement}}{\text{Time interval}}. \tag{F.4}$$

This will be strictly so only if the velocity is constant. Otherwise, the expression gives us the average velocity over the 1-s time interval.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

Because 1 s is small compared to the period (though not very much smaller), the velocity does not change very much in 1 s. Then the above equation becomes a reasonable first approximation, albeit a crude one.

We then have (with a subscript 1 referring to a time $t = l$ s):

$$v_0 \cong \frac{y(\text{at } 1\text{ s}) - y_0}{1\text{ s}} = y_1 - y_0. \tag{F.5}$$

or

$$y_1 \cong y_0 + v_0 = 0 + 1 = 1. \tag{F.6}$$

Next, we seek the velocity v_1 after 1 s. Recall that

$$\text{Acceleration} = \frac{\text{Change in velocity}}{\text{Change in time}}. \tag{F.7}$$

This will be strictly so if the acceleration is a constant. Otherwise, the expression gives us the *average* acceleration over the 1-s interval. However, if the acceleration does not change much during that interval, we can use (F.7) as an approximation.

The initial acceleration can be expressed approximately as

$$a_0 \cong \frac{v_1 - v_0}{1\text{ s}} = v_1 - v_0, \tag{F.8}$$

so that

$$v_1 \cong v_0 + a_0. \tag{F.9}$$

But

$$a_0 = -\frac{y_0}{10} = 0, \tag{F.10}$$

so that

$$y_2 \cong y_1 + v_1. \tag{F.11}$$

Similarly, after 2-s

$$v_1 = v_0 - \frac{y_0}{10} = 1 - 0 = 1, \tag{F.12}$$

and

$$v_2 = v_1 + a_1 = v_1 - \frac{y_1}{10}. \tag{F.13}$$

After 3-s:

$$y_3 \cong y_2 + v_2, \tag{F.14}$$

and

$$v_3 \cong v_2 - \frac{y_2}{10}. \tag{F.15}$$

etc.

You see how knowledge of y and v at any time allows you to find y and v 1 s later. This process is known as **numerical integration**.

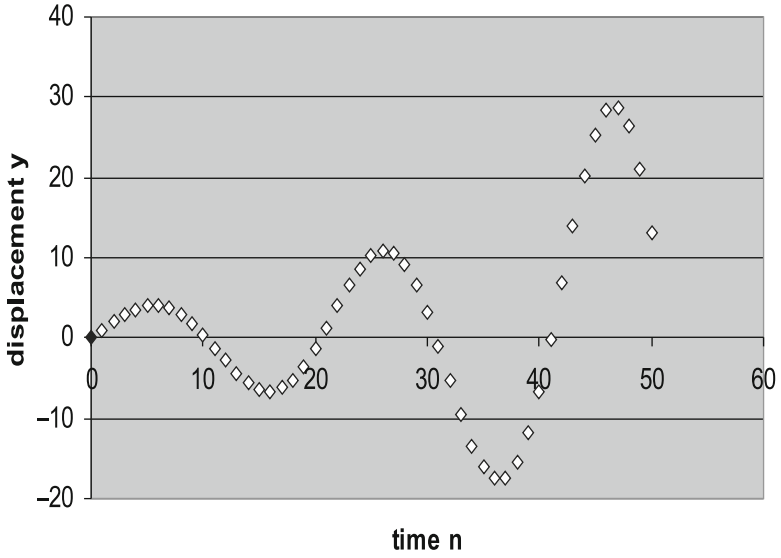


Fig. F.1 Resulting plot of the displacement: a crude numerical integration

For homework:

a. Make a table, listing y and v after 1 s, 2 s, etc., at least up to the time when you obtain $1 - 1/2$ oscillations. Round off all numbers to the nearest hundredth. Prepare your table as follows:

Time (s)	y (m/s)	v (m/s^2)
0	0	1.00
1	1.00	1.00
2	2.00	0.90
3	2.90	?
4	?	?

WARNING: Any error you make is *propagated* on to the numbers which follow. So, be careful.

b. Make a graph of your results for both y vs. t and v vs. t. Compare your results with the plots shown.

Above is a plot of the displacement vs. the time n (Fig. F.1).

Next we plot the velocity vs. time n (Fig. F.2).

c. Note that because of the approximations made, the numerical integration is *unstable*. That is, the displacement y oscillates with ever increasing amplitude. The displacement ultimately diverges to infinity.

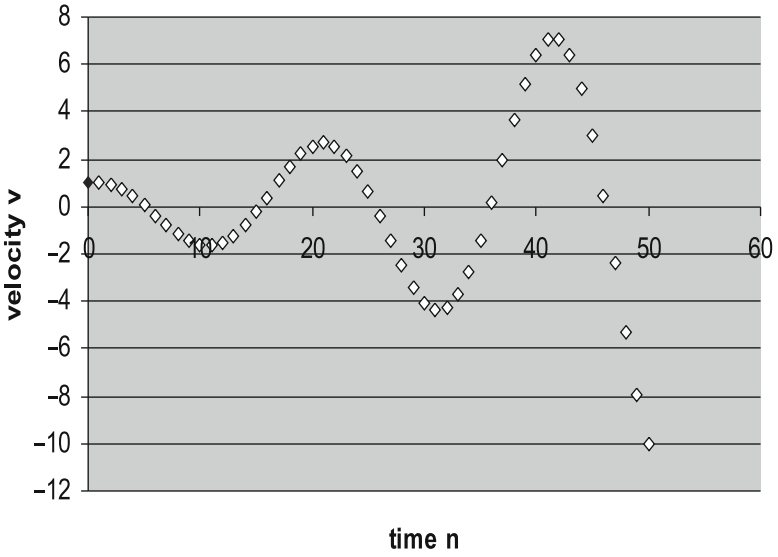


Fig. F.2 Resulting plot of the velocity: a crude numerical integration

An Improved Approximation

Let n be the time in seconds. That is, $n = 1$ refers to 1 s, $n = 2$ refers to 2 s, etc. Our previous approximation can be expressed as

$$y_n \cong y_{n-1} + v_{n-1} \tag{F.16}$$

and

$$v_n \cong v_{n-1} + a_{n-1} \tag{F.17}$$

along with

$$a \cong -\frac{y_n}{10}. \tag{F.18}$$

This last equation is exact.

We obtain a *much* better approximation if we use the following approximation, which replaces (F.16) and (F.17):

$$y_n \cong y_{n-1} + \frac{v_{n-1} + v_n}{2} \tag{F.19}$$

and

$$v_n \cong v_{n-1} + \frac{a_{n-1} + a_n}{2}. \tag{F.20}$$

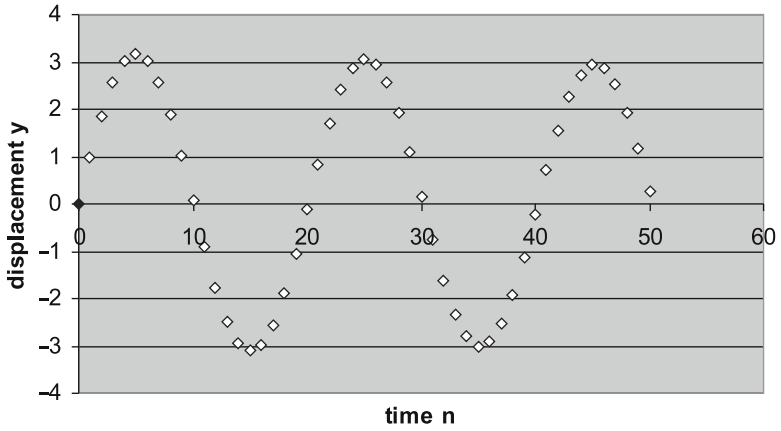


Fig. F.3 Resulting plot of the displacement: an improved numerical integration

Question: What do you suppose is the basis for these equations?

These equations, along with $a_n = -y_n/10$, may be solved for y_n and v_n. They lead to:

$$y_n \cong 0.95y_{n-1} + 0.98v_{n-1} \tag{F.21}$$

and

$$v_n \cong 0.95v_{n-1} - 0.096y_{n-1}. \tag{F.22}$$

Tabulate and graph these equations. Use the initial conditions

$$y_0 = 0 \qquad \text{and} \qquad v_0 = 1.00.$$

Just to show you how the above equations work. We have:

$$y_1 = 0.95y_0 + 0.98v_0 = 0 + 0.98 = 0.98 \tag{F.23}$$

$$v_1 = 0.95v_0 - 0.096y_0 = 0.95 - 0 = 0.95 \tag{F.24}$$

$$y_2 = 0.95y_1 + 0.98v_1 = 0.95(0.98) + 0.98(0.95) = \ldots \tag{F.25}$$

$$v_2 = 0.95v_1 - 0.096y_1 = \ldots. \tag{F.26}$$

Compare your results with the plots below. Then **repeat** the above with the initial values: $y_0 = 1.00$ and $v_0 = 0.0$.

Below are plots of the displacement vs. the time n and of the velocity vs. time n. (Figs. F.3 and F.4).

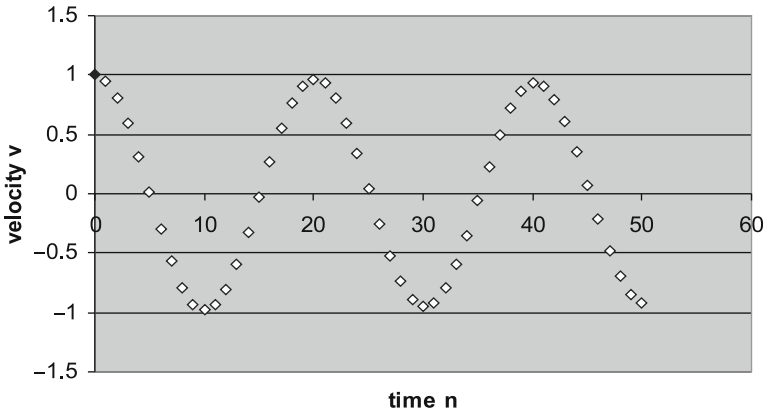


Fig. F.4 Resulting plot of the velocity: an improved numerical integration

Problems with Numerical Integration and Oscillators

1. In this problem, you will check the independence of the frequency of an SHO with amplitude. To do so, use the initial conditions $y_0 = 0$ and $v_0 = 4.00$. Use the refined set of equations to produce a plot of the displacement vs. time and compare.
2. In this problem you will study an oscillator whose restoring force is not Hooke's Law. Instead, let the force be proportional to the cube of the displacement. Thus, the acceleration is now given by

$$a = -0.1y^3. \tag{F.27}$$

Find the displacement vs. time for the two initial conditions used for the SHO.

For both Hooke's Law (linear restoring force) and the cubic restoring force, the force increases with increasing displacement. However, the cubic restoring force increases faster with increasing displacement. This leads to a suggestion: Before obtaining your graphs, try to predict whether the frequency for a "cubic oscillator" should increase or decrease with increasing amplitude.

Appendix G
Magnifying Power of an Optical System**

In Chap. 8, we introduced the parameter **magnification** of a lens as the ratio of the optical image size of a lens to the object size. If we have a compound lens system, such as a telescope or a microscope, the magnification would be the ratio of the final optical image size to the input object size. On the other hand, the ultimate purpose of an optical instrument such as a magnifying glass (a single converging lens), a telescope, or a microscope is to increase the size of the image of the object on the retina over the size of the image on the retina in the absence of the instrument.[1] The maximum possible ratio for an optical instrument is referred to as the **magnifying power**, to which we will assign the symbol $\mathcal{M}$:

$$\mathcal{M} = \frac{\text{Image on the retina with the instrument}}{\text{Image on the retina in the absence of the instrument}}. \tag{G.1}$$

By increasing the image size on the retina, we have two effects:

(1) The object appears to be larger.
(2) The details of the object are clearer: there is increased **resolution**.

We can increase the magnifying power by bringing the eye closer to what we will refer to as the **ultimate eye object** for the eye as a compound lens. Without the instrument in place, the *ultimate eye object* is the object itself. With the instrument in place, the *ultimate eye object* for the eye is the final image produced by the instrument.

We can maximize the image on the retina, with or without the instrument in place by having the *ultimate eye object* be at the **near point of the eye**.[2] This is the closest that an object can be to the eye and still be in focus. For the so-called normal eye, this distance is about 25 cm and this value is usually used as a standard in evaluating optical instruments.

[1] Note that while a lens might produce a large optical image, this image might be so far away from the eye as to lead to a smaller image on the retina!

[2] See Chap. 12 for details.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

Another way to increase magnifying power, that is accessible for a magnifying glass and a microscope, but *not* typically for a telescope, is to vary the distance of the object from the eye.

G.1 Image with the Naked Eye and with a Magnifying Glass

In Fig. G.1, we see the image at R produced on the retina by a specific object at O of height h_o when the object is at the near point d_{np}. The small dot with a label C is the center of the eye-lens system. The angle θ is referred to as the **angle subtended by the center of the lens of the object**. Clearly, this angle is a direct measure of the height of the image on the retina, since the distance d_{ie} is fixed. We also exhibit the central ray (dashed) from the object to the retina when the object is further from the eye than the near point; we see clearly that the image height h_{ie} is reduced.

If we try to increase the image on the retina by moving the object closer to the eye, we have the problem of not being able to bring the image into focus. However, by inserting a converging lens between the object and the eye, we can maintain the angle θ and therefore the desired image height and yet be able to bring the image into focus by having the image produced by this lens be located at the near point. This image becomes the *ultimate object* for the eye-lens system. See Fig. G.2.

In Fig. G.3, we see the magnifying glass up against the eyeball. We also see the image of the magnifying glass, which becomes the **ultimate eye object** (in black), placed at the position of the near point. For comparison, we see the actual object repositioned in blue at the near point, where it would have to be placed in the absence of the magnifying glass to be seen clearly. Thus, we can appreciate the ability of the magnifying glass to "magnify" in the applied sense. Essentially, the magnifying has enlarged the ultimate eye object located at the near point from h_o to h_i.

The magnifying power of the magnifying glass is given by

$$\mathcal{M} = \frac{h_o/d_o}{h_o/d_{np}} = \frac{d_{np}}{d_o} \tag{G.2}$$

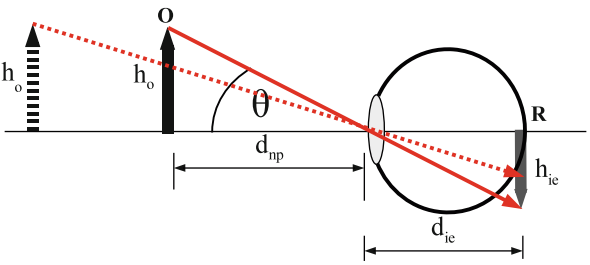


Fig. G.1 The image on the retina for various object distances with the naked eye

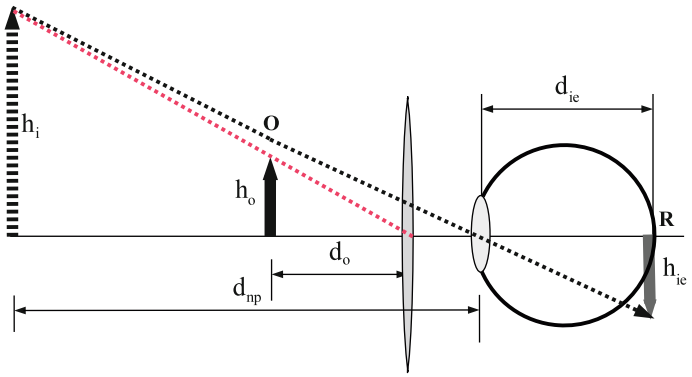


Fig. G.2 Magnification of the potential image on the retina by a lens

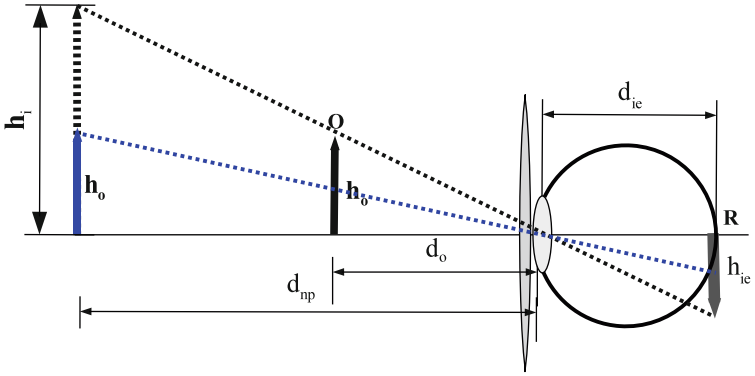


Fig. G.3 Magnifying glass up against the eyeball

Sample Problem 7-1

1. Show that the magnifying power of a magnifying glass that is held up against the eye is given by

$$\mathcal{M} = \frac{d_{\text{np}}}{f} + 1, \tag{G.3}$$

where f is the focal length of the magnifying glass.

Solution

We use the thin lens equation to obtain an expression for the object distance d_{o} in terms of f and the near point distance d_{np}. Noting that

$$d_{\text{i}} = -d_{\text{np}}, \tag{G.4}$$

we obtain

$$\frac{1}{d_o} = \frac{1}{f} - \frac{1}{d_i} = \frac{1}{f} + \frac{1}{d_{np}}. \tag{G.5}$$

Substituting into (G.2) we obtain

$$\mathcal{M} = \frac{d_{np}}{d_o} = d_{np}\left[\frac{1}{f} + \frac{1}{d_{np}}\right] = \frac{d_{np}}{f} + 1. \tag{G.6}$$

1. Show that the magnifying power is given by

$$\mathcal{M} = \frac{h_i}{h_o}. \tag{G.7}$$

Solution

left to the reader.

2. (a) Assuming that $d_{np} = 25\,\text{cm}$, calculate the magnifying power for two values of the focal length: 25 cm.

Solution

From (G.6) we have

$$\mathcal{M} = \frac{d_{np}}{f} + 1 = \frac{25}{25} + 1 = 2. \tag{G.8}$$

(b) Determine the focal length necessary to produce a magnifying power of 40×.

Solution

We have

$$\mathcal{M} = 40 = \frac{d_{np}}{f} + 1 = \frac{25}{f} + 1. \tag{G.9}$$

Then

$$\frac{25}{f} = 40 - 1 = 39 \tag{G.10}$$

from which we obtain

$$f = \frac{25}{39} = 0.64\,\text{cm}. \tag{G.11}$$

G.2 The Microscope

While the purpose of a telescope is to produce magnification of a typically huge object that is extremely far away, the purpose of a microscope is to produce magnification of an extremely small object that can be brought extremely close to our eyes. For both devices there are two lens systems, with the final lens serving as a magnifying glass.

Figure G.4 is a schematic of a microscope. The first lens of a microscope is referred to as the **objective**. It has an extremely small focal length f_{ob} that allows us to bring the lens very close to the **microscopic** object. The second lens, with a focal length f_{e}, is the **eyepiece** or **ocular**, situated close to our eye.

We see that the objective of the microscope produces a real image that is situated just within the focal length of the **eyepiece**. We have magnification by the objective: This first image is much larger than the object. Next, the eyepiece serves as a magnifying glass of the first image, producing a virtual image that is the ultimate "object" of the eye itself. One can obtain maximum magnification by having the ultimate object located at the near point of the eye, as shown in the figure.

The overall magnification of the microscope is the **product** of the magnification of the objective and the magnifying power of the eyepiece:

$$\mathcal{M}_{\text{microscope}} = \frac{d_{\text{i}}}{d_0}\left(\frac{d_{\text{np}}}{f_{\text{e}}} + 1\right). \tag{G.12}$$

Since $d_{\text{i}} \gg f_{\text{ob}}$, according to the thin lens formula,

$$\frac{1}{d_{\text{o}}} = \frac{1}{f_{\text{ob}}} - \frac{1}{d_{\text{i}}} \approx \frac{1}{f_{\text{ob}}}. \tag{G.13}$$

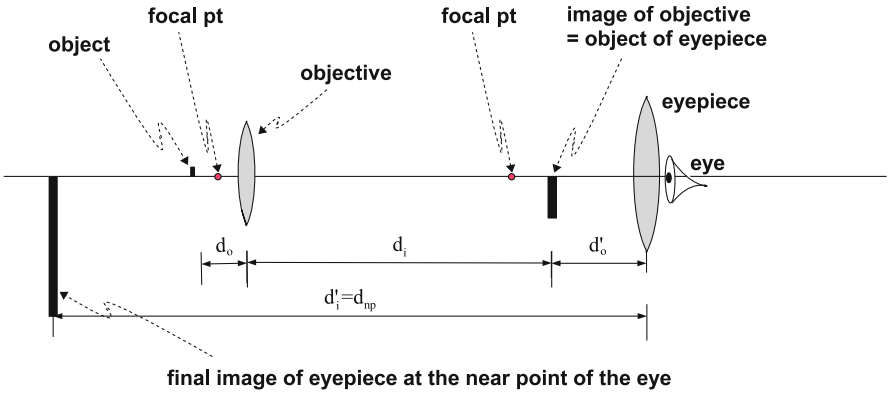


Fig. G.4 Schematic of a microscope

Thus, $d_0 \approx f_{\text{ob}}$ and we can rewrite the above magnification as

$$\mathcal{M}_{\text{microscope}} \approx \left(\frac{d_{\text{np}}}{f_{\text{e}}} + 1 \right) \frac{d_{\text{i}}}{f_{\text{ob}}}. \tag{G.14}$$

Unfortunately, there is a problem that we have to confront: People do not all have the same near point! Therefore, with the above design, the location of the image of the objective would have to vary from individual to individual. How can we design a microscope that takes this fact into account?

The answer is that we can change the design of a microscope so that the image of the objective is at the focal point of the eyepiece. In this case the final image moves to infinity, which is visible to any "reasonable eye." In this case, it can be shown that the magnifying power of the eyepiece changes:

$$\mathcal{M} = \left(\frac{d_{\text{np}}}{f_{\text{e}}} + 1 \right) \Rightarrow \frac{d_{\text{np}}}{f_{\text{e}}}. \tag{G.15}$$

Since typically d_{np} is only somewhat greater than f_{e}, this change is not too great. The overall magnification of the microscope will then be

$$\mathcal{M}_{\text{microscope}} \approx \frac{d_{\text{np}}}{f_{\text{e}}} \frac{d_{\text{i}}}{f_{\text{ob}}}. \tag{G.16}$$

G.3 Problems on Magnifying Power

1. Assuming that $d_{\text{np}} = 25\,\text{cm}$, calculate the magnifying power for two values of the focal length: 2.5 cm.
2. The magnifying power is defined as a standard in terms of the near point of 25 cm. However, in practice, the near point depends upon the individual.

   Suppose that a magnifying glass is labeled as having a magnifying power of 10×. The author of this book has settled down to a stable near point of 59 cm.

   Calculate the effective magnifying power for my eyes by first solving (G.6) for the focal length of the magnifying glass.

Appendix H
Threshold of Hearing, Threshold of Aural Pain, and General Threshold of Physical Pain**

If we look carefully at the set of equal loudness curves in Chap. 10, we find two curves that bound the curves from above and below. The curve at the top is referred to as the **threshold of hearing**; the curve at the bottom is referred to as the **threshold of pain**. The former represents the minimum intensity of a pure tone that can be heard as a function of frequency. The latter corresponds to intensities that produce a sense of pain as opposed to sound. We will refer to this latter curve as the **threshold of *aural* pain** so as to distinguish it from a threshold of a more general sort of physical pain that one might experience in a body structure such as one's leg or back and normally find annoying. This curve is therefore a boundary between two types of sensation.[1] See Fig. H.1.

We will begin our discussion by characterizing a bit the threshold of hearing in physical terms. We have pointed out in Chap. 3 that sound corresponds to a variation of pressure in a medium. In the case of air, this pressure is produced by a huge rate of collisions of molecules of air against a surface. At a pressure of one atmosphere, there are about one trillion–trillion $(1,000,000,000,000,000,000,000,000)$ collisions each second on an eardrum, which has a surface area of about $1\,\text{cm}^2$. The sound that we hear reflects a **difference** in the forces on the two sides of an eardrum.

In Fig. H.2, we see the results of a computer simulation of the pressure fluctuations. To the left we see the positions of an ensemble of dots representing molecules within a box. To the right is plot vs. time of the pressure produced by the collisions of the molecules with the surface of the walls of the box.

In Fig. H.3, we depict the force on a sub-microscopic area of an ear drum due to these collisions over a short interval of time. Each spike represents a collision of a single molecule in the air, lasting about one-trillionth of a second.[2] In Fig. H.3a we

[1] The closest analogous threshold for a large body part that I can think of the boundary between a tickle and an ache.

[2] In order to get a feeling as to what the number one trillionth is, suppose we were to cover an area the size of a football field with the dots over the letter "i" in this print. There would be about one-trillion such dots. Therefore, each dot takes up one-trillionth of the area of a football field.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

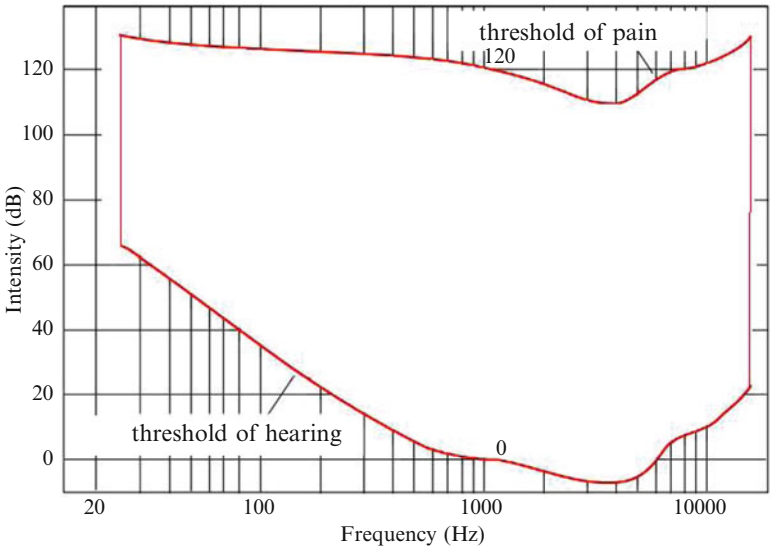


Fig. H.1 Hearing thresholds enclosing the set of equal loudness curves

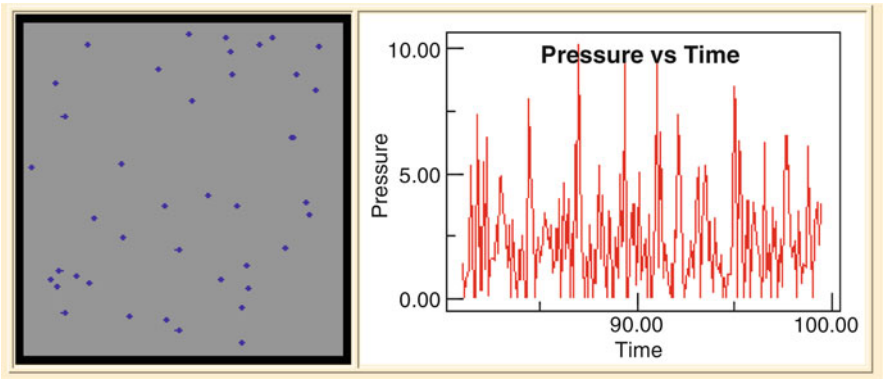


Fig. H.2 Computer simulation of the fluctuations of the pressure on a surface (source: http://webphysics.davidson.edu/applets/Molecular/Pressure.html)

see the force on the outside of the eardrum, while in Fig. H.3b we see the force on the inside of the eardrum. The latter force is shown downward in a negative direction to represent the fact that the above two forces have opposite directions.

In Fig. H.3c we see the two forces together on one graph. Because each collision is so minuscule in its effect and because this rate over the entire eardrum is so huge, we experience a force that is uniform over the surface of the eardrum and is extremely steady over time. To appreciate this fact, imagine a short time interval of 100 trillionths of a second. Imagine that there were 100 collisions during this

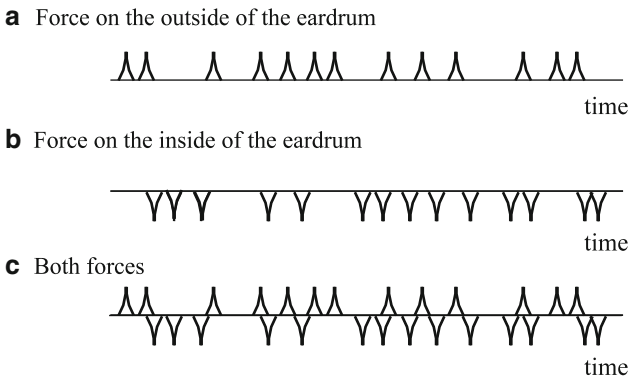


Fig. H.3 Force due collisions of individual molecules on an eardrum

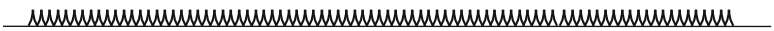


Fig. H.4 A segment of adjacent collisions

time interval and that they were spread uniformly over this time interval. The spikes would then touch each other, as shown in Fig. H.4.

In fact, during an interval of 100 trillionths of a second there are 100 trillion collisions, not merely 100. Imagine how dense the spikes actually are! It is then easy to appreciate why the force is extremely close to being constant in time.

Because of the discreteness of the collisions, the force produced is not exactly zero or constant in time. In Fig. H.3, representing the collisions over an extremely minute area of an eardrum, we see that in that short time interval there are 13 collisions on the outside along with 14 collisions on the inside, making a difference of one collision. The situation for the force on an entire eardrum is different. Over a period of about 1 s, we might have a total of about one trillion–trillion collisions on each side of the eardrum. The **difference** in the number of collisions will be on the order of one part in one trillion – that is the still huge number of one trillion collisions! The resultant variation in the overall pressure is referred to as **pressure fluctuations**.[3] As a result, even when there is no sound wave present, the forces on the two sides of an eardrum do not cancel each other. How does this net force compare to the force that is necessary to produce an audible sound?

In principle, with sufficient aural sensitivity, we should well wonder whether we can hear the individual collisions of molecules! In fact, the sensing apparatus for hearing is designed so that we cannot hear these collisions. Their presence is in

[3]For comparison sake, imagine rain drops colliding with a window pane and pitter patter sound they produce. Now imagine what would happen if the density of raindrops were to increase greatly and their rate of collision increases greatly. Ultimately, we would describe the sound produced as a steady continuous sound.

fact mirrored in the fluctuations of nerve impulses mentioned in Chap. 10. We hear sounds that produce nerve impulses that are over and above these fluctuations.

Still, we would like to get back to the question as to how the fluctuations compare to the sounds that are at the threshold of hearing. It is not too difficult to calculate the sound that is produced as a result of the variation of the force on the eardrums as a result of the discreteness of the collisions. There is a frequency spectrum to this force that ranges from zero frequency to a frequency corresponding to the duration of a single collision (∼one-trillionth of a second) – therefore, to a frequency of about one-trillion Hz. The sound is a uniform mixture of frequencies (as in the case of white light) and is often referred to as **white noise**. It is not straightforward, if at all possible to compare the resulting intensity with the intensities corresponding to the threshold of hearing among the equal loudness curves since these curves correspond to sounds having single frequencies. Ideally, one should test people for their threshold of hearing white noise. The best I can think of doing is to calculate the total intensity of the white noise over a range of audible frequencies – say in the most audible range of 1,000–3,000 Hz. The result is a total intensity of about $10^{-12}\,\mathrm{W/m^2}$, equal to the threshold of hearing at 2,000 Hz. Thus, it is likely that the sensitivity of the ear is as small as possible, being close to being able to hear the collisions of individual molecules of air on the ear drums! The *inaudible* sound due to the randomness of the collisions is referred to as **background noise**. Our entire nervous system is wired up so that normally we do not sense the background noise that tends to excite our nerves.[4]

What is the pressure at the threshold of hearing? It is a minimum at a frequency to which we are most sensitive: about one-ten-billionth of an atmosphere at a frequency of about 3,000 Hz. **Georg von Békésy**, in his study of cats' ears, found that at the threshold, the eardrum has an amplitude of about one-tenth of an Ångstrom, a size that corresponds to about one-tenth the diameter of a single molecule!

What about pain in general? Recently, I visited a physiatrist to deal with back pain that had developed over the past few months – another one of my many episodes. The main issues were whether the source of the pain was due to a relatively simple problem of sprained ligaments or strained muscles or a more serious problem of a herniated disc of the spine, or an even more serious disease such as cancer. The fact was that I had no recollection of having had any incident that might have done damage to my back.

My visit to his office informed me of a fascinating phenomenon of the human body. The brain can reset the level of background noise that can produce a sense of pain. My physiatrist's diagnosis to my backache was simple: I am getting old! My discs are wearing down. Old MRIs of my back were evidence that my level of deterioration might well be belated in life since they revealed at my then current age of 55 the back of a 40 year old. What should be my response? I will paraphrase his

[4]The same situation holds for vision. The eye is capable of detecting the incidence of only about 100 photons over a period of a few seconds. Any greater sensitivity could lead to problems due to noise within the nervous system for vision.

response. “Simple,” he said. “Stop worrying about your back. People are fortunate to have a unique organ in the brain called the **amygdala**. One of its functions is to control the response in your brain to nerve impulses that can generate sensations of pain. When we have a steady or continual input of such impulses, the amygdala can change the brain’s response to these sources of pain by causing the nervous system to treat them as our new level of background noise! And, if we are fully fortunate, we will eventually not sense any pain. If you dwell upon the pain, you will weaken the amygdala’s ability to perform this function and therefore prolong your pain. So ignore the pain and move on with your life.”

Appendix I
Transformation Between Tables of Color-Matching Functions for Two Sets of Monochromatic Primaries**

In Chap. 14, we provided a Table 14.2 of color-matching functions – henceforth referred to as **TCMF** – that was produced by Judd and Wyszecki by studying the color vision of a set of individuals with normal color vision for a specific set of primaries, $\lambda_B = 435.8\,\text{nm}$, $\lambda_G = 546.1\,\text{nm}$, and $\lambda_R = 700.0\,\text{nm}$. The table consists of three columns of numbers, $\overline{r}(\lambda)$, $\overline{g}(\lambda)$, and $\overline{b}(\lambda)$ – the **color matching functions**. What if we have a different set of monochromatic primaries? How should we mix these primaries so as to produce the same colors? In this appendix, we will derive a set of nine numbers that will allow us to determine a corresponding TCMF – starting with the original TCMF – that should allow the same set of individuals to match a spectral intensity using any other set of monochromatic primaries. We will label their wavelengths as $\{\lambda'_R, \lambda'_G, \text{and } \lambda'_B\}$.

In the TCMF, there are 16 different wavelengths for each of the three primaries.[1] One might assume that each of the set of $3 \times 16 = 48$ numbers of the new TCMF depends upon the entire set of 48 numbers of the original TCMF. We would then need $48 \times 48 = 2{,}304$ numbers to specify the relationship between the two TCMFs. We will see shortly that in fact, only $3 \times 3 = 9$ numbers are sufficient to determine the relationship.[2] The nine numbers are exhibited in the following three equations:

$$\begin{aligned}
\overline{r'}(\lambda) &= U_{RR}\overline{r}(\lambda) + U_{GR}\overline{g}(\lambda) + U_{BR}\overline{b}(\lambda) \\
\overline{g'}(\lambda) &= U_{RG}\overline{r}(\lambda) + U_{GG}\overline{g}(\lambda) + U_{BG}\overline{b}(\lambda) \\
\overline{b'}(\lambda) &= U_{RB}\overline{r}(\lambda) + U_{GB}\overline{g}(\lambda) + U_{BB}\overline{b}(\lambda).
\end{aligned} \tag{I.1}$$

[1] We realize that there are an infinite number of possible wavelengths spanning the range 400–700 nm. Our table essentially samples a given spectral intensity at a discrete values simply for convenience. In fact, our table was taken from another that had double the number of wavelengths, with values halfway between the values in our table. For ideal sampling, we would need color-matching functions for the continuum of visible wavelengths. The tristimulus values would be integrals: $R = \int \overline{r}(\lambda) I(\lambda) \mathrm{d}\lambda$, $G = \int \overline{g}(\lambda) I(\lambda) \mathrm{d}\lambda$, and $B = \int \overline{b}(\lambda) I(\lambda) \mathrm{d}\lambda$.

[2] In fact, if we do not care about maintaining a specific requirement on the unit intensities, we need only eight numbers.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

The nine numbers are represented by the symbols U_{RR}, U_{GR}, U_{BR}, U_{RG},U_{GG}, U_{BG}, U_{RB}, U_{GB}, and U_{BB}. They are commonly exhibited as an array of numbers, together forming what is referred to as a **matrix**. A common symbol for a matrix is $\mathbb{U}$, here shown for the letter U. The matrix $\mathbb{U}$ is exhibited below.[3]

$$\begin{pmatrix} U_{RR} & U_{GR} & U_{BR} \\ U_{RG} & U_{GG} & U_{BG} \\ U_{RB} & U_{GB} & U_{BB} \end{pmatrix}.$$

The essential reason for this extraordinary simplification is *physiological*: color vision is based on a *single* set of three receptors with three corresponding independent nerve impulse rates, as opposed to such a set for each possible set of primaries. This fact will be demonstrated in the last subsection of the appendix.

Explicit Expression for the Transformation Matrix $\mathbb{U}$

We will later show that the transformation matrix $\mathbb{U}$ can be expressed in terms of a **sub-matrix** obtained from the TCMF.

Suppose that the wavelengths of the second set of primaries is given by $\lambda_{R'}$, $\lambda_{G'}$, and $\lambda_{B'}$.

Then this sub-matrix, here symbolized by $\mathbb{V}$, is given by:

$$\begin{pmatrix} V_{RR} & V_{RG} & V_{RB} \\ V_{GR} & V_{GG} & V_{GB} \\ V_{BR} & V_{BG} & V_{BB} \end{pmatrix} \equiv \begin{pmatrix} \overline{r}(\lambda_{R'}) & \overline{g}(\lambda_{R'}) & \overline{b}(\lambda_{R'}) \\ \overline{r}(\lambda_{G'}) & \overline{g}(\lambda_{G'}) & \overline{b}(\lambda_{G'}) \\ \overline{r}(\lambda_{B'}) & \overline{g}(\lambda_{B'}) & \overline{b}(\lambda_{B'}) \end{pmatrix}.$$

Note, for example, that $V_{RG} = \overline{g}(\lambda_{R'})$.

We will show that

$$U_{\alpha\beta} = V_{\alpha\beta}^{-1}/u'_{\beta}. \tag{I.4}$$

---

[3]For those who are familiar with matrices, we will rewrite (I.1) in a simpler form. We introduce two vector functions, (λ) and $'(\lambda)$. They represent the color-matching functions for the respective two sets of primaries have the following components:

$$\begin{aligned} &\overline{c}_R(\lambda) = \overline{r}(\lambda), \quad \overline{c}_G(\lambda) = \overline{g}(\lambda), \quad \overline{c}_B(\lambda) = \overline{b}(\lambda) \\ &\overline{c'}_R(\lambda) = \overline{r}'(\lambda), \quad \overline{c'}_G(\lambda) = \overline{g}'(\lambda), \quad \overline{c'}_B(\lambda) = \overline{b}'(\lambda). \end{aligned} \tag{I.2}$$

We will use subscript notation, with subscripts $\alpha, \beta, ... = R, G, or B$. Then (I.1) can be written as

$$\overline{c'}_{\alpha}(\lambda) = \sum_{\beta=R,G,B} U_{\alpha\beta}\overline{c}_{\beta}(\lambda). \tag{I.3}$$

where the unit intensities are given by

$$u'_\beta = \sum_{\alpha=\mathrm{R,G,B}} V^{-1}_{\alpha\beta}, \tag{I.5}$$

so that

$$U_{\alpha\beta} = \frac{V^{-1}_{\alpha\beta}}{\sum\limits_{\sigma=\mathrm{R,G,B}} V^{-1}_{\sigma\beta}} \tag{I.6}$$

I.1 Application of the Transformation: Determining an Ideal Set of Primaries

We have tried to determine an ideal set of primaries starting with the TCMF produced by Judd and Wyszecki for the primaries 436, 546, and 700 nm. Our goal is to minimize the total area that encompasses negative primaries. The process we will use involves a bit of trial and error. We first note that a change of primaries can be looked at as taking a piece of rubber on which we draw the TCMF and stretching it in various directions so that the points corresponding to the desired primaries lie at the corners of the triangle entirely encompassing positive color coordinates. The TCMF is best seen as a whole through the corresponding chromaticity diagram encompassed by the **horseshoe perimeter**. See Fig. I.1.

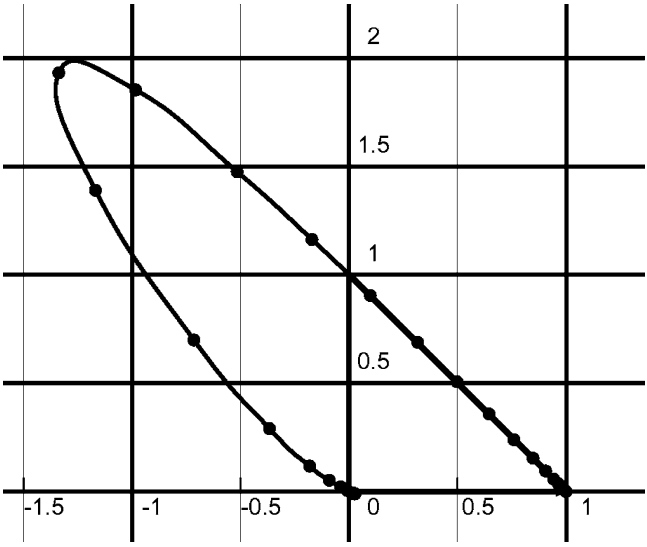


Fig. I.1 The Horseshoe perimeter for the Judd and Wyszecki primaries 436, 546, and 700 nm

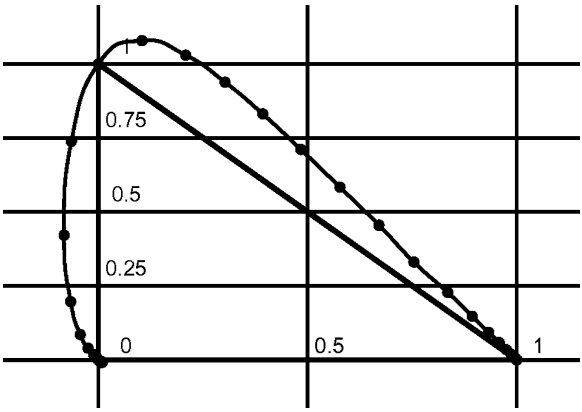


Fig. I.2 The horseshoe perimeter for the primaries 436, 510, and 700 nm

We notice that the bulge to the upper left represents a large region having negative red coordinates. Consequently, this choice of monochromatic primaries is far from ideal in enabling one to match colors entirely with positive color coordinates. On the other hand, the perimeter from the green primary to the red primary is extremely close to being straight, so that there is an extremely small region having a negative blue primary coordinate. Finally, the red primary is at the end of the perimeter while the extreme blue end of the perimeter, – at 400 nm – is extremely close to the blue primary, so that we have an extremely small region having a negative green primary coordinate.

Since the extreme left end, with the greatest negative red coordinate, corresponds to a wavelength of about 510 nm, it is reasonable to study a new set of primaries, with the green primary at a wavelength of 510 nm and with the same red and blue primaries. The resulting horseshoe is shown in Fig. I.2.

The improvement is dramatic. We have significantly reduced the area with a negative red coordinate. However, we now have a significant region with a negative blue coordinate. We therefore experiment with a green primary a bit closer to the original 546 nm, hoping that the reduction in the region with a negative blue coordinate will not lead to a significant region with a negative green coordinate. We next switch to a green primary of 515 nm. The resulting horseshoe is shown in Fig. I.3. The area of negative blue primary is now about the same as the area of negative red primary. With mathematical optimization techniques we could, perhaps, make further improvements; nevertheless, we will stop here and accept what we have now.

Procedure

We now summarize the procedure for obtaining the new TCMF, specifically for the set 436, 510, and 700 nm. We first extract out of the original TCMF of Judd and

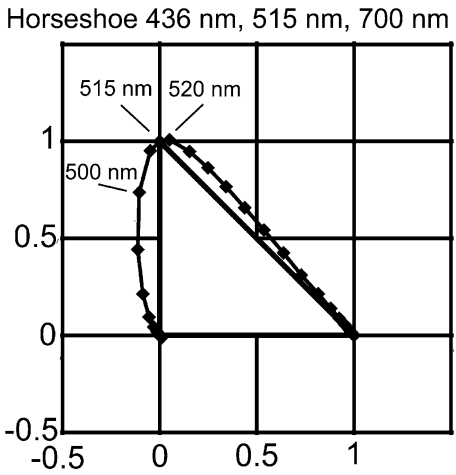


Fig. I.3 The horseshoe perimeter for the primaries 436, 515, and 700 nm

λ(nm)	R	G	B	R	G	B	R	G	B
700	0.0041	0	0	243.9	0	4E-17	0.591	0	0
510	–0.089	0.1286	0.027	168.82	7.776	–0.21	0.409	1	–0.2655
436	0	0	1	0	0	1	0	0	1.2655
			u' =	**412.72**	**7.776**	**0.7902**			

Fig. I.4 Matrices for transforming from the Judd-Wyszecki primaries 436, 546, and 700 nm to the primaries 436, 510, and 700 nm

Wyszecki that is found in Williamson and Cummins those rows having to do with the new primaries.[4] We arrive at the 3 × 3 sub-matrix $\mathbb{V}$ that is shown in **light blue** in Fig. I.4. The wavelengths 700 and 510 nm are present in the TCMF. However, the wavelength 436 nm of the blue primary, which is the same for both the original and the new primaries, is absent. That is not a problem for us since the red and the green color coordinates of the blue primary must be identically zero. The blue color coordinate can be obtained by interpolation between the wavelengths 420 and 440 nm.

Note that while the original TCMF was organized with the wavelengths running from 400 nm at the top to the highest at 700 nm, we have rearranged the three wavelengths in reverse order so that we have R, G, and B running from top to bottom.

We next produce the **inverse of the matrix** $\mathbb{V}$, shown in **yellow** using the Excel.[5] The *unit intensities* of the new primaries are shown in magenta and are the respective

[4]This text has only wavelengths that are multiples of 20, while the original TCMF found in Williamson and Cummins has all multiples of 10.

[5]We highlight a 9 × 9 block of cells. We then type in the command line: **=MINVERSE(C7:E9)**, where (C7:E9) identifies the matrix $\mathbb{V}$ to be inverted – here C7 is the cell ID of V_{RR} and E9 is the cell ID of V_{BB}. Of course, your cell IDs might be different.

sums of the columns above. Finally, we see the transform matrix, $\mathbb{U}$, in **light green**. It is the ratio, cell by cell, of the matrix $\mathbb{V}$ divided by the unit intensity corresponding to the column. Note that the sum of each of the three columns of $\mathbb{U}$ is unity.

I.2 Proof of Equations (I.1) and (I.6)

Let us begin by understanding better the content of the tables. The color of a spectral intensity is produced by mixing sources of the set of given primaries. Physiologically, the three tristimulus values calculated for that spectral intensity are NOT simply proportional to the corresponding rates at which nerve impulses are sent to the brain by the cones. In Chap. 14, we introduced the following functions:

The response functions $S_{\mathrm{R}}(\lambda)$, $S_{\mathrm{G}}(\lambda)$, and $S_{\mathrm{B}}(\lambda)$ are the respective rates at which the cones, R, G, and B emit nerve impulses per *unit* intensity of wavelength λ.

The nerve impulse rate for a given spectral intensity $I(\lambda)$ are: $N_{\mathrm{R}}(\lambda) = S_{\mathrm{R}}(\lambda)I(\lambda)$, $N_{\mathrm{G}}(\lambda) = S_{\mathrm{G}}(\lambda)I(\lambda)$, and $N_B(\lambda) = S_B(\lambda)I(\lambda)$.

For a given spectral intensity, the total nerve impulse rates from the respective cones are given by [6]

$$\begin{aligned} N_{\mathrm{R}} &= \sum_{\lambda} S_R(\lambda)I(\lambda) \\ N_{\mathrm{G}} &= \sum_{\lambda} S_G(\lambda)I(\lambda) \\ N_{\mathrm{B}} &= \sum_{\lambda} S_B(\lambda)I(\lambda), \end{aligned} \tag{I.8}$$

where the right-hand sides are sums over all the wavelengths.

To produce a match by mixing the primary sources, the sources have to produce the same set of nerve impulse rates. Therefore, we next need to obtain expressions for how these rates depend upon the primary sources. We note that generally each of the three primaries produces nerve impulse rates from all three cones. We therefore introduce the following nine quantities:

[6] See Chap. 14. With matrix and vector notation, we have

$$N_\alpha = \sum_{\lambda} S_\alpha(\lambda)I(\lambda). \tag{I.7}$$

$$
\begin{aligned}
S_{\mathrm{RR}} &= \text{Nerve impulse rate of the R-cones per unit intensity of R-primary} \\
S_{\mathrm{RG}} &= \text{Nerve impulse rate of the R-cones per unit intensity of G-primary} \\
S_{\mathrm{RB}} &= \text{Nerve impulse rate of the R-cones per unit intensity of B-primary} \\
S_{\mathrm{GR}} &= \text{Nerve impulse rate of the G-cones per unit intensity of R-primary} \\
S_{\mathrm{GG}} &= \text{Nerve impulse rate of the G-cones per unit intensity of G-primary} \\
S_{\mathrm{GB}} &= \text{Nerve impulse rate of the G-cones per unit intensity of B-primary} \\
S_{\mathrm{BR}} &= \text{Nerve impulse rate of the B-cones per unit intensity of R-primary} \\
S_{\mathrm{BG}} &= \text{Nerve impulse rate of the B-cones per unit intensity of G-primary} \\
S_{\mathrm{BB}} &= \text{Nerve impulse rate of the B-cones per unit intensity of B-primary.} \qquad \text{(I.9)}
\end{aligned}
$$

The nine quantities can be treated as a 9×9 matrix $\mathbb{S}$, exhibited below:

$$
\begin{pmatrix}
S_{\mathrm{RR}} & S_{\mathrm{GR}} & S_{\mathrm{BR}} \\
S_{\mathrm{RG}} & S_{\mathrm{GG}} & S_{\mathrm{BG}} \\
S_{\mathrm{RB}} & S_{\mathrm{GB}} & S_{\mathrm{BB}}
\end{pmatrix}.
$$

The nerve impulse rates depend upon the **tristimulus values**, **R**, **G**, and **B** and the matrix $\mathbb{S}$:

$$
\begin{aligned}
N_{\mathrm{R}} &= \mathbf{R}\, S_{\mathrm{RR}} + \mathbf{G}\, S_{\mathrm{RG}} + \mathbf{B}\, S_{\mathrm{RB}} \\
N_{\mathrm{G}} &= \mathbf{R}\, S_{\mathrm{GR}} + \mathbf{G}\, S_{\mathrm{GG}} + \mathbf{B}\, S_{\mathrm{GB}} \\
N_{\mathrm{B}} &= \mathbf{R}\, S_{\mathrm{BR}} + \mathbf{G}\, S_{\mathrm{BG}} + \mathbf{B}\, S_{\mathrm{BB}}. \qquad \text{(I.10)}
\end{aligned}
$$

We can now use the equations from Chap. 14 for the dependence of the tristimulus values on the color-matching functions, namely,[7]

$$
\begin{aligned}
\mathbf{R} &= \sum_{\lambda} \overline{r}(\lambda) I(\lambda) \\
\mathbf{G} &= \sum_{\lambda} \overline{g}(\lambda) I(\lambda) \\
\mathbf{B} &= \sum_{\lambda} \overline{b}(\lambda) I(\lambda). \qquad \text{(I.12)}
\end{aligned}
$$

[7] In matrix notation, we can define the vector representing the three tristimulus values as $C_\alpha = (\mathbf{R}, \mathbf{G}, \mathbf{B})$. Then

$$
C_\alpha = \sum_{\lambda} \overline{c}_\alpha(\lambda) I(\lambda). \qquad \text{(I.11)}
$$

These expressions, substituted into (I.10), provide us with an expression for the nerve impulse rates in terms of the matrix $\mathbb{S}$. Both this set of equations and the set of (I.9) in terms of the functions $S_R(\lambda)$, $S_G(\lambda)$, and $S_B(\lambda)$ must hold for any spectral intensity and therefore **must hold for any specific wavelength**. This fact will allow us to show how visual physiology reduces the number of independent variables necessary to relate matching with one set of primaries with matching by another set of primaries, as expressed by (I.1).

To clarify the above, let us join the first equations (I.9) and (I.10) and all three equations of (I.12). We obtain

$$\begin{aligned} N_{\mathrm{R}} &= \sum_{\lambda} S_{\mathrm{R}}(\lambda) I(\lambda) = \mathbf{R}\, S_{\mathrm{RR}} + \mathbf{G}\, S_{\mathrm{RG}} + \mathbf{B}\, S_{\mathrm{RB}} \\ &= \sum_{\lambda} \overline{r}(\lambda) I(\lambda)\, S_{\mathrm{RR}} + \sum_{\lambda} \overline{g}(\lambda) I(\lambda)\, S_{\mathrm{RG}} + \sum_{\lambda} \overline{b}(\lambda) I(\lambda)\, S_{\mathrm{RB}}. \end{aligned} \tag{I.13}$$

We obtain a similar equation for N_{G} and N_{B}. Ultimately, we have for each wavelength[8]

$$\begin{aligned} S_{\mathrm{R}}(\lambda) &= \overline{r}(\lambda) S_{\mathrm{RR}} + \overline{g}(\lambda) S_{\mathrm{RG}} + \overline{b}(\lambda) S_{\mathrm{RB}} \\ S_{\mathrm{G}}(\lambda) &= \overline{r}(\lambda) S_{\mathrm{GR}} + \overline{g}(\lambda) S_{\mathrm{GG}} + \overline{b}(\lambda) S_{\mathrm{BG}} \\ S_{\mathrm{B}}(\lambda) &= \overline{r}(\lambda) S_{\mathrm{BR}} + \overline{g}(\lambda) S_{\mathrm{BG}} + \overline{b}(\lambda) S_{\mathrm{BB}}. \end{aligned} \tag{I.15}$$

There are, correspondingly, nine quantities for the second set of primaries, the matrix $\mathbb{S}'$ as well as the second set of color-matching functions, $\overline{r'}$, $\overline{g'}$, and $\overline{b'}$. We also have a set of equations parallel to (I.15):

$$\begin{aligned} S_{\mathrm{R}}(\lambda) &= \overline{r'}(\lambda) S'_{\mathrm{RR}} + \overline{g'}(\lambda) S'_{\mathrm{RG}} + \overline{b'}(\lambda) S'_{\mathrm{RB}} \\ S_{\mathrm{G}}(\lambda) &= \overline{r'}(\lambda) S'_{\mathrm{GR}} + \overline{g'}(\lambda) S'_{\mathrm{GG}} + \overline{b'}(\lambda) S'_{\mathrm{GB}} \\ S_{\mathrm{B}}(\lambda) &= \overline{r'}(\lambda) S'_{\mathrm{BR}} + \overline{g'}(\lambda) S'_{\mathrm{BG}} + \overline{b'}(\lambda) S'_{\mathrm{BB}}. \end{aligned} \tag{I.16}$$

[8] With matrix notation we have

$$S_{\alpha}(\lambda) = \sum_{\beta = R,G,B} S_{\alpha\beta} \overline{c}_{\beta}(\lambda). \tag{I.14}$$

Therefore,[9]

$$\begin{aligned}
\overline{r'}(\lambda)S'_{\mathrm{RR}} + \overline{g'}(\lambda)S'_{\mathrm{GR}} + \overline{b'}(\lambda)S'_{\mathrm{BR}} &= \overline{r}(\lambda)S_{\mathrm{RR}} + \overline{g}(\lambda)S_{\mathrm{GR}} + \overline{b}(\lambda)S_{\mathrm{BR}} \\
\overline{r'}(\lambda)S'_{\mathrm{RG}} + \overline{g'}(\lambda)S'_{\mathrm{GG}} + \overline{b'}(\lambda)S'_{\mathrm{BG}} &= \overline{r}(\lambda)S_{\mathrm{RG}} + \overline{g}(\lambda)S_{\mathrm{GG}} + \overline{b}(\lambda)S_{\mathrm{BG}} \\
\overline{r'}(\lambda)S'_{\mathrm{RB}} + \overline{g'}(\lambda)S'_{\mathrm{GB}} + \overline{b'}(\lambda)S'_{\mathrm{BB}} &= \overline{r}(\lambda)S_{\mathrm{RB}} + \overline{g}(\lambda)S_{\mathrm{GB}} + \overline{b}(\lambda)S_{\mathrm{BB}}. \qquad \text{(I.18)}
\end{aligned}$$

We see above that the two matrices, $\mathbb{S}$ and $\mathbb{S}'$, determine the relationship between the two sets of color-matching functions and that the relationship is identical for each wavelength. The algebra of matrices leads to an expression for the transformation matrix $\mathbb{U}$ of (I.1) that involves the so-called inverse matrix of the matrix $\mathbb{S}'$:

$$\mathbb{U} = \mathbb{S}\mathbb{S}'^{-1}. \qquad \text{(I.19)}$$

For those who are not familiar with these symbols, we will exhibit one of the matrix elements of $\mathbb{U}$:

$$U_{\mathrm{RG}} = S_{\mathrm{RR}}S'^{-1}_{\mathrm{RG}} + S_{\mathrm{RG}}S'^{-1}_{\mathrm{GG}} + S_{\mathrm{RB}}S'^{-1}_{\mathrm{BG}}. \qquad \text{(I.20)}$$

Here, for example, S'^{-1}_{RG} is the RG element for the matrix $\mathbb{S}'^{-1}$.

> **Note**
>
> When the two sets of primaries are identical, we expect the transformation matrix to yield the same TCMF as the original TCMF so that $\mathbb{U}$ should be the so-called identity matrix. Then all the diagonal elements ($U_{\mathrm{RR}}, U_{\mathrm{GG}}$, and U_{BB}) are unity while the remaining six elements vanish. Equation (I.19) confirms this result since in this case $\mathbb{S} = \mathbb{S}'$

We will next prove (I.4)

$$\mathbb{U}_{\alpha\beta} = \mathbb{V}^{-1}_{\alpha\beta}/u'_{\beta}, \qquad \text{(I.21)}$$

where the parameters u'_{β} are the unit intensities for the second set of primaries, given by

$$u'_{R} = V^{-1}_{\mathrm{RR}} + V^{-1}_{\mathrm{GR}} + V^{-1}_{\mathrm{BR}} \qquad \text{(I.22)}$$

with corresponding expressions for the other two unit intensities.

Thus, the second TCMF is determined by the 9 × 9 sub-matrix $\mathbb{V}$ of the original TCMF.

9

$$\sum_{\beta=R,G,B} \overline{c}'_{\beta}(\lambda)S'_{\alpha\beta} = \sum_{\beta=R,G,B} \overline{c}_{\beta}(\lambda)S_{\alpha\beta} \qquad \text{(I.17)}$$

Proof

First let us recall (I.1), which we rewrite here:

$$\begin{aligned}\overline{r'}(\lambda) &= U_{\mathrm{RR}}\overline{r}(\lambda) + U_{\mathrm{GR}}\overline{g}(\lambda) + U_{\mathrm{BR}}\overline{b}(\lambda)\\ \overline{g'}(\lambda) &= U_{\mathrm{RG}}\overline{r}(\lambda) + U_{\mathrm{GG}}\overline{g}(\lambda) + U_{\mathrm{BG}}\overline{b}(\lambda)\\ \overline{b'}(\lambda) &= U_{\mathrm{RB}}\overline{r}(\lambda) + U_{\mathrm{GB}}\overline{g}(\lambda) + U_{\mathrm{BB}}\overline{b}(\lambda).\end{aligned} \tag{I.23}$$

In Chap. 14, we pointed out that if we sum any color-matching function, $\overline{r}(\lambda)$, $\overline{g}(\lambda)$, or $\overline{b}(\lambda)$ over all of the wavelengths, we must obtain the same number so that a constant spectral intensity will produce equal energy white. This fact obviously holds true for the second set of primaries too, except that the constant common to the three sums can be different. We now recall that if we were to multiply every color-matching function in a TCMF by the same number, all tristimulus values are multiplied by that number but the color coordinates are unchanged. We therefore are free to choose the sums to be equal for the two different sets of primaries.

If we carry out this sum in each of the above three equations, we will obtain the following three equations for the transformation matrix[10]:

$$\begin{aligned}U_{\mathrm{RR}} + U_{\mathrm{GR}} + U_{\mathrm{BR}} &= 1\\ U_{\mathrm{RG}} + U_{\mathrm{GG}} + U_{\mathrm{BG}} &= 1\\ U_{\mathrm{RB}} + U_{\mathrm{GB}} + U_{\mathrm{BB}} &= 1.\end{aligned} \tag{I.25}$$

Next, we recall that (see Problem 19 in Chap. 14)

$$\overline{r}(\lambda_{\mathrm{R}}) = \frac{1}{u_{\mathrm{R}}},\quad \overline{g}(\lambda_{\mathrm{G}}) = \frac{1}{u_{\mathrm{G}}},\quad \overline{b}(\lambda_{\mathrm{B}}) = \frac{1}{u_{\mathrm{B}}} \tag{I.26}$$

with the remaining functions, e.g., $\overline{r}(\lambda_{\mathrm{G}})$, vanishing.

Let us introduce the matrix $\mathbb{C}$ defined by

$$\mathbb{C} = \begin{pmatrix}\overline{r}(\lambda_{\mathrm{R}}) & \overline{r}(\lambda_{\mathrm{G}}) = 0 & 0\\ 0 & \overline{g}(\lambda_{\mathrm{G}}) & 0\\ 0 & 0 & \overline{b}(\lambda_{\mathrm{B}})\end{pmatrix}.$$

[10]The three equations can be expressed in matrix notation as

$$\sum_{\alpha} U_{\alpha\beta} = 1 \tag{I.24}$$

for all β.

Similarly, we have a corresponding matrix for the second set of primaries, λ'_{R}, λ'_{G}, and λ'_{B}, we have

$$\overline{r'}(\lambda'_{\mathrm{R}}) = \frac{1}{u'_{\mathrm{R}}}, \quad \overline{g'}(\lambda'_{\mathrm{G}}) = \frac{1}{u'_{\mathrm{G}}}, \quad \overline{b'}(\lambda'_{\mathrm{B}}) = \frac{1}{u'_{\mathrm{B}}} \tag{I.27}$$

with the remaining functions, e.g., $\overline{r}(\lambda_{\mathrm{G}})$, vanishing. We also define the matrix

$$\mathbb{C}' = \begin{pmatrix} \overline{r'}(\lambda'_{\mathrm{R}}) & \overline{r'}(\lambda'_{\mathrm{G}}) = 0 & 0 \\ 0 & \overline{g'}(\lambda'_{\mathrm{G}}) & 0 \\ 0 & 0 & \overline{b'}(\lambda'_{\mathrm{B}}) \end{pmatrix} = \begin{pmatrix} 1/u'_{\mathrm{R}} & 0 & 0 \\ 0 & 1/u'_{\mathrm{G}} & 0 \\ 0 & 0 & 1/u'_{\mathrm{B}} \end{pmatrix}.$$

According to (I.1), we have

$$\begin{aligned} C'_{\mathrm{RR}} &= \overline{r'}(\lambda'_{\mathrm{R}}) = \frac{1}{u'_{\mathrm{R}}} = U_{\mathrm{RR}}\overline{r}(\lambda'_{\mathrm{R}}) + U_{\mathrm{GR}}\overline{g}(\lambda'_{\mathrm{R}}) + U_{\mathrm{BR}}\overline{b}(\lambda'_{\mathrm{R}}) \\ C'_{\mathrm{GG}} &= \overline{g'}(\lambda'_{\mathrm{G}}) = \frac{1}{u'_{\mathrm{G}}} = U_{\mathrm{RG}}\overline{r}(\lambda'_{\mathrm{G}}) + U_{\mathrm{GG}}\overline{g}(\lambda'_{\mathrm{G}}) + U_{\mathrm{BG}}\overline{b}(\lambda'_{\mathrm{G}}) \\ C'_{\mathrm{BB}} &= \overline{b'}(\lambda'_{\mathrm{B}}) = \frac{1}{u'_{\mathrm{B}}} = U_{\mathrm{RB}}\overline{r}(\lambda'_{\mathrm{B}}) + U_{\mathrm{GB}}\overline{g}(\lambda'_{\mathrm{B}}) + U_{\mathrm{BB}}\overline{b}(\lambda'_{\mathrm{B}}). \end{aligned} \tag{I.28}$$

These equations can be rewritten so as to exhibit the matrix $\mathbb{V}$:

$$\begin{aligned} C'_{\mathrm{RR}} &= V_{\mathrm{RR}}U_{\mathrm{RR}} + V_{\mathrm{RG}}U_{\mathrm{GR}} + V_{\mathrm{RB}}U_{\mathrm{BR}} \\ C'_{\mathrm{GG}} &= V_{\mathrm{GR}}U_{\mathrm{RG}} + V_{\mathrm{GG}}U_{\mathrm{GG}} + V_{\mathrm{GB}}U_{\mathrm{BG}} \\ C'_{\mathrm{BB}} &= V_{\mathrm{BR}}U_{\mathrm{RB}} + V_{\mathrm{BG}}U_{\mathrm{GB}} + V_{\mathrm{BB}}U_{\mathrm{BB}}. \end{aligned} \tag{I.29}$$

These equations can be expressed as a multiplication of matrices:

$$\mathbb{C}' = \mathbb{V}\mathbb{U}. \tag{I.30}$$

or

$$C'_{\alpha\beta} = \sum_{\sigma=\mathrm{R,G,B}} V_{\alpha\sigma}U_{\sigma\beta}. \tag{I.31}$$

We know so far only the matrix $\mathbb{V}$. We can obtain the unit intensities (u'_α) as follows. First, we solve (I.31) for the matrix $\mathbb{U}$:

$$\mathbb{U} = \mathbb{V}^{-1}\mathbb{C}'. \tag{I.32}$$

In matrix element form, we have

$$U_{\alpha\beta} = \sum_{\sigma=\mathrm{R,G,B}} V^{-1}_{\alpha\sigma}C'_{\sigma\beta} = V^{-1}_{\alpha\beta}/u'\beta. \tag{I.33}$$

Next we recall (I.25). It can be written as

$$\sum_{\alpha=\mathrm{R,G,B}} U_{\alpha\beta} = 1 \tag{I.34}$$

for all β. Therefore,

$$u'_\beta = \sum_{\alpha=\mathrm{R,G,B}} V^{-1}_{\alpha\beta}. \tag{I.35}$$

Finally we have our expression for the transformation matrix $\mathbb{U}$ in terms of the submatrix of the original TCMF, (I.6):

$$U_{\alpha\beta} = \frac{V^{-1}_{\alpha\beta}}{\sum_{\sigma=\mathrm{R,G,B}} V^{-1}_{\sigma\beta}}. \tag{I.36}$$

I.3 Problems on the Transformation of TCMFs

1. Suppose that the matrix $\mathbb{V}$ has only diagonal matrix elements. We say that the matrix is a **diagonal matrix**.

   Explain why the matrix $\mathbb{U}$ is diagonal, with all its matrix elements equal to unity. What does this imply about how the new set of primaries is related to the old set?
2. Below is the TCMF produced by Stiles and Burch (See Chap. 14) for the primaries $\lambda_{\mathrm{B}'} = 444.44\,\mathrm{nm}$, $\lambda_{\mathrm{G}'} = 526.32\,\mathrm{nm}$, and $\lambda_{\mathrm{R}'} = 645.16\,\mathrm{nm}$ (Table I.1).

   (a) According to the theory of color vision we have presented the data collected by the two groups, Stiles and Burch vs. Judd and Wyszecki, should be consistent. For example, if we were to apply our transformation matrices to the TCMF of Stiles and Burch, we should obtain the TCMF of Judd and Wyszecki.

   Carry out this process and compare the TCMFs. Beware that the two might differ by a constant factor that multiplies all the coordinates. Why is this so? On the other hand, the color coordinates do not depend upon such a constant factor, so compare the horseshoe perimeters.

   (b) Derive the transformation matrix $\mathbb{U}$ for the set of primaries 436, 520, and 700 nm that is based on the Stiles and Burch TCMF.

   Find the TCMF for this second set of primaries and the corresponding horseshoe perimeter.

   Compare this perimeter with the one we derived by a transformation from the Judd and Wyszecki TCMF of the set of primaries 436, 546, and 700 nm.

Table I.1 Table of color-matching functions based on the spectral primaries 444.44, 526.32, and 645.16 nm

λ (nm)	$\overline{r}(\lambda)$	$\overline{g}(\lambda)$	$\overline{b}(\lambda)$
400	0.0089	−0.0025	0.04
410	0.035	−0.0119	0.1802
420	0.0702	−0.0289	0.467
430	0.0763	−0.0338	0.6152
440	0.0561	−0.0276	0.8778
450	−0.0044	0.0024	1.0019
460	−0.097	0.0636	0.9139
470	−0.2235	0.1617	0.7417
480	−0.3346	0.2796	0.472
485	−0.3776	0.3428	0.3495
490	−0.4136	0.4086	0.2564
500	−0.4452	0.5491	0.1307
510	−0.414	0.7097	0.058
520	−0.2845	0.8715	0.02
530	−0.0435	0.9945	0.0007
540	0.3129	1.0375	−0.0064
550	0.7722	1.039	−0.0094
560	1.271	0.9698	−0.0097
570	1.8465	0.8571	−0.0087
580	2.425	0.6953	−0.0073
590	2.9151	0.5063	−0.00537
600	3.1613	0.336	−0.00357
610	3.1048	0.1917	−0.00208
620	2.7194	0.0938	−0.00103
630	2.17	0.0371	−0.00044
640	1.5179	0.0112	−0.00014
650	1.007	0.000078	0
660	0.5934	−0.001988	0
670	0.3283	−0.002006	0
680	0.1722	−0.001272	0
690	0.0853	−0.000683	0
700	0.0408	−0.000337	0

3. Prove that the tristimulus values C'_α for a second set of monochromatic primaries is related to the tristimulus values C_α of the first set by the equation

$$C'_\alpha = \sum_\beta U_{\alpha\beta} C_\beta. \tag{I.37}$$

4. (a) Prove that the sum $S = \sum_\alpha C_\alpha$ is independent of the set of monochromatic primaries.
   (b) Prove that the color coordinates satisfy the equation

$$\frac{\overline{c'}_\alpha}{C'_\alpha} = \frac{\overline{c}_\alpha}{C_\alpha}. \tag{I.38}$$

Appendix J
Hommage to Pierre-Gilles de Gennes: Art and Science**

In July 2007, I came to know that the physicist **Pierre-Gilles de Gennes** had passed away on May 18, 2007. I received the news with deep sadness since de Gennes was my supervisor when I was a young Post-Doc at the Laboratory of the Physics of Solids in Orsay, France, in the academic year 1966–1967, and I had crossed paths with him many times since then. While only 7 years older than me, he had already achieved fame as a leader of a group of physicists studying superconductivity. De Gennes had a great influence on all those working in the field of Condensed Matter. Whatever subject he touched, he transformed. Frequently he attacked problem areas for which physicists and chemists felt that they had reached the limits thought possible conceptually. De Gennes discovered sets of concepts that produced a new exciting level of activity of research in the fields (Fig. J.1).

One meeting I had with him stands out in my memory: I was working on a problem and was stuck because of not knowing one of the essential parameters of the problem. **de Gennes** listened to me attentively with all due patience. He then proceeded to ask me one simple question, which I will recall with clarity: "What is the approximate wavelength of the electrons in bismuth in comparison to the roughness of the surface of the sample." The reader need not understand the meaning of the question. It was the utter simplicity and directness of his question that mattered. For that question was all I needed to hear for me to proceed to complete my study. Our meeting was very brief and yet extremely productive.

Of what relevance is this encounter to this book, in which our goal is to appreciate the connection between physics and the arts? It is that some people have an extraordinary gift to see order and simplicity in the complex. De Gennes was one of those who applied this ability to many fields, whose numerous familiar applications depend upon the results of his research. This ability to create order from a suspended blank state of our normal awareness is manifest in artists as well. De Gennes is quoted as describing the deep influence that the painter **Pablo Picasso** had on his own scientific studies. To appreciate this point, I am including an excerpt of one of a number of obituaries written after his death. For details see: http://www.nature.com/nature/journal/v448/n7150/pdf/448149a.pdf.

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

Fig. J.1 Pierre-Gilles de Gennes (source: http://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/materials/public/DeGennes/DeGennesintro_fr1252.htm)

The obituary makes reference to a film on Picasso that left a great mark on de Gennes' research as a physicist. You can view an excerpt of this film on the following YouTube site: http://www.youtube.com/watch?v=5tn5uTTCRCg.

Excerpt of an Obituary of de Gennes by Françoise Brochard-Wyart

"Chacun en nous a son trésor d'images entrevues dans un instant mais jamais oubliées. Un exemple pour moi: Picasso peignant grands traits blancs sur une vitre et filmé par Clouzot. Tout ce que j'ai essayé de dessiner laborieusement plus tard est né de ces moments."

"Every one of us has a treasure of images caught in glimpses but never forgotten. A personal example: Picasso painting white lines on glass using large strokes, filmed behind the glass by Clouzot. Everything that I tried painting laboriously later was born from such glimpses."
(Pierre-Gilles de Gennes, from L'émerveillement by Thibaut de Wurstemberger, Saint-Augustin, 1998.)

With his strikingly simple yet pioneering ideas, Pierre-Gilles de Gennes drew "white lines in large strokes" that defined the physics of soft matter – liquid crystals, polymers, colloids, and surfactants. He died on 18 May.

Educated at the École Normale Supérieure in Paris during 1951–1955, de Gennes learned theoretical physics from the greatest masters of his time. He obtained his Ph.D. in 1957 while at the French Atomic Energy Commission, specializing in magnetism and neutron scattering. During a stay at the University of California, Berkeley, in 1959, he studied with the solid-state physicist Charles Kittel, who taught him how to communicate ideas in physics using plain language, so avoiding the use of daunting equations....

De Gennes fostered a collective research effort that is scarcely imaginable today. Papers were signed not with the names of individuals, but with the name of the group. Theoreticians would spend half their time contemplating liquid crystals

under the microscope and discussing practical experiments. Researchers would often arrive in the morning to find a note from de Gennes that would launch them in yet another ground-breaking direction. Calling on his vast knowledge of physics, de Gennes drew analogies between different fields. For example, he realized that laws developed to describe superconductivity phenomena could be used to understand phase transitions in liquid crystals.

...

De Gennes pursued his research with extraordinary imagination, insatiable curiosity, and an ability to grasp facts rapidly. But he also gave his time to others and helped them develop their ideas. A keen ambassador of science to the public, he generated passionate debates on subjects as diverse as "Physics and Medicine," "Inventors," and "Primo Levi." He inspired generations of students to pursue careers in physics and played an active role in establishing the L'Oréal-UNESCO Awards for Women in Science.

...

End of Excerpt of the Obituary

Here is another expression of a likening of de Gennes revelations to the product of an artist in an excerpt of a description of de Gennes' book on Soft Interfaces: "(The book provides us with) an impressionistic tour of the physics of soft interfaces by Nobel Laureate Pierre-Gilles de Gennes. Full of insight and interesting asides, it not only provides an accessible introduction to this topic, but also lays down many markers and signposts for interesting new research possibilities."

I will end this essay with a personal remark:

Over the years, de Gennes and I corresponded and saw each other only once every few years. Yet, in spite of his fame and his multitude of activities and acquaintances, he remembered even the most trivial of our encounters. Most vivid in his mind when I last saw him at a talk he gave at Harvard was about a dinner that his wife and he had prepared for my family in 1967. They had forgotten to buy fresh cheese for our dessert; this mistake required deep apologies and remained in his mind for 40 years.

Appendix K
MAPPINGS as a Basis for Arriving at a Mutually Agreed Upon Description of Our Observations of the World – Establishing 'Truths' and 'Facts'

This book addresses the subject of physics, sound and light along with their relationship to our own experience of sound and light. We have introduced many concepts and equations that provide relationships among various physical quantities. Physics is all about relationships. And so is a piece of music or work of fine art: There are the relationships we perceive about the components of a given piece of music or of a given work of art. In addition, there are relationships of these components and their sum total that produce the full composition with our personal emotional responses to these compositions. All of these relationships are examples of what are more generally called **mappings**.

All forms of communication involve mappings. Moreover, physical laws are mappings of observations - that we share - onto mathematical equations. Satisfactory communication, as well as satisfactory laws of physics, require agreement among those who share them. In this appendix, we will explore this subject a bit and relate mappings to the complex philosophical questions of truth and fact. According to my colleague George Smith of the Department of Philosophy at Tufts University, who is an expert on Isaac Newton, my orientation towards the nature of scientific investigation is within the framework of Newton's proposed system thereof.

I was led to consider mappings seriously because of my study of **color vision**, as this subject compels us to think with great clarity about the nature of mappings. One of the first sets of words parents teach their babies is colors. This process provides us with a wonderful example of how people learn how to share a common mapping of human experience. The parent shows the baby an object with a uniform surface of color, points, and says the word for the color of the surface - for example, *red*. The term color is technically better referred to as the **hue**, the term that we will henceforth use in this appendix. The parent then points to another object and says *green*. The baby must learn that it is the hue that the word is distinguishing and not another aspect of the object such as its shape or size. How is this aspect provided? – by using a number of objects that hopefully differ essentially in all ways except for the hue of their surfaces.

We realize that there are different hues that are similar but not identical, with different levels of saturation. Note that, as we pointed out earlier in Chap. 14, there

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

is no way to tell how the actual sensations compare among people. The same would be true for the baby vis a vis the parent.

There are cases when a baby will be confused: The parent shows the baby two surfaces, e.g. one red, the other green, and assigns these two different hues to the surfaces. The baby, on the other hand, seems to jump around, randomly assigning one or the other hues to both surfaces. The baby doesn't seem to differentiate. As you might guess, the baby is color blind. How does the baby handle this confusion?[1] I bring up color blindness here just to point out that there are situations wherein people are not always able to establish mappings that they can agree on. Imagine what the situation would be like if the prevalence of various types of colorblindness were close to 100%!

Note

Suppose that an infant is fitted with a device placed over its eyes that inverts all images throughout infancy. Consider how the infant would map observations onto language:

1. Can you think of situations where there might be confusion in communication having to do with up and down?
2. How would the child draw itself as it sees itself in a mirror. Would the child draw an image that is upside down to us?

[1] In color blindness, two words, red and green, are perceived to represent the same experience - perceived color. Ultimately, the child will be told that the two words represent different colors that between which he is incapable of distinguishing. I had problems of confusion of two words of a different sort in hearing Yiddish as a child. In a number of cases, two different pronunciations or even words were randomly assigned to what appeared to me to represent the same idea. I was confused and blamed my difficulty on my own inability to remember or learn the correct pronunciation or to distinguish between two 'different' words. For example, the word for the number 'two' was pronounced as either 'tsvay' (as in the English word 'say') or 'tsvy' (as in the English word 'my'). I heard them as two different words. Ultimately, as an adult, I was told that the reason that my relatives were jumping back and forth between two pronunciations was that they naturally spoke with a 'Galitsianer' accent (close to a German accent). However, Litvaks (from Lithuania), with their Litvak accent, were regarded as being more cultured. As a result, my relatives were sometimes embarrassed about their natural Galitsianer accent.

Of course there is a difference between the case of color blindness and a confusion between two dialects: While both involve a mapping of two different words onto what is conceived as representing the same experience, contrary to the latter situation, the former involves an intrinsic deficiency in perception that cannot be cured by explanation.

K.1 MAPPINGS as Central to Organizing Human Experience

Essentially all human experience is dominated by **mappings** of one kind or another. In the context of these notes, a mapping involves an association between two aspects of human experience.

Examples are:

1. written letters that spell words and their verbal counterpart as expressed words
2. words that refer to classes of objects referred to as nouns
3. images that we perceive in our conscienceness and the scenes that produce physical responses on the retina of an eye
4. printed musical notes and the tones produced by a musical instrument or the human voice
5. frequency and intensity of a pure tone and a sense of pitch
6. spectral intensities and the corresponding sensations associated with color (hue and saturation) and brightness.
7. a sequence of positions of an object and the perception by the eye and brain as 'motion' of the object.
8. the memory a person has of various perceptions of past inputs that correspond to actual physical inputs to a person's senses
9. words that classify many objects that produce an experience that is common in some respect or respects - such as the appearance of tigers, lions, humans, or apples or love, or anger. Sometimes there is disagreement as to how objects are related to the words we ascribe to them. Severe arguments can arise, often merely as a result of people having different mappings. In these cases, ultimately what is important is how such classifications affect the way we use them - that is, how they are mapped onto other actions or attitudes. The important thing is for people to clarify as best then can the mappings they are using. [2]

K.2 NUMBERS as a Mapping

A number of years ago, my wife, my then nine year old son Avi, and I were in Grenoble, France for one of my sabbaticals. Avi went to l'École Houille-Blanche, a public school whose student body was 50% French and 50% foreigners from all over the world. Avi was placed in a class with foreign children ranging in age from about 6 to 10, none of whom knew French. Few shared any particular language.

[2]Recently (2008) the International Astronomical Union decided to demote Pluto to the status of being a "dwarf planet". See the article in the National Geographic News, July, 2008. http://news.nationalgeographic.com/news/2006/08/060824-pluto-planet.html

It seems to me to ludicrous to regard astronomers of the past as having been mistaken in labeling Pluto as a planet. All we can say is that this new label allows astronomers to make statements about the now regarded 'true' planets that will not be applied to Pluto.

How are such children to be taught and be prepared to join the rest of the student body in classrooms that used essentially only French? All I will mention here is the following: The very first subject that students were taught was mathematics - numbers being the first of this subject. Why was this so? Because it is relatively easy to teach and discuss the concept of numbers without using a particular verbal language. All one has to do is to present a number of objects such as one's fingers and assign a word to each finger: "One", "two", "three", "four", . Or: "un", "deux", "trois", "quatre", ... We are observing the establishment of a one to one correspondence between an ordered set of objects (such as our fingers) and words expressed verbally or in written script. This numerical one to one correspondence is perhaps the simplest example of a "mapping".

Most of us would appreciate the probability that the first elements of communication between an earthling and an extraterrestrial would be the sharing of our 'words' for numbers. The reason is the simplicity of this mapping and the small chance that the mapping will not be correctly communicated.

K.3 The Concept of TIME as a Mapping

What is **time**? The first level of consideration and observation regarding **time** is the existence of an ordered sequence of observations. We refer to this observation as **time order**. This ordering is preserved in the patterns that our minds provide in what we call **memory**. Imagine what would happen if our brains destroyed the order or direction of this sequence! The next level in establishing or characterizing our sense of so-called 'time' requires that the physicist observe a system behaving in a cyclic way: A pattern is observed to repeat itself again and again, with negligible observable change in the pattern. A sense of equality in the evolving pattern leads one to associate a time interval to a single occurrence of the pattern and to then assign a numerical value to an evolution of patterns - we number and order the patterns. The patterns are observed to be occurring simultaneously with other physical observations so that we can assign a value to the time interval of a sequence of physical observations. This special cyclic system becomes our "clock". Any time an event takes place, such as hearing a pulse of a sound or noting the position of a car on the highway, we can **correlate**, that is map, that event onto the numerical value of the number of cycles of a clock has made since we assigned an initial time. We can express the time interval between events by noting how many cycles took place between the two events. Here we have a mapping between two events and number of cycles of a clock.

Now consider that astronomers used the rotation of the earth and its revolution about the sun in order to measure time. These processes were believed to be periodic. People could thus count the days or years by making reference to the position of the sun or moon or stars in relation to the earth. Galileo is understood to have studied the motion of objects with respect to time at one point by relying on his trust of

good time keeping by a musician acquaintance.[3] Later, Galileo used clocks that were still quite crude compared to what we would demand today. On the basis of the measurements of astronomers and scientists like Galileo, Newton was led to his three laws of mechanics and his Universal Theory of Gravitation. **Christian Huyghens** was able to improve on the limitations of the **pendulum clock** by extending the observations of Galileo on an inclined plan through his contributions to **mathematics**.[4] Pure mathematics, along with Galileo's experiments, justified the trust he had in his clock.

Later developments in the improvement of clocks with respect to precision and accuracy depended upon the Laws of Mechanics. Their quality was based on **theory**. We now have the **Cesium clock**, which is understood to have an accuracy of one nanosecond (10^{-9} seconds) per day, or about one part in 100-trillion![5] The quoted accuracy is based upon an application of quantum theory.

The logic behind clocks is a bit confusing: Experiment based upon crude clocks led to theory; theory then led to more accurate clocks. Where lies the ultimate basis of evidence? Experiment or theory? Is the logic circular and therefore flawed?

It might seem as if we use theory as our ultimate judge, so that **circular reasoning** is not present. But that is not exactly so. The situation is more complex.

While Newton proposed his laws on the basis of a restricted domain of observations, his laws have been ultimately applied to a vast set of interconnected physical phenomena - for example, all the developments in engineering and medicine and in sending rockets to the moon. The Laws of Physics weave a network, an edifice, such that if any component were to fail to fit the theory, the structure would lose its reliability. It is because of the solidity of this edifice that physicists have such a high degree of faith in the laws of Physics - yes **faith** in the laws.[6] How did we end up with quantum theory? The answer is that new experiments destroyed our total trust by revealing that the edifice was flawed in a domain that takes into account the behavior of systems the size of atoms or smaller. Classical physics misses certain fine details and therefore had to be refined. Ultimately, Quantum Theory wove an intricate edifice that became the basis for a new level of trust as did the classical laws - hence the trust in the accepted accuracy of the Cesium clock. Nevertheless, physicists still use Classical Laws to account for or describe most behavior in the

[3] See Drake, S., *The Role of Music in Galileo's Experiments*, Scientific American, p. 98, June 1975. Also see the website (2-4-2011): http://www.joakimlinde.se/java/galileo/, which contains a beautiful applet that enables us to appreciate Drake's conjecture as to how Galileo might have used a musician to arrive at his law that when a ball rolls down an inclined plane, its speed increases linearly with time.

[4] A pendulum bob moves along a curved path that can be analyzed in terms of an infinite sequence of infinitesimal inclined planes having different angles of inclination.

[5] (2-5-2001): http://en.wikipedia.org/wiki/Atomic_clock

[6] Reader beware: the faith to which I am referring is not the same as the faith in religion, which has no such edifice and yet has its great benefits in helping some people handle the complexity of life's experience.

large, recognizing that corrections sometimes have to be made to take into account Quantum Theory.[7]

Note: I have raised the issue of time here because it is an example wherein the nature of a mapping can be complex and obscure.

K.4 Mappings as the Essential Goal of Physics

We observe the world about us. These observations are summarized by mappings within our brains. We communicate with others in words that represent these mappings, hoping to summarize these mappings in such a way that we can establish a one-to-one correspondence between our words and our observations that are shared among our fellow human beings. I will repeat a statement to be found in Chap. 5:

The essential goal of physics is to establish a theoretical framework for describing in a quantitative way what we decide to and are able to measure. That framework makes use of models, concepts, and images. However, its ultimate content is a set of mathematical equations, which we call laws. The laws are as simple and all-encompassing as possible, and provide relationships among measurable quantities.

One of the most remarkable examples of such a mapping is the following: We observe an enormous variety of materials made out of a relatively small number of different kind of atoms (fewer than 100) arranged in a multitude of ways. We have millions of different organic compounds, metals and alloys, complex materials like wood, and so on. They have a variety of physical properties with respect to pliability, density, color, texture, and so on. And yet, it is understood by physicists that this entire variety of properties is describable, that is, can be *mapped* onto a set of a small number of mathematical equations.

For comparison sake: Note that a finite set of coupled algebraic equations are incredibly simple in content. For example, suppose that we have to solve the two equations, $x + y = 1$ and $x - y = 3$ for x and y. The solution is $x = 2$ and $y = -1$. Such equations cannot provide us with the richness of content that is associated with the behavior of materials.

Having discussed mappings, we need to answer the question of the relationship of mappings to questions of **truth** or **fact**. In my opinion, these issues are not subject to being defined by science. They are purely philosophical. In practice, I would say that we tend to use these terms in science to describe mappings for which there is essentially universal agreement under the rules that are used by scientists

[7] I must warn the reader that the above view of Physics as being ultimately dependent upon faith is not shared by many physicists. Interestingly, this issue does not seem to arise among mathematicians because they recognize that a mathematical theory is dependent upon a set of axioms that is not provable.

to test for acceptability. Any person has a *right* not to accept these rules, often to their own detriment. To add to this list of what I consider a **non-scientific issue** is the question of **reality**. "Do photons really exist?" I heard recently this question argued at a colloquium at Harvard University, during which Nobel prize winners couldn't agree![8] At best, we can say that the photon is a conceptual tool that is represented mathematically in physical laws that are mapped onto observations. Interestingly: While physicists might debate and disagree about the issues of truth, fact, and reality, these disagreements don't seem to affect their ability to conduct the discipline of physics.

A Final Remark

Let us recall the opening chapter of this book, wherein we exhibited a graph of a wave for a piece of music. It can be a joy to contemplate that this curve has all the content that maps onto our incredibly rich, sensual, and emotional experience when listening to the sound associated with the wave pattern. The graph maps onto a sound wave that ultimately produces nerve impulses that are processed in our programmed brains so as to produce our musical experience.

[8] I ought to be specific: While Max Planck based his theory of Black Body radiation on the assumption that electromagnetic radiation is absorbed and/or emitted by atoms in discrete units, he didn't believe that the radiation itself was quantized. In 1905, Einstein produced a theory of the so-called **photoelectric effect**, which involves electromagnetic radiation knocking electrons out of a metal. Einstein's theory assumes that discrete units of radiation collide with the electrons. As a result it has been commonly understood that this experiment along with Einstein's theory provide proof as to the photon's existence. Someone in the Harvard audience asked whether the photoelectric effect does indeed prove the photon's existence. The vote was overwhelmingly in the negative, but not unanimously. Interestingly, no one who voted in the negative proposed an experiment that does prove its existence. Another example of existence questions has to do with atoms. Planck rejected their existence until as late as the early 1900's. Brownian motion (the motion of micron sized particles in a water suspension, due to collisions of water molecules with the particle) is given credit for **proving** their existence.

Index

Symbols
Čerenkov radiation, 276
24-bit color, 448

A
absorbed, 245
absorption, 119, 189
absorption coefficient, 119
absorption spectra, 453
accelerating charge, 175
acceleration, 175, 485
accommodation, 385
action-at-a-distance, 128, 129
additive mixing, 410
additive primaries, 415
after images, 390
after-image, 460
air pressure, 65
air resonance, 82
alychne, 443
AM - amplitude modulation, 49
ambient pressure, 310
Ampère, André Marie, 154
amplifier, 70
amplitude, 24, 26, 30, 318
amplitude modulation frequency, 49
amygdala, 505
analysis, 9
angle of incidence, 242, 249
angle of reflection, 242
angle of refraction, 249
anisotropic, 131
antinode, 21, 73, 76
aqueous humor, 387
Archimedes, 311
Archimedes' Principle of Lever Action, 311
atom, 127, 130
attenuate, 98
attenuation, 87, 110, 116
attenuation constant, 115
attenuation length, 114
attenuation time, 110
auditory canal, 310
auditory nerve, 317, 323
Australopithecus bosei, 1
axis of the lens, 260

B
background noise, 504
bandwidth, 398
bar magnet, 131
base, 106
basilar membrane, 315
bat, 295
battery, 143
beat frequency, 221, 339, 375
beating, 339
beats, 221, 227
Bell, Alexander Graham, 105
bending force, 42
biconcave lens, 260
biconvex, 299
biconvex lens, 260
Big Bang, 201, 287
biosphere, 291
birefringence, 281
bleaching of rods, 390
blind spot, 394
blue color of sea, 341
blue cone, 390, 415
Boethius, 8
Bohr, Neils, 184

L. Gunther, *The Physics of Music and Color*, DOI 10.1007/978-1-4614-0557-3,

brain, 305
brightness, 101, 401, 413, 430
British thermal unit, 88
bulk modulus, 71
burning, 95
buzzer, 141

C
calcite, 281
Carl E. Seashore, 356
carrier frequency, 50
carrier wave, 49
centi, 477
central ray, 263, 264
cents, 371, 374
Cesium clock, 529
chemical energy, 93, 98
chiral, 288
chiral biosphere, 291
chirality, 288
Chladni plates, 59
Chladni, Ernst, 58
chromatic aberration, 262
chromatic scale, 355, 365
chromaticity, 419, 420, 424, 434
chromaticity coordinates, 424
chromaticity diagram, 415, 419
CIE tristimulus values, 439
ciliary muscles, 385
circular reasoning, 529
classical physics, 182, 529
clockwise rotation, 288
closed pipe, 76
cochlea, 306, 315
cochlear fluid, 306
cochlear implant, 319, 324
coherent, 109, 197, 213, 216
color, 413, 419
color blindness, 414
color coordinates, 419, 424, 429, 430
color matching functions for Wright–Guild primaries, 432
color monitor, 445
color patch, 408
color vision, 525
color-matching functions, 415, 432
combination modes, 341
combination tone, 306, 335, 350
compass, 127
complement, 461
complementary color pairs, 416
complementary colors, 390, 430
complementary hue, 432
complements, 430
compound lens, 268, 383
concave lens, 260
condensation, 68
conduction electrons, 130
conductor, 130
cone response, 455
cones, 387
consonance, 348, 360
constructive interference, 210, 229
Convergence, 391
converging lens, 259
convex lens, 260
convex-concave lens, 260
Cornea, 383
Cornsweet, Tom N., 470
corrective lens, 396
correlate, 528
Coulomb's Law, 129
counterclockwise rotation, 288
critical angle, 252, 298
critical bandwidth, 334
crystal, 220
cycle, 15

D
damping, 110
dark adaptation, 391
day vision, 390
dB, 105
de Gennes, Pierre-Gilles, 521
Debussy, Claude, 373
decay, 52
decibel, 105
degree of saturation, 432
Dennett, Daniel, 327
density, 73
depth perception, 383, 391
desaturated, 429
destructive interference, 211, 229
determinism, 486, 489
deuteranope, 458, 464
deuteranopia, 458
Deutsch, Diana, 345
diagonal matrix, 518
diatonic scale, 355, 359
dichromats, 458, 464
differential and integral calculus, 319
diffraction, 231, 233
diffraction angle, 233
diffraction broadening, 235
diffraction grating, 180, 218
diffuse reflection, 241, 244

diffusely, 295
diopter, 395
Disparity of eye position, 392
dispersion, 180, 257, 258, 260
dispersive, 257
displacement, 69
displacement current, 167, 168
dissipation, 100, 110
dissonance, 360
diverging lens, 259
dominant wavelength, 432
Doppler effect, 231, 272, 379
dull surface, 241
dwarf planet, 527

E
eardrum, 306
echo-location, 295
effective focal length, 396
efficiency, 101
Einstein's Theory of Special Relativity, 277
Einstein, Albert, 5, 199, 531
electric battery, 131
electric charge, 127, 129
electric current, 130
electric dipole, 153
electric field, 149, 150
electric field lines, 152
electric generator, 143, 147
electric motor, 142
electrical energy, 93
electrical resistance, 143
electricity, 127
electromagnet, 140
electromagnetic force, 127
electromagnetic radiation, 93, 193
electromagnetic wave, 127, 168
electron, 130, 531
electron-volt, 88
electrostatic potential energy, 93
EMF, 143, 144
empirical, 182
enantiomorphs, 288
end correction, 79
energy, 87, 88
energy level, 193
energy level diagram, 193
energy spectrum, 193
enharmonic equivalents, 366
envelope, 52, 110, 114, 317
envelope of waves, 319
equal energy, 424
equal energy spectrum, 401
equal energy white, 429
equal temperament, 18, 356, 359, 369
equilibrium, 96
equilibrium state, 26
Erg, 88
Escher, M.C., 345
Estill method, 331
Eustachian tube, 310
exchange frequency, 45
excited state, 193
eye-brain system, 243
eyepiece, 499

F
f-hole, 82, 83
fact, 530
faith, 529
far-sighted, 385
Faraday's Law, 159
Faraday, Michael, 143
Feynman, Richard, 3
field, 149
Fifth, 358
first harmonic, 20
fix tuned, 354
FM - frequency modulation, 49
FM modulation frequency, 50
FM radio wave, 50
focal length, 260, 385
focal point, 260
Food calorie, 88
footplate, 306
force, 34
four primary colors, ix
Fourier amplitude, 46
Fourier analysis, 46
Fourier spectrum, 46
Fourier synthesis, 47
Fourier's Theorem, 46
Fourier, Jean Baptiste Joseph, 46
fourth, 358, 363
fovea, 390
frequency, 15
frequency spectrum, 46
fringe, 214
fructose, 288
fulcrum, 311
function, 105
fundamental, 52
fundamental forces, 127
fundamental mode, 20
fusion of harmonics, 344

G
Galilei, Galileo, 528
gamma correction, 444
gamut, 427, 438
gene, 390
general fluorescence, 196
General Relativity, 202
Georg von Békésy, 317, 504
giga, 477
glaucoma, 387
gram, 128
gravitational force, 127
Gravitational Potential Energy, 90
gravitational redshift, 279
green cone, 390, 415
Gregory, R. L., 387, 461
ground state, 189, 193
guitar pickup, 166

H
Haiku, 305
hair cells, 315, 317
half-silvered mirror, 223
harmonic distortion, 335
heat, 94
heat transfer, 96
helicotrema, 315, 316
Helmholtz resonator, 82
Helmholtz, Hermann, 82, 319, 361
Hertz, Heinrich, 15, 171
homochiral, 292
Hooke's Law, 71
Hooke, Robert, 27, 71
horsepower, 100
horseshoe magnet, 132
horseshoe perimeter, 509
horseshoe perimeter of chromaticity diagram, 426
hs color, 448
hsb-color, 448
Huang Chung, 7
hue, 401, 413, 418, 431, 432, 525
human ear, 305
Huyghens, Christian, 199, 252, 529
hyperopic, 385

I
ideal polarizer, 283
image, 263
image point, 263
impedance, 248, 297, 309
in phase, 207
incoherence, 429
incoherent, 109, 216
incoherent sources, 429
incoming flux, 201
incompletely polarized, 284
index of refraction, 246, 297
induced electromotive force, 143
inertia, 128
inner ear, 315
insulator, 130
intensity, 87, 101, 102, 402
intensity density, 402
intensity level, 329
interface, 249
intonation, 354
inverted and real image, 266
ion, 130
ionized, 130
iris, 383
irreversible, 269
Isaac Newton, 92
isotropic, 131

J
Joule, 88
Joule, James Prescott, 88
Just noticeable difference in frequency, 348, 414
Just noticeable difference of loudness, 348
just noticeable difference of pitch, 381
Just Noticeable Differences in Color, 414
Just tuning, 359, 361, 363

K
key note, 355
kilo, 477
kilocalorie, 88
kilogram, 128
kilowatt-hour, 88
kinetic energy, 89, 96

L
largest common denominator (LCD), 51
Laser, 197
Law of Action and Reaction, 138
Le Grand, Y., 413
Leakey, Louis, 1
left handed, 288
Leibniz, Gottfried WIlhelm, 319
length scale of roughness, 241
length scales, 244

lens, 259, 383
lens-humor interfaces, 387
Lenz's Law, 164
Leu Buhwei, 7
lever action, 311
light, 127
light intensity, 383
light receptors, 387
lightness, 440
line of purples, 426
line spectrum, 180
linear mass density, 34, 36
linear response, 336
localized photon, 187
lodestone, 127, 131
logarithm, 105
longitudinal, 13
Lord Rayleigh, 196
loudness, 17, 87, 101
loudspeaker, 141
lRGB gamut, 450
luminance, 440
luminosity, 103

M
Mach bands, 323
Mach one, 276
Mach's Law of Simultaneous Contrast in Vision, 321, 322
Mach, Ernst, 321
macroflow, 130
macroscopic bodies, 93
magenta, 423
magnet, 131
magnetic field, 149
magnetic force, 149
magnetic polarization, 166
magnetism, 127
magnetized, 127, 133
magnification, 266, 495
magnifying glass, 266, 496
magnifying power, 266, 495
main air resonance, 82
major second, 358
major sixth, 358, 363
major third, 358, 363
Malus' Law, 283
mapped, 413
mapping, 421, 525, 527
Masaoka Shiki, 305
masking, 348
mass, 26, 91, 128
mass density, 63
mathematics, 529
matrix, 508
Maxwell, James Clark, 167
Mayer, Julius Robert, 319
mean free path, 64
mechanical advantage, 311, 313
mechanical energy, 93
mega, 477
mel scale, 371
memory, 528
metal, 130
metamers, 420, 424, 464
micro, 477
micron, 477
microphone, 147
microscope, 268, 499
MIDI, 371
milli, 477
minimum image diameter, 236
minor second, 358
minor sixth, 358
minor third, 358, 362
mirage, 255
mirror image, 242
mirror reflection, 242
mismatch of impedances, 310
mistuned consonances, 348
modes, viii
modulation index, 346
monochromatic, 434
monochromatic light, 398
monochromators, 404
monotonically increasing, 456
musical interval, 358
musical scales, 354
musical staff, 357
myopic, 385

N
nano, 477
near point of an eye, 385, 394, 495
near-sighted, 385
negative charge, 129
negative feedback, 166
negative terminal, 131
nerve fibers, 317
nerve signals, 305
neutron, 130
Newton's Third Law, 138, 139
Newton, Isaac, 34, 128, 319, 525
Nexium, 292
night vision, 389
nodal lines, 58

node, 21, 73, 76
non-dispersive, 257
non-scientific issue, 531
nonlinear response, 336, 337
normal to a surface, 242
north pole, 131
nuclear energy, 93
nuclear force, 127
nuclei, 130
number density, 64
number of distinguishable colors, 450
numerology in tuning, 354

O
object distance, 264
objective, 413, 499
objective input, 348
octave, 356
ocular, 499
Oersted, Hans, 135, 143
Olduvai Gorge, 1
one atmosphere pressure, 65
one-to-one correspondence, 414
open pipe, 76
ophthalmoscope, 387
optic nerve, 383
optically active, 288
order of interference, 211
organ of Corti, 318
oscilloscope, 70
ossicles, 306
out of phase, 207
oval window, 306
overtone, 23

P
parallel ray, 263, 264
partial, 23, 50
partial frequency analysis, 315
partially polarized, 285
pendulum clock, 529
pentatonic scale, 355, 362
period, 15, 23, 26, 29, 30
periodic, 25
periodic wave, 354
permanent magnet, 141
permeability of free space, 171
permittivity of free space, 171
phase difference, 207
phon, 329, 331, 440
phon level, 350
phosphorescence, 197
photoelectric effect, 531
photon, 185, 531
photopic, 391
photopic vision, 390
piano keyboard, 53, 356
Picasso, Pablo, 521
pigments, 409, 453
pinna, 306
pitch, 16
pitch discrimination, 319
Place Theory of Pitch Perception, 319
Planck's Constant, 186
Planck, Max, 186, 199, 531
plane wave, 231
pluck, 15
point mass, 128
point source, 103, 209
pointillist, 414
polarization, 231
polarized light, 280
Polaroid, 281
positive charge, 129
positive feedback, 166
positive terminal, 131
potential energy, 90, 96
power, 87, 99, 313
pressure fluctuations, 503
primary illuminants, 441
Principle of Conservation of Energy, 88
Principle of Relativity, 147
principle of superposition, 205
prism, 180, 256
probability density, 191
probability mode, 192, 193
protanope, 458
protanopia, 458
proton, 130
psychoacoustics, 328
pulse, 12, 20
pure tone, 349, 354
purity, 432
purple colors, 464
Pythagorean comma, 367
Pythagorean intervals, 362
Pythagorean tuning, 359, 362

Q
quadratic nonlinear response, 337
quantum corral, 192
quantum energy level, 193
quantum state, 190
quantum theory, 190, 529
quarter comma, 381

R
radiation decoupling after the Big Bang, 287
rainbow, 180
rarefaction, 68
Ravel, Maurice, 6, 373
Rayleigh scattering, 196
real image, 265, 271
reality, 531
recovery time, 390, 460
rectangular wave, 70
red cone, 390, 415
redshift, 278
reducing to the octave, 362
reference level, 107
reflectance, 245
reflected, 245
reflection, 231
refraction, 180, 231, 249
Reissner's membrane, 315
relative humidity, 115
relative phase, 47, 207
relative probability, 191
relative velocity, 277
resolution, 495
resonance, 43, 195
resonance fluorescence, 196
restoring force, 36, 41, 66, 486
Retina, 383
reverberation time, 118
reversibility, 269
rhodopsin, 390
Rhythm Theory of Pitch Perception, 319, 321, 324
right handed, 288
rods, 387
Roederer, Juan, 329
Rossi, Salomon de, 9
rotatory power, 289
Rutherford, Ernest, 183

S
Sabine's Law, 119
saturated, 401
saturation, 413, 418, 431
sawtooth wave, 70
scala media, 315
scala tympani, 315
scala vestibuli, 315
scattered wave, 238
scattering, 231
scotopic, 391
scotopic vision, 389
second harmonic, 20
Second Law of Thermodynamics, 95
second order beats, 348
selective absorption, 403
semitone, 358
sense of color, 383
seventh, 359
Shepard's Staircase, 345
Shepard, R.N., 345
shiny surface, 241
shock wave, 276
silk screen, 218
sine function, 23, 24, 29
sine wave, 23
sinusoidal, 24, 27
solar constant, 103
solenoid, 139
sone, 331, 440
sound density, 69
sound level, 105, 107, 329, 331, 350
sound pressure, 69, 73, 310
sound pressure level, 105
sound pulse, 13
south pole, 131
sparkling surface, 241
specific rotatory power, 289
spectral colors, 399
spectral intensity, 10, 397, 402, 429
specular reflection, 242, 244
specularly, 295
spherical aberration, 260, 262
spherical wave, 263
spontaneous emission, 186, 188, 197
spontaneous transition, 188
spring constant, 27, 59, 336
squillo, 331
sRGB, 450
standing wave, 23
stapes, 306
static electricity, 127
steps (musical), 356
stereoscope, 392
stiffness, 41, 257
stiffness of basilar membrane, 322
stimulated emission, 197
stirrup, 306
stretch tuning, 54
stretch tuning of pianos, 379, 380
stroboscope, 122, 203
sub-matrix, 508
subjective, 413
subjective perception, 348
subtractive mixing, 410, 463
subtractive primaries, 416
surface, 249

symmetry, 269
synthesizer, 344
syntonic comma, 367, 378, 380

T
table of color matching functions, 432, 507
tectorial membrane, 321
telescope, 268
temperament, 354
tension, 11, 34, 39
Terhardt, Ernst, 344
tetrachromacy, 464
tetrachromats, 464
tetranope, 458
Thales of Miletus, 127
thalidomide, 292
thermal energy, 88, 93, 94, 98
thin-lens equation, 264
three primary colors, 414
three-primary representation, 10
threshold of *aural* pain, 501
threshold of hearing, 331, 501
threshold of pain, 331, 501
timbre, 17
time, 528
time order, 528
total internal reflection, 252, 255
transformation between sets of primaries, 437
transmission coefficient, 386
transmittance, 246, 283, 284, 386, 403
transmitted, 245
transmitted wave, 249
transverse, 12, 20
traveling wave, 20, 23
trichromats, 464
tristimulus value, 425, 429
tritanopia, 458
tritone, 375
truth, 530
tuning, 354, 359, 373
twang, 331
tympanic membrane, 306

U
ultimate eye object, 495, 496
unilateral dichromats, 458
unpolarized light, 280
upright image, 266

V
vector field, 150
velocity, 175
velocity amplitude, 96
vibrato, 48, 50, 346, 368
virtual image, 243, 266, 396
virtual pitch, 344
viscosity, 315
visual purple, 390
vitreous humor, 387

W
Wald, George, 383, 384, 390
Watt, 87, 100
Watt, James, 100
wave generator, 70
wave packet, 187, 429
wave phenomena, 11
wave propagation, 12
wave velocity, 12, 30
wavelength, 26, 30
weak force, 127
weight, 91
well tempered tunings, 370
Well-Tempered Clavier, 370
Werkmeister I(III) Temperament, 380
Werkmeister, Andreas, 370, 380
Western Music, 360
whip, 276
white, 415, 422, 429
white noise, 504
Whitman Walt, 1
whole tone, 358, 362
work, 96, 98

Y
Young's modulus, 41, 42